EXAMINATION OF THE ASSESSMENT CENTER CONSTRUCT-CRITERION RELATIONSHIP: SITUATIONAL SPECIFICTY AS IT RELATES TO BANDWIDTH-FIDELITY AND PREDICTING JOB PERFORMANCE

Andrew B. Speer

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts

Department of Psychology

Central Michigan University
Mount Pleasant, MI 48858
August, 2011

Accepted by the Faculty of the College of Graduate Studies,

Central Michigan University, in partial fulfillment of

the requirement for the master's degree

Thesis Committee:

Neil Christiansen, Ph.D.	Committee Chair
Matthew Prewett, Ph.D.	Faculty Member
Kevin Love, Ph.D.	Faculty Member
August 11, 2011	Date of Defense
Roger Coles, Ed.D.	Dean College of Graduate Studies
October 13, 2011	Approved by the College of Graduate Studies

ACKNOWLEDGEMENTS

I want to acknowledge my friends, family, and loved ones for their support during this endeavor called graduate school. After many long days of work they still stand by side. I also want to specifically thank my advisor, Neil Christiansen, for his guidance throughout the course of this project. Without his inspiration and insights this thesis would not have grown into what it is. Much appreciated to all who are part of my life.

ABSTRACT

EXAMINATION OF THE ASSESSMENT CENTER CONSTRUCT-CRITERION RELATIONSHIP: SITUATIONAL SPECIFICITY AS IT RELATES TO BANDWIDTH-FIDELITY AND PREDICTING JOB PERFORMANCE

by Andrew B. Speer

After years of research regarding the dimension consistency and exercise effects commonly seen within Assessment Centers (AC, e.g. Lievens, 2008), there have still been few attempts made to connect AC construct theory to AC criterion theory. The current study tested why, despite poor dimension consistency across exercises, Assessment Center ratings still predict managerial performance. Assuming psychological differences in situational demands affect both the consistency of behavioral ratings and the bandwidth of assessment, it was predicted that pairs of exercises with very different demands would result in less consistent ratings but have higher predictive validities. Likewise, the ensuing inconsistency was thought to be reflective of a more comprehensive assessment of individual differences.

Archival data from two operational Assessment Centers were used, with the first sample consisting of 342 managers who underwent four exercises and the second sample consisting of 99 managers who underwent five exercises. Exercise characteristics were measured using Trait Activation Potential (TAP) ratings. Using the TAP ratings, pairs of exercises were ordered according to similarity and composites of AC ratings for these pairings were then formed. Optimally inconsistent and consistent composites of AC ratings were also created by examining post-exercise dimension-ratings. These sets of composites were then examined against supervisory ratings of job performance.

As expected and in both samples, behavioral consistency was highest across pairs of similar exercises. When exercise similarity decreased, so did dimension convergence. In line with the hypotheses, it was found that composites of two psychologically dissimilar exercises resulted in higher predictive validities than did pairs of two similar exercises. This was the case in both samples. Thus, exercises designed to reflect a variety of demands should act as better selection devices. This should occur even though they result in less consistent ratings. Lastly, when composites of optimally consistent AC ratings and optimally inconsistent AC ratings were examined with job performance, it was found that inconsistency did not at all hinder predictive validity, and in fact may have enhanced it in Sample 1. Overall, findings support a "situational bandwidth" approach to maximize predictive validity in Assessment Centers.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Despite the long history of Assessment Center (AC) construct research, the specific nature of AC constructs has not fully been explored in regards to how they relate to criterion-related validity. Assessment Centers are clearly supported by high content and high criterion validities, and yet disagreement exists over what it is Assessment Centers are actually measuring (see Jackson, Stillman & Atkins, 2005; Lance, Foster, Nemeth, Gentry & Drollinger, 2007). In recent years research regarding the internal structure of AC ratings has helped shed light on some issues. It is now apparent that exercise effects are not composed exclusively of error, they represent actual differences in candidate performance across situations, and they explain more variance in candidate ratings than do dimensions (e.g. Lance, Newbolt, Gatewood, Foster, French & Smith, 2000; Bowler & Woehr, 2006). The coupling of poor dimension consistency with the predominance of exercise effects has led some to challenge the continued use of Assessment Center dimensions at all, instead favoring the adoption of task based rating systems (e.g. Lowry, 1997; Jackson et al., 2005; Jackson, Barney, Stillman, & Kirkley, 2007).

However, in the midst of all this research few have examined how Assessment Center construct validity actually relates to the predictive validity of job performance. The issue of inconsistency –displayed in the form of low monotrait-heteroexercise correlations – needs to be examined simultaneously with job criteria. If across exercise dimension consistency truly is required for construct validity, then shouldn't those consistent and assumed to be, more reliable measures, better predict job performance?

Most contemporary Assessment Center theory actually explains this behavioral inconsistency as a result of situational demands and not as error (e.g. Neidig & Neidig, 1984; Lance, Foster, Gentry & Thoresen, 2004). Less explicit though, is how situational specificity relates to the predictive nature of AC ratings. Drawing from Assessment Center and personality psychology literatures, the more diverse the situational demands the greater the variety of behavioral expression (e.g. Bem & Funder, 1978; Funder, 2006). If behavior is in part caused by situational presses (Murray, 1938), then trait measurement can be maximized by assessing behavior in the most varied of contexts and under the most varied of demands (Mischel & Peake, 1982).

