IMPACT OF INTERRATER RELIABILITY ON THE CONSTRUCT VALIDITY OF ASSESSMENT CENTERS POST-EXERCISE DIMENSION RATINGS (PEDRS) USING SINGLE VERSUS MULTIPLE RATERS

by

SABRINA DROLLINGER

(Under the Direction of Charles Lance)

ABSTRACT

The purpose of the present research was to re-examine the traditional and experimental methods used by Kolk, Born, and van der Flier (2002) for their impact on the construct validity of assessment centers (AC). Data for this study were AC ratings for law enforcement officers. I calculated the reliability of the multiple raters for each dimension within an exercise, then using these reliabilities I corrected the correlations in the multi-trait multi-method (MTMM) matrix for attenuation due to unreliability in the single ratings for the different dimension-same exercise correlations. Results indicate there were no differences between the multiple raters model and any of the single rater models. Results are discussed in terms of construct validity of ACs and future direction for investigating the construct validity problems of ACs.

INDEX WORDS: Assessment Center, Construct Validity, Multiple Raters

IMPACT OF INTERRATER RELIABILITY ON THE CONSTRUCT VALIDITY OF ASSESSMENT CENTERS POST-EXERCISE DIMENSION RATINGS (PEDRS) USING SINGLE VERSUS MULTIPLE RATERS

by

SABRINA DROLLINGER

B.S., Missouri State University, 2003

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

2007

© 2007

Sabrina Drollinger

All Rights Reserved

IMPACT OF INTERRATER RELIABILITY ON THE CONSTRUCT VALIDITY OF ASSESSMENT CENTERS POST-EXERCISE DIMENSION RATINGS (PEDRS) USING SINGLE VERSUS MULTIPLE RATERS

by

SABRINA DROLLINGER

Major Professor:	Charles Lance
Committee:	Lillian Eby
	Richard Marsh

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007

DEDICATION

I dedicate this work to my mother. You have been my rock. I would not be the person I am today without your constant support and love. You have earned this right along with me. Thank you for everything. You have made a world of difference in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

INTRODUCTION

Purpose of This Study

There has been continued discussion and research on the construct validity paradox of assessment centers (AC). Researchers have been working to understand why ACs have good predictive validity but fail to show construct validity. Much research has been done to improve the construct validity of ACs with findings providing little evidence of improvement. One study was done to investigate the usefulness of using a single rater to rate a dimension within one exercise only. This study will be done to evaluate the usefulness of using a single rater compared to using multiple raters to rate all dimensions within an exercise. It is expected that the multiple rater design will be more statistically appropriate, as well as, more practical than the single rater design.

IMPACT OF RELIABILITY ON CONSTRUCT VALIDITY OF ASSESSMENT CENTERS POST-EXERCISE DIMENSION RATINGS (PEDRS) USING SINGLE VERSUS MULTIPLE RATERS

Assessment centers (AC) are used by many organizations for multiple purposes including selection and placement, promotion, and the identification, training and development of managers and potential managers (Gaugler, Rosenthal, Thornton, & Bentson, 1987). ACs are useful to organizations due to their demonstrated predictive validity (Schmitt & Hunter, 1998) and low rate of adverse impact (Hoffman & Thornton, 1997). Thus, there has been a consistent increase in the use of ACs by organizations as a tool to identify those individuals expected to perform the specific duties and jobs within an organization well.

However, the preponderance of research over the past couple of decades indicates the intended structure of ACs may not be an accurate description of the actual structure (Lance, in press). While the AC is predictive of external criteria such as job performance, the constructs being measured do not seem to be the stable abilities (i.e. leadership, problem solving, etc.) that ACs were originally thought to measure. This line of logic has been supported by the lack of evidence for the construct validity of ACs (Lance, in press). Attempts have been made to modify the AC structure in order to improve construct validity. Some modifications have resulted in improved evidence for construct validity, though, researchers have been unable to obtain the evidence necessary to concluded ACs are construct validity (Silverman, Dalessio, Woods, & Johnson, 1986; Bycio, Alvares, & Hahn, 1987; Gratton, & Sharpley, 1987; Brannick, Michaels, & Baker, 1989; Gaugler & Thornton, 1989; Robertson, Gratton, & Sharpley, 1987; Schneider & Schmitt, 1992; Harris, Becker, & Smith, 1993; Chan, 1996; Kleinmann & Köller, 1997; Lance, Foster, Thoresen, & Gentry, 2004).

Robertson and his colleagues (1987) argued that the lack of construct validity in AC research was due to consistent biases in within-exercise dimension ratings or post-exercise dimension ratings (PEDRs) which are given by the same raters within the same exercise at the completion of an exercise. Additionally, these PEDRs are given by a different set of raters in different exercises with each set of raters only providing PEDRs for one exercise. Rater biases would be consistent across different dimensions within the same exercise because PEDRs for each dimension are provided by the same set of raters. However, these biases would not be consistent within the same dimension across exercises because a different set of raters provide PEDRs in different exercises. This then would lead to higher correlations of PEDRs within an exercise between different dimensions (across the same raters) than the correlations between PEDRs of the same dimension across different exercises (across different raters). Additionally, AC participants might also perform differentially from one exercise to another and PEDRs within an exercise might reflect a participant's overall performance not just performance of a specific dimension (Robie, Osburn, Morris, Etchegaray, & Adams, 2000; Silverman et al., 2000).

In 2002, Kolk, Born, and van der Flier created an experimental design to eliminate these rater and exercise biases by having a single rater provide a PEDR for a single dimension within a single exercise (see Table 2.3 for an example of this design). They did find improvements in discriminant validity evidence using this design. However, there is an alternative explanation for the obtained results. Using only one rater for each exercise by dimension combination could be reducing the reliability of the PEDRs. Reducing the number of ratings obtained reduces the reliability of the ratings and could be attenuating the correlations between the ratings of different dimensions across different exercises instead of eliminating biases (Hoyt, 2000).

The current study is important for several reasons. First, it is an attempt to show using single raters instead of the traditional multiple rater design is not an improvement in the structure of ACs and instead reduces the reliability of the AC within-exercise dimension ratings. Second, this study makes a contribution to the literature by showing attempts to fix the design of the AC is unlikely to improve the construct validity of the AC. Finally, the findings from this study are expected to support a new view of the AC structure indicating exercise variance is not just bias as Kolk, Born, and van der Flier (2002) might argue but is true variance reflecting performance that is cross-situational specific (Lance, et. al., 2000).

History of Assessment Centers

Assessment centers were originally created in the mid-1940s during World War II. The United States Office of Strategic Services (OSS) needed to quickly evaluate both military personnel and civilians for intelligence work during the war. Dr. Henry Murray developed a new method for testing individuals' abilities to handle the hazardous work conditions of intelligence gathering. This new method required an individual to participate in various exercises designed to elicit the individual abilities believed to be relevant to successful spying, (e.g., ability to handle stress), and observers then rated the extent to which individuals effectively displayed these abilities (Moses, 1977).

Following the war, Robert Greenleaf and Douglas Bray altered the AC model used by the OSS and created an assessment center for the American Telegraph and Telephone (AT&T) company as part of the Management Progress Study. This longitudinal study tracked the career progress of a sample of young men beginning their careers at AT&T. The AC was a measure of the men's initial effectiveness as managers and used strictly for research purposes to predict their later career success (Bray, Campbell, & Grant, 1974; Moses, 1977). The AC predicted career

success so well that other companies began to use the AC as a selection tool, and it has since become a widely used method within many organizations (Gaugler et al, 1987; Kudisch, Ladd, Dobbins, 1997; Spychalski, Quiñones, Gaugler, & Pohley, 1997).

The original design of the AC was defined by multiple exercises that were meant to tap certain dimensions (Moses, 1977). Successful completion of the AC exercises was believed to be contingent upon the possession of relevant abilities. Expert observers would examine the participant's performance and then rate the extent to which the participant displayed each trait or dimension the exercise was meant to elicit. The ratings were then used to select those individuals that most effectively displayed the traits required of the job for which the AC was created (Moses, 1977).

Over the past few decades two types of AC ratings have been used. The first type is known as within-dimension ratings. Using this type of rating, overall dimension ratings are made after all exercises have been completed and the assessors have had a chance to discuss which dimension rating is most appropriate for each candidate. The second type is known as within-exercise dimension ratings or postexercise dimension ratings (PEDRs) where dimension ratings are given at the completion of each exercise. Table 2.1 provides an example of the architecture of within-exercise dimension ratings. From this table, the within-exercise dimension ratings or PEDRs would be those that are given for each dimension by exercise combination. For example, following the in-basket exercise dimension ratings would be given for the dimensions energy, fact finding, and oral communication. The assessors then use these PEDRs to discuss and come to a consensus on final dimension ratings and then provide a summary, overall rating (OAR) of the candidate's AC performance. Assessors using within-dimension ratings only provide consensus-based final dimension ratings and OARs once all exercises have been completed.

Using this type of rating, only the final column of Table 2.1 would be completed whereas, the entire table is completed using the within-exercise dimension ratings.

Construct Validity Paradox

The validity of the AC was originally established by content-related validity and criterion-related validity. In 1987, Sackett reviewed the content-related validity AC literature. He concluded that content-related validity had been established for ACs. However, he noted there were components of AC construction that needed to be considered and were not covered by simply establishing content-related validity. For example, such things as using job analysis as the bases for AC construction, instructions to candidates in an AC, the method for scoring dimensions, and other various design issues could impact validity. Sackett (1987) warned that claims of content-related validity made without regard for these important issues could lead to a lack of other types of validity (i.e. construct-related validity). As an example of this, Sackett (1987) highlighted the construct validity paradox which has over the past few decades become the focus of much of the research on ACs.

Sackett and Dreher (1982) were among the first researchers to investigate the construct validity of ACs using Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach. Using MTMM, a correlation matrix is computed across traits and methods using ratings of AC candidates' performances. Table 2.2 provides an example of a MTMM matrix. Traits are defined as the dimensions of the AC and methods are the exercises used. Once the ratings have been correlated, a visual examination or a confirmatory factor analysis (CFA) of the correlations is then done to determine whether there is support for convergent and discriminant validity of the AC. For the purposes of clarity, the terms dimension and exercise are used here in lieu of the MTMM terms traits and methods when discussing the validity of ACs. To establish convergent

validity, the same dimension-different exercise (SDDE) correlations, those correlations between the same dimensions in different exercises, should be significant and relatively high when compared to the rest of the correlations matrix. In Table 2.2, an example would be the correlation between ratings for Dimension 1 in Exercise 1 and ratings for Dimension 1 in Exercise 2. To establish discriminant validity, SDDE correlations should be larger than the different dimension-different exercise (DDDE) correlations, those correlations between different dimensions in different exercises, demonstrating discriminability between different dimensions. From Table 2.2, an example of DDDE correlations would be the correlation between ratings for Dimension 1 in Exercise 1 and ratings for Dimension 2 in Exercise 2. A more stringent test of discriminant validity requires the SDDE correlations to be larger than the different dimension-same exercise (DDSE) correlations, those correlations between different dimensions in the same exercise. An example of DDSE correlations from Table 2.2 would be the correlation between ratings for Dimension 1 in Exercise 1 and ratings for Dimension 2 in Exercise 1. Sackett and Dreher examined three organizations' ACs and failed to find evidence supporting the convergent and discriminant validity of these ACs.

With the exception of one study (Arthur, Woehr, & Maldegen, 2000), there continues to be a lack of support for both convergent and discriminant validity and consequently construct validity for AC PEDRs. While some studies have found support for convergent validity with very specific construction requirements (e.g. limiting the number of dimensions) for the ACs used (Silverman, et. al. , 1986; Brannick,, et. al., 1989; Gaugler & Thornton, 1989), these same studies have failed to find corresponding support for discriminant validity. Further, other studies have found a lack of evidence for both convergent and discriminant validity (Bycio, et. al., 1987; Robertson, et. al, 1987; Schneider & Schmitt, 1992; Harris, et. al., 1993; Chan, 1996; Kleinmann

& Köller, 1997; Lance, Foster, et. al., 2004). Based on these and other findings, overwhelming evidence suggests that ACs clearly *do not* demonstrate construct validity (see Lance, Lambert, Gewin, Lievens, & Conway, 2004 for a review).

The apparent paradox is that although there has been little evidence to support the construct validity of the AC, there continues to be support for criterion-related validity. Research indicates ACs are predictive of career success over time (Jansen & Stoop, 2001). Both a narrative review (Sackett & Tuzinski, 2001) and a meta-analysis (Gaugler et al., 1987) of AC literature also provide support for the criterion-related validity of ACs for various criteria (i.e. job performance, salary progress, etc.). There seems to be only one study to date that has not found evidence supporting the criterion-related validity of ACs (Chan, 1996). Therefore, organizations and researchers alike have concluded that ACs are useful predictors of job performance related criteria and overall career success.

Modifications to Improve Construct Validity

There has been some speculation about the reason for the lack of congruency between content-related validity, criterion-related validity and construct-related validity. One suggestion has been that the cognitive demand on observers is so great that observers are not able to adequately discriminate between the dimensions within an exercise. Therefore, if observers within an AC find it difficult to manage the amount of information about the behaviors displayed in an exercise, they will reduce the information so that it can more easily be categorized (Gaugler & Thornton, 1989).

Researchers have investigated the impact of different modifications in the design of ACs to test the theory of cognitive overload. Gaugler and Thornton (1989) investigated whether fewer dimensions would reduce the cognitive demand and therefore improve the discriminability

between dimensions. They examined the difference in dimension ratings for three groups, with groups rating three, six, or nine dimensions. While rating fewer dimensions did improve the accuracy of classification of behaviors into dimensions, it did not improve the accuracy of observing behaviors nor did it impact convergent or discriminant validity (Gaugler & Thornton, 1989). Although Gaugler and Thornton failed to find an impact on construct validity, in a meta-analysis of the AC literature, Woehr and Arthur (2003) did find having fewer dimensions improved convergent validity. However, these results may differ from Gaugler and Thornton's (1989) due to the greater and more variable number of dimensions used in the studies that were part of the meta-analysis.

Another modification to reduce cognitive load investigated has been the use of behavioral checklists. It was argued that if the specific behaviors were categorized for the observer, then the observer could more easily discriminate between different dimensions without having to recognize and categorize the expected behaviors merely from a definition of the dimension (Donahue, Truxillo, Cornwall, & Gerrity, 1997; Reilly, Henry, & Smither, 1990). Findings indicated behavioral checklists did improve discriminant validity (Donahue et al., 1997), reduced between-judge variability (Hennessy, Mabey, & Warr, 1998), and improved convergent validity (Reilly et al., 1990) but none resolved the construct validity problem. Finally, two other modifications meant to reduce cognitive load investigated were the transparency of the dimensions and the participant-to-assessor ratio. The more transparent the dimension was to an assessor the more the construct validity of the AC improved (Kolk, Born, & van der Flier, 2003). However, a lower participant-to-assessor ratio has not been shown to improve construct validity (Gaugler et al, 1987; Woehr & Arthur, 2003). Overall, while some of the modifications to reduce

cognitive load have improved the construct validity of ACs, the improvements were not significant enough to eliminate the concerns over the construct validity problems.

Other modifications of the AC structure, unrelated to the cognitive overload theory, that have been investigated to improve construct validity include: the type of assessor, type and length of assessor training, and exercise form and content. Psychologists and experienced assessors, because they are familiar with the dimensions within an AC, do tend to provide ratings that increase the construct validity of the AC when compared to managers, supervisors, and inexperienced assessors (Kolk, Born, & van der Flier, & Olman, 2002; Woehr & Arthur, 2003). Although results for Woehr and Arthur's (2003) meta-analysis indicated longer training of assessors shows increased convergent and discriminant validity, Gaugler and colleagues (1987) determined from their meta-analysis "that within the range of number of days of training studied (.5-15), more training does not lead to high validities" (p. 505). Additionally, it has been shown that using frame-of-reference training to train assessors improves the accuracy and reliability of ratings as well as the criterion-related validity, discriminant validity, and to some extent the convergent validity of ACs (Schleicher, Day, Mayes, & Riggio, 2002). Although there were some improvements in the construct validity using these modifications, the results of these studies have still failed to demonstrate construct validity for ACs.

Types of ratings in ACs have also been manipulated as a means to improving convergent validity. It has been argued that within-exercise ratings increase the likelihood that assessors rate dimensions based on overall performance within an exercise, introducing halo error in the dimension ratings (Robie et. al. 2000; Silverman et al., 2000) which is known as common exercise bias. Common exercise bias results in correlations between different dimension ratings in the same exercise (DDSE) being larger than the correlations between similar dimension

ratings across different exercises (SDDE). As mentioned before, this pattern of correlations indicates a lack of construct-related validity. One proposed solution has been to use the within-dimension rating strategy to lower or eliminate this form of halo error. If the assessors are required to make ratings within a dimension then they will be less prone to use the overall performance in any one exercise as a basis for the dimension ratings.

There have been mixed results concerning the impact of within-dimension ratings compared with within-exercise ratings on the construct validity of ACs. Harris et al. (1993) as well as Silverman et al. (1986) found within-dimension ratings did not provide evidence of construct validity although there was some indication that discriminant validity was slightly better using within-dimension ratings. Conversely, Woehr and Arthur (2003) found some evidence to support the contention that within-dimension ratings improve construct validity. However, they state that these findings may be due to the fact that within-dimension ratings in the studies that were part of their meta-analysis typically required the same assessors to observe the same candidates across exercises. The increase in the correlations between the dimension ratings could then be attributed to common rater variance and not an improvement in the dimension ratings. That is, dimension ratings given by the same raters will have more consistent biases then the ratings given by different raters. This consistency in biases could then lead to increase in the variance accounted for in a dimension that is unrelated to more accurate dimension ratings but instead is a reflection of the common rater bias. Robie, et al. (2000) also found support for convergent and discriminant validity using within-dimension ratings but with this same limitation of using the same raters across exercises mentioned by Woehr and Arthur (2003).

As an extension of the research performed by Robie et al. (2000), Kolk, Born, and van der Flier (2002) further investigated the impact of the within-dimension rating method compared with the within-exercise rating method on construct validity. To disentangle the effect of variance associated with "true" effects and the variance that may be associated with common exercise bias and common rater bias, they conducted two studies. In the first study, they had two groups of assessors. One group of assessors rated the candidates on all dimensions within an exercise, the traditional within-exercise rating method. The other group of assessors rated the candidates on only one dimension within an exercise (experimental method). Table 2.3 is an illustration of the experimental design with each rater providing PEDRs independently across dimensions and across exercises. Assessors were rotated so that they only rated a candidate on one exercise by dimension combination. By having each rater only rate each candidate on one exercise by dimension combination; raters evaluated each candidate on only one dimension for only one exercise. This design then eliminates the possibility of common exercise variance by having each candidate evaluated by each specific rater in only one exercise eliminating the carry over of biases (due to dimension ratings based on overall performance in an exercise) in dimension ratings within an exercise. Additionally, this design eliminated the possibility of common rater variance by having each candidate evaluated by each specific rater on only one dimension rating eliminating the possibility of carry over of biases associated with a specific rater in dimension ratings both within and between exercises. Results indicated there was no difference in convergent validity between the within-exercise rating method and experimental method. However, there were significant differences found in discriminant validity. The experimental group had smaller DDSE correlations indicating improved discriminant validity over the within-exercise rating method, although the DDSE correlations were still larger than the

SDDE correlations indicating a general lack of discriminant validity. They posited that the differences between the two methods may have been due to greater cognitive demand for the traditional method group because they were required to rate a larger number of dimensions (3 versus 1) within an exercise. Kolk, Born, and van der Flier's (2002) second study was created to eliminate this explanation as a possible reason for the differences found between the two methods.

In the second study, one group of assessors rated all dimensions within an interview simulation exercise, another group rated all dimensions within a client interview exercise, and another group rated all the dimensions within an analysis and presentation exercise. When analyzing the data, the dimension ratings from all assessors (3 ratings per dimension per exercise) were included in the final MTMM matrix to represent the within-exercise rating method. For the experimental method, dimension ratings were randomly chosen from the ratings collected to be included in the final MTMM matrix. For this method, only one PEDR was used from each rater for a candidate. This would then create a matrix that parallels the design in the first study. See Table 2.3 for an illustration of the experimental design. The results again indicated no differences in convergent validity between the two methods. Also like the first study, neither method provided total support for discriminant validity although the DDSE correlations were smaller for the experimental group. Referring back to Table 2.2, the DDSE correlations (in italics) should be relatively small to support discriminant validity of the AC indicating low correlations between different dimensions within the same exercise. While the DDSE correlations were still larger than the SDDE correlations, the average difference between the DDSE and SDDE correlations were smaller for the experimental group. This was due to smaller DDSE correlations in the experimental group and not to differences in the SDDE

correlations between the two groups. From Table 2.2, the SDDE correlations (in bold) should be larger than the DDSE (in italics) providing additional support for discriminant validity of ACs indicating a stronger relation between ratings of the same dimension across exercises (SDDE) and a weaker relation between ratings of different dimensions within the same exercise (DDSE). This then indicates the assessors are able to discriminate between the conceptually distinct dimensions so that dimension ratings within an exercise produce smaller correlations than correlations across exercise within the same dimension. An example using Table 2.1 would be the correlation between ratings given for energy in the in-basket and energy in the leaderless group discussions should be larger than the correlation between the ratings given for energy in the in-basket and ratings given for fact finding in the in-basket to support discriminant validity.

The design problems with the Kolk, Born, and van der Flier (2002) studies are two-fold. First, the use of one rater per dimension per exercise is not a practical application for most organizations using ACs. Most organizations do not have the available resources to provide enough raters to cover all dimensions across all exercises. The AC is already considered an expensive tool (Hoffman & Thornton, 1997) when compared to other tools utilized by organizations. It would therefore be difficult to convince organizations to increase the expense of using an AC in order to improve construct validity, especially when the AC has already been shown to have good predictive validity. The second design flaw with the Kolk, Born, and van der Flier (2002) study is the use of a single rater for a dimension within an exercise. First, the use of multiple ratings in the traditional method and the use of a single rater in the experimental method could be an explanation for the differences found between the groups. Using the single rater design, Kolk, Born, and van der Flier (2002) found smaller DDSE correlations indicating assessors provided less similar ratings for different dimensions within an exercise. This lead to

smaller differences in the SDDE correlations (which were not significantly different between the two designs) and DDSE correlations using the single rater design. They argued the improvements in discriminant validity evidence were due to the elimination of common rater variance because each PEDR was made independently across the dimensions and across the exercises by a single rater. They argued that the DDSE correlations are inflated in the traditional within-exercise dimension rating method because the same rater is rating the different dimensions within the exercise thus contributing to common rater variance. However, another possible explanation for smaller DDSE correlations using the single rater design could be attributed to attenuation of the DDSE correlations due to decreases in reliability when using a single rater design. A well-known result in the psychometric literature is that, all other things being equal, the more test items that comprise a test, the more reliable a test is. Accordingly, the mean of multiple assessors' ratings will, all other things being equal, be more reliable than a single assessor's rating. Thus, reduced reliability and consequent attenuation of the DDSE correlations is a viable alternative explanation for Kolk, Born, and van der Flier's (2002) findings. The focus of the present research is to determine whether attenuation due to unreliability or common rater variance in the multiple rater design is the explanation for the differences found and whether the use of multiple raters or single raters is the more appropriate design.

Interrater Reliability Issues Concerning Single Raters

While the use of a single rater for each dimension within an exercise may seem to improve the construct validity of ACs, research indicates that the use of only one source of information for a construct is a methodological weakness (Hoyt, 2000). The use of multiple raters contributes to the objectivity of the information obtained (Moses, 1977). Difference in

ratings given by different raters can be attributed to differences in exposure to target behavior, differences in interpretation of target behavior, and the extent to which raters incorporate irrelevant information into the ratings (Kenny, 1991). If only one rater provides a rating for a dimension in an exercise, then these differences in rating behavior may lead to a deficiency of relevant information and contamination in information due to biases of the rater that can not be corrected when there is only one rater (Bock, Brennan, & Muraki, 2002). Instead the more appropriate design using multiple raters allows for correction due to attenuation and aids in reducing biases associated with the raters (Hoyt, 2000). Therefore I hypothesized that

H1: The single rater design with correlations not corrected for unreliability would have significantly lower DDSE correlations than the multiple rater design.

H2: The single rater design correlations corrected for unreliability would no longer have significantly different DDSE correlations when compared to the multiple rater design.

METHOD

Data

The participants in this study were law enforcement officers participating in an AC used for promotion. This sample of data was collected as part of on-going investigation of AC structure. The data included three dimension ratings for each participant for each exercise. The three dimensions investigated were (a) perception (PER)—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, (b) judgment (JUD)—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 field operations and (c) organizing and planning (O&P)—establishing a course of action for self and/or others in order to accomplish a mission or work assignment, including planning the proper assignments of personnel and the appropriate allocation of resources and the organization of such personnel and resources. The three exercises consisted of (a) a role play (RP)—candidates were given a packet of information in advance and during the RP exercise were asked to provide answers using this information to a simulated supervisor, (b) an oral presentation (OP)—the participant were given information which had to be used to summarize the problems in the information and present a proposed plan of action to resolve the problems, and (c) a work history report (WH)—in which candidates were given 8 hr to provide a summary of their credentials for promotion.

The actual structure of this AC measures additional dimensions. Within the RP and OP exercises, the dimensions decisiveness, oral communication, and leadership were also measured. Within the WH exercise, the dimension written communication was also measured. As can be seen, the additional dimensions measured in the RP and OP exercises were not measured in the WH exercise and likewise the dimension measured in WH was not measured in the RP and OP exercises. To answer the empirical questions for this study, all dimensions included in the analyses had to be measured in all of the exercises. This allows for comparisons of correlations between the same dimensions across all 3 different exercises. Therefore, only dimensions measured in all exercises were included as part of the study.

Assessors

Assessors were law enforcement officers that were at least the rank candidates were being evaluated on for promotion. There were three groups with three assessors per group. One group of assessors evaluated participants in the RP exercise, one group evaluated participants in

the OP exercise, and one group evaluated participants in the WH exercise. Assessors participated in an 8-hour training session where the activities of candidates and assessors were explained. During the first four hours of training, assessors were instructed on general information about conducting assessments. For example, explanations were given about each dimensions and exercise, rater error, and the use of behavioral checklists.

For the second four hours of training, each assessor was given practice both observing the behaviors of the exercise they would be assessing and actually evaluating these behaviors using frame-of-reference training. Assessors were first instructed on taking copious notes of behavior exhibited by a candidate in the RP and OP exercises. Notes were not taken during the WH exercise because this involved reading a candidate's written report and could be referred back to when needed whereas this could not happen in the RP and OP exercises. Assessors were then instructed on providing task-based ratings for the behavioral checklists within each dimension. Following this, assessors were then instructed on providing dimension ratings. Finally, assessors were instructed on comparison of dimension ratings to identify and resolve any discrepancies more than one scale unit apart. This type of frame-of-reference training is similar to the type of training described by Schleicher and her colleagues (2002). This type of training was used because it has been shown to help assessors apply more consistent standards, increasing the reliability and accuracy of ratings (Schleicher et. al, 2002).

Traditional and Experimental Method

In line with the Kolk, Born, and van der Flier (2002) study, the traditional (multiple rater) design utilized the mean of all three ratings for each dimension within each exercise (mean of PER, JUD, O&P from Assessor 1, 2, 3 in RP; mean of PER, JUD, O&P from Assessor 4, 5, 6 in OP; mean of PER, JUD, O&P from Assessor 7, 8, 9 in WH). The experimental (single rater)

design utilized the dimension rating of only one rater for each dimension within an exercise. The raters for the experimental method were randomly chosen using a random number table. For this design, only one dimension rating within one exercise was used from each assessor (e.g. PER from Assessor 1 in RP; JUD from Assessor 2 from RP; O&P from Assessor 3 in RP; PER from Assessor 4 from OP; JUD from Assessor 5 from OP, and so on). This created a design parallel to the Kolk, Born, and van der Flier (2002) (see Table 2.3).

Interrater reliabilities were calculated using data from the traditional design using all three assessor ratings within an exercise. Thus, reliabilities were calculated for all dimension by exercise combinations (e.g. PER x RP, PER x OP, and PER x WE, etc.). To provide both lower and upward bound estimates of reliability three different types of reliability estimates were calculated. Interrater correlations were calculated between the three assessors for each dimension by exercise combinations and then averaged. For example, for the Perception x RP combination, three interrater reliabilities (Pearson correlations across candidates between raters 1 and 2, rater 1 and 3, and raters 2 and 3) were calculated and then averaged. Next, two types of intraclass correlation (ICC) were calculated for each dimension by exercise combination. The ICC (1,1) and the ICC (1,k) were calculated from one-way analysis of variance (ANOVA) results with candidates as the between factor and raters as the within factor (Lahey, Downey, & Saal, 1983). These reliabilities were then used to correct the DDSE correlations for attenuation for the experimental design.

Confirmatory factor analyses were done to determine if there were differences between the DDSE correlations for the traditional and the two experimental designs (corrected and uncorrected for attenuation). First, MTMM matrices of correlations for exercise by dimension combinations were created for both the traditional and the experimental design. The DDSE

correlations for the experimental design were corrected for attenuation due to unreliability creating a new set of correlation vectors.

Using a CFA framework, the corresponding DDSE correlations (i.e. correlations between PER ratings and JUD in the RP exercise) were constrained to be equal between the traditional design and the uncorrected experimental design. The following fit indices were used for model evaluation: chi-square statistic (χ^2), comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMSR). Based on Hu and Bentler's (1999) recommendations the following cutoff criteria were used for model evaluation: .95 or higher for CFI and TLI, .06 or lower for RMSEA, and .08 or lower for SRMSR.

RESULTS

First, a visual examination of the multiple rater MTMM matrix indicated the SDDE correlations (range: .19 - .40) were not high when compared to the rest of the correlations in the matrix (see Table 4.1 SDDE correlations are in bold). This indicates a lack of convergent validity. Also, these SDDE correlations were about the same size as the DDDE (range: .18 - .38) (see Table 4.1 DDDE correlations are those neither bolded nor italicized) and were smaller than the DDSE correlations (range: .63 - .89) (see Table 4.1 DDSE correlations are italicized). This indicates a lack of discriminant validity of this AC which is consistent with previous research findings (Sackett & Dreher, 1982; Bycio, et. al., 1987; Robertson, et. al, 1987; Schneider & Schmitt, 1992; Harris, et. al., 1993; Chan, 1996; Kleinmann & Köller, 1997; Lance, Foster, et. al., 2004). This pattern of SDDE correlations being larger than the DDSE correlations has been assumed to indicate method (i.e. exercise) effects, according to Campbell and Fiske's criteria (1959).

The same pattern of results was found with the single rater design without correction for attenuation due to unreliability. SDDE correlations (range: .18 - .38) (see Table 4.2 SDDE correlations are bolded) were not high relative to the rest of the correlations in the matrix indicating a lack of convergent validity. The SDDE correlations were about the same size as the DDDE correlations (range: .18 - .38) (see Table 4.2 DDDE correlations are those neither bolded nor italicized) and were lower than the DDSE correlations (range: .58 - .83) (see Table 4.2 DDSE correlations are italicized). Again, these results have been considered both a lack of discriminant validity and an indication of method effects. Overall, the correlations using both multiple and single raters were consistent with previous research findings indicating a lack of support for convergent and discriminant validity and therefore a lack of support for construct validity (Sackett & Dreher, 1982; Bycio, et. al., 1987; Robertson, et. al, 1987; Schneider & Schmitt, 1992; Harris, et. al., 1993; Chan, 1996; Kleinmann & Köller, 1997; Lance, Foster, et. al., 2004).

I hypothesized that the DDSE correlations using single raters uncorrected for attenuation due to unreliability would be significantly smaller than the DDSE correlations using multiple raters. Results of the CFA comparing the multiple rater model and the single rater model with DDSE correlations uncorrected for attenuation indicated no significant differences between the models (χ^2 (9, N = 217) = 2.96, *n.s.*) Additional, all fit indices were within the range commonly accepted as indicating no differences in models (see Table 4.3). These findings do not support hypothesis 1 predicting significant differences between the DDSE correlations for the two models.

Although the first hypothesis was not supported, I ran the analyses to compare the multiple raters model to the single rater models with corrected DDSE correlations. MTMM

matrices with corrected DDSE correlations can be found in the appendices in Tables A, B, and C. I hypothesized that the single rater DDSE correlations corrected for attenuation due to unreliability would no longer be significantly different from the multiple rater DDSE correlations. The same analysis and fit indices were used to compare the multiple rater model and the single rater models with corrected DDSE correlations using interrater reliability, ICC(1,1) and ICC(1,k). The results of the CFA indicated there were no significant differences in the multiple rater model and the single rater models with DDSE correlations corrected for attenuation due to unreliability using interrater correlations (χ^2 (9, N = 217) = 2.05, *n.s.*) , ICC (1,1) (χ^2 (9, N = 217) = 1.66, *n.s.*), and ICC (1,k) (χ^2 (9, N = 217) = 2.19, *n.s.*). Again, all fit indices were within the ranges commonly accepted as indicating no differences in models (see Table 4.3). Although these results do indicate no significant differences between the models, lack of support for the first hypothesis limits the interpretation of these analyses in reference to the second hypothesis. Further discussion of these findings is presented later in this paper.

DISCUSSION

The purpose of this study was to examine whether differences in the DDSE correlations between PEDRs provided by multiple raters versus single raters were due to elimination of rater and exercise biases or decreased reliability using single raters. There were two main findings from this study. First, DDSE correlations between single rater PEDRs were not significantly smaller than DDSE correlations between PEDRs provided by multiple raters. Second, DDSE correlations corrected for attenuation due to unreliability between single rater PEDRs were not significantly different than DDSE correlations between PEDRs provided by multiple raters.

First, DDSE correlations between single rater PEDRs were not significantly smaller than DDSE correlations between PEDRs provided by multiple raters. While this finding is surprising,

one possible explanation for this finding differing from the findings in Kolk, Born, and van der Flier's (2002) study was the differences in structure of their AC and the AC used in this study. In Kolk, Born, and van der Flier's (2002) study, the assessors did not discuss the PEDRs following the exercise and only meet to discuss and create an overall assessment rating for each AC candidate. The assessors also did not use behavioral checklists when assessing candidates. For the data used in this study, the assessors took behavioral notes during the exercise. At the completion of the exercise, assessors then rated the candidate on specific task-based behaviors. Then, assessors individually provided PEDRs for all dimensions in the exercise. Finally, the assessors discussed differences in PEDRs and modified ratings that were more than one scale unit different. This type of design likely resulted in PEDRs that were less variable between the raters than those PEDRs in Kolk, Born, and van der Flier's (2002) design in which final PEDRs were not discussed nor modified to within one scale unit of each other. Additionally, the use of behavioral checklists has been shown to reduce between-assessor variability (Hennessy et. al., 1998), another possible explanation for lack of variability in assessor ratings. The lack of variability in the PEDRs leads to similar DDSE correlations in the single rater model with the DDSE correlations using all raters, regardless of which rater was chosen for the single rater model. This then leads to findings indicating no significant differences in DDSE correlations between the single and multiple rater models and therefore a lack of support for my first hypothesis predicting a difference.

To further support this explanation, it would be expected that PEDRs that are more consistent and less variable should lead to higher correlations. A post hoc visual examination of the DDSE correlations from this study and the DDSE correlations in the Kolk, Born, and van der Flier (2002) study was conducted. The average DDSE correlations for the Kolk, Born, and van

der Flier (2002) study using multiple raters for the interview simulation, client interview, and analysis and presentation exercises across all dimensions were .56, .59, and .59 respectively. The average DDSE correlations for this study for the RP, OP, and the WH exercises across all dimensions were .63, .76, and .68 respectively. None of their DDSE correlations was even in the range of my DDSE correlations. It would appear the PEDRs in this study are more consistent than those used in the Kolk, Born, and van der Flier (2002) study.

Although the first hypothesis was not supported, I continued the analyses to examine the impact of correcting the DDSE correlations for unreliability. Interrater reliability was used as a lower-bound estimate of reliability, ICC (1,1) was the next lowest-bound estimate of reliability used, and then finally ICC (1,k) was used as an upper-bound estimate of reliability. The comparison of the multiple rater model and the single rater models with corrected DDSE correlations using the three estimates of reliability resulted in no significant differences between the models. This was unsurprising considering there were no difference between the multiple rater model and the single rater model with DDSE correlations not corrected for unreliability. Hypothesis 2 predicted there would no longer be differences between DDSE correlations from the multiple rater design and the DDSE correlations from the single rater design once corrected for unreliability. Due to the fact that the first hypothesis was not supported and no differences were found before correcting the DDSE correlations in the single rater design, conclusions can not be drawn from the results from the analysis to examine my second hypothesis, as it was a conditional hypothesis.

Strengths, Limitations, and Directions for Future Research

Strengths of this study that are noteworthy are related specifically to the design of the AC used during data collection. The AC used for data collection was a highly developed, well-

established AC that incorporates recent research findings for improving ACs. Behavioral checklists were used in the AC which have been shown to improve rating by reducing the cognitive load of the assessors (Reilly et. al., 1990; Donahue et al., 1997; Hennessy et. al, 1998). Additionally, the assessors were well-trained using frame-of-reference training which has been shown to improve the accuracy of ratings, reliability of ratings, and criterion-related validity (Schleicher et. al, 2002). Most of the assessors that participate in this AC were experienced as well, adding to the accuracy of the ratings (Kolk, Born, van der Flier, & Olman, 2002; Woehr & Arthur, 2003).

There are also limitations of this study that should be noted. First, the PER and JUD dimensions required some similar types of behavior from the candidate for successful performance which could have made it more difficult for assessors to discriminate between these two dimensions. However, the DDSE correlations between PER and JUD for each exercise were close to the other DDSE correlations in the same exercise as can be seen in Table 4.1 and Table 4.2. In fact, in the OP exercise, the DDSE correlations between PER and JUD for both multiple (r = .84) and single rater designs (r =.79) were smaller than the DDSE correlations between JUD and OP for both the multiple (r =.89) and single rater designs (r =.83)

Finding no differences between the multiple rater design DDSE correlations and the single rater design DDSE correlations uncorrected for reliability also limited the findings. As mentioned earlier, hypothesis 2 could only be tested if hypothesis 1 was supported. Although there has been support for differences between these two designs in previous research (Kolk, Born, & van der Flier, 2002) I was unable to find such differences. Possible future research could examine data from an AC that is more similar to the AC used in the Kolk, Born, and van der Flier (2002) study. PEDRs that are potentially less reliable might provide attenuated correlations

and therefore could lead to significant differences between the multiple rater DDSE correlations and the single rater DDSE correlations. Then the DDSE correlations could be corrected for attenuation due to unreliability and compared to the multiple rater DDSE correlations. This then might lead to a better examination of the two different designs.

Another possible goal for future research would be to use Kolk, Born, and van der Flier's (2002) data and correct their DDSE correlations for unreliability. This would eliminate the need to create a design that would replicate their study. Using their data, eliminates the need to create an AC that provides less reliable PEDRs which would not be advisable considering the psychometric goal for establishing reliability and validity of an AC. However, if other individuals already possess data from an AC similar to Kolk, Born, and van der Flier (2002), then it would not be absolutely necessary to use their data.

Implications for Practice

The results of this study, while not supporting the original idea to examine the reason for the differences, did support the usefulness of using multiple raters rather than single raters. First, as mentioned earlier, the use of single raters, in lieu of multiple raters, is not only impractical but also psychometrically unsound. The lack of significant differences between the multiple rater DDSE correlations and the single rater DDSE correlations indicates using single raters is not an improvement on the design of an AC. Therefore, it is advisable for AC designers to continue to incorporate multiple raters as part of their ACs. In addition to this, AC designers should incorporate other design features empirically shown to improve PEDRs, such as behavioral checklists.

Implications for Theory

It would seem from the findings of this study and others that there are no design "fixes" that will lead to evidence for construct validity. In this study (see Tables 5 and 6), as well as those studies where design fixes have been investigated (Silverman et. al., 1986; Bycio et. al., 1987; Gratton, & Sharpley, 1987; Brannick et. al., 1989; Gaugler & Thornton, 1989; Robertson et. al., 1987; Schneider & Schmitt, 1992; Harris et. al., 1993; Chan, 1996; Kleinmann & Köller, 1997; Lance, Foster, et. al., 2004), the DDSE correlations continue to be larger than the SDDE correlations which does not support discriminant validity. Without the establishment of discriminant validity, construct validity can not be established. Even Kolk, Born, and van der Flier (2002) failed to support discriminant validity using single raters, but instead only found a reduction in the DDSE correlations when using single raters instead of multiple raters.

Researchers should now consider revising the original view of AC structure. Originally, the ACs were designed to measure different abilities that were expected to be cross-situationally consistent (Moses, 1977) but perhaps performance is cross-situationally specific as suggested by recent research (Bycio, et al., 1987; Robertson, et. al., 1987; Lievens & Conway, 2001; Lance, Foster, et. al., 2004; Lance, Lambert et. al, 2004; Lance, in press). It is time to move away from trying to fix the AC as it is and begin to view the AC as it should be. As mentioned early, while some design fixes have shown improvement in construct validity evidence (Silverman et. al., 1986; Bycio et. al., 1987; Brannick et. al., 1989; Gaugler & Thornton, 1989; Robertson et. al, 1987; Schneider & Schmitt, 1992; Chan, 1996; Kleinmann & Köller, 1997), it seems as though the gamut of possible improvements has now been tested. Exercise variance does not appear to be related to method biases but instead is a portion of the true variance in performance (Lance, et. al., 2000). AC design improvements are seemingly not going to lead to evidence supporting

construct validity. Researchers now need to investigate these new theories of AC structure and move toward understanding the true structure of the AC.

REFERENCES

Arthur, W., 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.

Bock, R. D., Brennan, R. L, & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364-375.

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

Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives*. New York: Wiley.

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

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

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

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*(5), 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.

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.

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

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.

Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*, 455-470.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

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.

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

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, *92*, 155-163.

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.

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

Kolk, N. J., Born, M. Ph., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325-337.

Kolk, N. J, Born, M. P., van der Flier, H., & Olman, J. M. (2002). Assessment center procedures: Cognitive load during the observation phase. *International Journal of Selection and Assessment, 10*, 271-278.

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

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93*, 586-595.

Lance, C. E. (in press). Why assessment centers don't work they way they are suppose to. *Interact I/On*.

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.

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 post-exercise dimension ratings. *Journal of Applied Psychology*, *89*, 377-385.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith,

D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323-353.

Lance, C. E., 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.

Moses, J. L. (1977). The assessment center method. In J. L Moses & W. C. Byham (Eds.) *Applying the assessment center method* (pp. 3- 11). New York: Pergamon.

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.

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

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M, & Adams, K. A. (2000). Effects of rating process on construct validity of assessment center dimension evaluation. *Human Performance, 13*, 355-370.

Sackett, P. R. (1987) Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 11-25.

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

Sackett, P. R., & 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.

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.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

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.

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

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.

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.

Table 2.1

Example Architecture of an AC That Uses the "Within-Exercise" Rating Method

	Exercises					
Dimensions	In-Basket	Leaderless Group Discussion	Budget Meeting	Subordinate Meeting	Analytic Problem	Final Dimension Ratings
Energy	XX	XX	XX		XX	XX
Fact Finding	XX	XX	XX	XX	XX	XX
Oral Communication	XX		XX	XX	XX	XX
Autonomy			XX	XX	XX	XX
Behavioral Flexibility			XX	XX	XX	XX
Supervision			XX	XX	XX	XX
Overall Rating						XX

Note. From "Why Assessment Centers Don't Work the Way They're Suppose To," by C. E. Lance, in press, p. 30.

Reprinted with permission from the author.

Table 2.2

Hypothetical Matrix of Correlations Between Three Dimensions as Measured in Each of Three Exercises

	Exercise 1			Exercise 2			Exercise 3		
	D1	D2	D3	D1	D2	D3	D1	D2	D3
Exercise 1:									
Dimension 1									
Dimension 2	*DDSE*								
Dimension 3	*DDSE*	*DDSE*							
Exercise 2:									
Dimension 1	**SDDE**	DDDE	DDDE						
Dimension 2	DDDE	**SDDE**	DDDE	*DDSE*					
Dimension 3	DDDE	DDDE	**SDDE**	*DDSE*	*DDSE*				
Exercise 3:									
Dimension 1	**SDDE**	DDDE	DDDE	**SDDE**	DDDE	DDDE			
Dimension 2	DDDE	**SDDE**	DDDE	DDDE	**SDDE**	DDDE	*DDSE*		
Dimension 3	DDDE	DDDE	**SDDE**	DDDE	DDDE	**SDDE**	*DDSE*	*DDSE*	

Note: DDSE = correlation between different dimensions in the same exercise; DDDE = correlation between different dimensions in different exercises; SDDE = correlation between the same dimension in different exercises. From "Why Assessment Centers Don't Work the Way They're Suppose To," C. E. Lance, in press, p. 30. Reprinted with permission from the author.

Table 2.3

Example of Study Design for Experimental Rating Method Used by Kolk, Born, and van der Flier (2002)

	Exercise 1	Exercise 2	Exercise 3
Dimension1	R1	R4	R7
Dimension 2	R2	R5	R8
Dimension 3	R3	R6	R9

Note: PEDRs were completed independently across dimensions and across exercises.

Table 4.1

MTMM Matrix for Multiple Rater Model

	RPPER	RPJUD	RPOP	OPPER	OPJUD	OPOP	WEPER	WEJUD	WEOP
RPPER	--								
RPJUD	*.74*	--							
RPOP	*.63*	*.71*	--						
OPPER	**.36**	.35	.18	--					
OPJUD	.31	**.40**	.23	*.84*	--				
OPOP	.30	.38	**.23**	*.76*	*.89*	--			
WEPER	**.23**	.25	.26	**.33**	.36	.35	--		
WEJUD	.27	**.29**	.27	.34	**.35**	.32	*.86*	--	
WEOP	.22	.19	**.19**	.29	.32	**.31**	*.77*	*.79*	--

Note. RPPER = perception dimension in role play, RPJUD = judgment dimension in role play, RPOP = organizing and planning dimension in role play, OPPER = perception dimension in oral presentation, OPJUD = judgment dimension in oral, OPOP = organizing and planning dimension in oral presentation, WEPER = perception dimension in written exercise, WEJUD = judgment dimension in written exercise, WEOP = organizing and planning dimension in written exercise.

Table 4.2

MTMM Matrix for Single Rater Uncorrected Model

	RPPER	RPJUD	RPOP	OPPER	OPJUD	OPOP	WEPER	WEJUD	WEOP
RPPER	--								
RPJUD	*.66*	--							
RPOP	*.58*	*.63*	--						
OPPER	**.31**	.31	.17	--					
OPJUD	.27	**.38**	.24	*.79*	--				
OPOP	.29	.38	**.26**	*.68*	*.83*	--			
WEPER	**.18**	.18	.22	**.27**	.30	.28	--		
WEJUD	.25	**.28**	.32	.33	**.32**	.31	*.70*	--	
WEOP	.18	.20	**.27**	.31	.31	**.30**	*.66*	*.69*	--

Note. RPPER = perception dimension in role play, RPJUD = judgment dimension in role play, RPOP = organizing and planning dimension in role play, OPPER = perception dimension in oral presentation, OPJUD = judgment dimension in oral, OPOP = organizing and planning dimension in oral presentation, WEPER = perception dimension in written exercise, WEJUD = judgment dimension in written exercise, WEOP = organizing and planning dimension in written exercise.

Table 4.3

Comparisons of Multiple Rater Design with the Single Rater Designs.

Comparison	df	χ^2	CFI	TLI	RMSEA	SRMSR
Model 1 vs 2A	9	2.97	1.00	1.02	0.0	0.038
Model 1 vs 2B	9	2.05	1.00	1.02	0.0	0.016
Model 1 vs 2C	9	1.66	1.00	1.02	0.0	0.010
Model 1 vs 2D	9	2.19	1.00	1.02	0.0	0.028

Note. Model 1 is the multiple rater model; Model 2A is the single rater model no corrected for unreliability; Model 2B is the single rater model corrected for unreliability using interrater reliability; Model 2C is the single rater model corrected for unreliability using ICC(1,1); Model 2D is the single rater model corrected for unreliability using ICC(1,k). df = model degrees of freedom; χ^2 = chi-square statistic; CFI = Comparative Fit Index; Tucker-Lewis Index; RMSEA = root mean squared error of approximation; SRMSR = standardized root mean squared residual.

APPENDICES

Table A

MTMM Matrix for Single Rater Corrected Using Interrater Correlations

	RPPER	RPJUD	RPOP	OPPER	OPJUD	OPOP	WEPER	WEJUD	WEOP
RPPER	--								
RPJUD	.75	--							
RPOP	.69	.76	--						
OPPER	.31	.31	.17	--					
OPJUD	.27	.38	.24	.85	--				
OPOP	.29	.38	.26	.75	.91	--			
WEPER	.18	.18	.22	.27	.30	.28	--		
WEJUD	.25	.28	.32	.33	.32	.31	.90	--	
WEOP	.18	.20	.27	.31	.31	.30	.84	.85	--

Note. RPPER = perception dimension in role play, RPJUD = judgment dimension in role play, RPOP = organizing and planning dimension in role play, OPPER = perception dimension in oral presentation, OPJUD = judgment dimension in oral, OPOP = organizing and planning dimension in oral presentation, WEPER = perception dimension in written exercise, WEJUD = judgment dimension in written exercise, WEOP = organizing and planning dimension in written exercise.

Table B

MTMM for Single Rater Corrected Using Intraclass Correlation (1,1)

	RPPER	RPJUD	RPOP	OPPER	OPJUD	OPOP	WEPER	WEJUD	WEOP
RPPER	--								
RPJUD	.71	--							
RPOP	.65	.71	--						
OPPER	.31	.31	.17	--					
OPJUD	.27	.38	.24	.83	--				
OPOP	.29	.38	.26	.72	.88	--			
WEPER	.18	.18	.22	.27	.30	.28	--		
WEJUD	.25	.28	.32	.33	.32	.31	.82	--	
WEOP	.18	.20	.27	.31	.31	.30	.77	.79	--

Note. RPPER = perception dimension in role play, RPJUD = judgment dimension in role play, RPOP = organizing and planning dimension in role play, OPPER = perception dimension in oral presentation, OPJUD = judgment dimension in oral, OPOP = organizing and planning dimension in oral presentation, WEPER = perception dimension in written exercise, WEJUD = judgment dimension in written exercise, WEOP = organizing and planning dimension in written exercise.

Table C

MTMM for Single Rater Corrected Using Intraclass Correlation (1,k)

	RPPER	RPJUD	RPOP	OPPER	OPJUD	OPOP	WEPER	WEJUD	WEOP
RPPER	--								
RPJUD	.68	--							
RPOP	.60	.66	--						
OPPER	.31	.31	.17	--					
OPJUD	.27	.38	.24	.80	--				
OPOP	.29	.38	.26	.69	.84	--			
WEPER	.18	.18	.22	.27	.30	.28	--		
WEJUD	.25	.28	.32	.33	.32	.31	.74	--	
WEOP	.18	.20	.27	.31	.31	.30	.70	.73	--

Note. RPPER = perception dimension in role play, RPJUD = judgment dimension in role play, RPOP = organizing and planning dimension in role play, OPPER = perception dimension in oral presentation, OPJUD = judgment dimension in oral, OPOP = organizing and planning dimension in oral presentation, WEPER = perception dimension in written exercise, WEJUD = judgment dimension in written exercise, WEOP = organizing and planning dimension in written exercise.