Turning back to Assessment Centers, it seems likely that the ability to meet a multitude of assessment requirements should correspond with the ability to meet a multitude of job demands. In this sense, a selection battery composed of very unique exercises would assess a broader range of behaviors and capture a more complete understanding of individual performance tendencies. At the same time though, an increase in exercise uniqueness should result in less consistent dimension ratings across participants (see Highhouse & Harris, 1993; Haaland & Christiansen, 2002) as behaviors should vary with situational differences. If the purpose is to assess a given trait or dimension, this behavioral inconsistency would then constitute as unreliable measurement according to psychometric theory, thus limiting the potential to predict other meaningful criteria (Cronbach, 1960). One is then left to determine whether this inconsistency hinders the predictive nature of an Assessment Center, or whether it is the inconsistency itself that reflects a situationally broad assessment of individual tendencies.

The current study tested why, despite poor dimension consistency across exercises, Assessment Center ratings still predict managerial performance. This relationship between construct and criterion validity was examined in two ways. The first considers one of the potential causes of the low monotrait heteromethod correlations - the effect of situational demands. It is assumed that similar situations require the same set of behavioral requirements, thus resulting in consistent ratings between exercises. Likewise, less similar exercises require a more varied set of behavioral competencies and should result in worse consistency but greater coverage of trait information. This method involves measuring the contextual demands as a reflection of assessment bandwidth and then comparing how this predicts job performance.

While the above method assumes that situational similarity causes consistency in exercise performance, an inspection of AC ratings measures this directly. Actual behavioral consistency at the group level can thus be examined by looking at which exercises each dimension was rated most similarly and in which situations it was most inconsistent. By doing this, the inconsistency that signifies the construct paradox can be examined directly with job performance.

Both of these approaches get at the same question through different methods. The first examines the cause of behavioral consistency as a reflection of assessment bandwidth, while the second directly assesses the consistency of ratings. The question then becomes - does poor dimension convergence across exercises represent measurement error and thus reduce the ability to predict job related variance? Or, does this inconsistency actually reflect a more broad assessment of candidate tendencies and therefore relate better to a wide-ranging criteria such as job performance?

This study differs from previous research in several ways. While the endeavor adds to an already large body of literature, it will do so by assessing the intricacies of Assessment Center constructs and how they are predictive measures. Very few studies have explicitly examined construct and criterion validity simultaneously, with those that have extending Assessment Center literature by connecting AC constructs with various external criteria. These constructs have included final dimension scores, exercise factors, latent general performance factors, salient dimension factors and so forth (e.g. Lance et al., 2000; Arthur, Day, McNelly & Edens, 2003; Lance et al., 2004; Lance et al., 2007). However, no attempt has been made to examine the issue of dimension consistency itself in lieu with job performance at this level. Doing so is a direct look at the construct paradox and how despite dimension inconsistency, Assessment Centers still predict job criteria. An inspection of this issue will lead to a greater understanding of how AC constructs operate and specifically whether dimensions can maintain utility in the face of poor construct validity. In turn, this information could greatly aid in Assessment Center design. Further, this study could provide additional support for why an MTMM approach may be inappropriate in establishing Assessment Center construct validity.

CHAPTER II

LITERATURE REVIEW

Assessment Center Methodology

In order to understand how constructs and predictive validity relate, one must first know how Assessments Centers work. Assessment Centers (AC's) are standardized evaluation procedures used for organizational selection, diagnosis and development (Take Force on Assessment Center Guidelines, 2009; Thornton III & Gibbons, 2009). Taking multiple observations of behavior in work-like settings, AC's allow raters to make inferences of individual differences and traits, thus providing relevant information for both selection and developmental purposes. As such, Assessment Centers have been utilized the past 50 years for managerial assessment (Thornton III & Gibbons, 2009; Spychalski, Quinones, Gaugler & Pohley, 1997).

In general, the AC procedure involves candidates undergoing one or more simulations that are designed to replicate real work settings and elicit work behaviors (Task Force, 2009). Also known as exercises, these simulations include but are not limited to in-baskets, leaderless group discussions, assigned role plays, interviews, problem analyses, and presentations (Spychalski et al., 1997). Exercises are considered to be representative of the job position in question and are designed from job analysis (Task Force, 2009). Within each exercise candidates are typically rated on several behavioral categories that represent stable individual differences and are known as dimensions. It is important to note that some dimensions may be rated in one exercise while not evaluated in another. Also, from Assessment Center to Assessment Center the name and number of dimensions greatly varies (see Arthur et al., 2003).

Assessment Centers require multiple trained assessors to rate and record candidate behaviors systematically (Task Force, 2009). Typically, this is done by rating a given candidate in each exercise on multiple behavioral dimensions. These are known as post-exercise-dimension-ratings (PEDR). These dimension ratings are then commonly averaged across assessors and across exercises so that for every dimension there is one single score (Gibbons & Rupp, 2009). To consolidate further, practitioners can also combine these dimension ratings into a single AC performance composite know as an Overall Assessment Rating (OAR).

The continued use of Assessment Centers is facilitated by strong evidence of content and criterion-related validity (e.g. Lance et al., 2007). When carefully designed through job analysis, AC's show good content validity and high job relatedness (Sackett, 1987). This is because exercises are designed to resemble actual work tasks and the behavioral dimensions to reflect competencies underlying successful job performance (Neidig & Neidig, 1984). Assessment Centers are therefore face valid simulations of actual jobs.

Assessment Centers are highly predictive selection devices, typically correlating between .37 and .45 with job performance measures (e.g. Gaugler, Rosenthal, Thornton III, & Bentson, 1987; Arthur et al., 2003; Meriac, Hoffman, Woehr & Fleisher, 2008). They further offer unique incremental variance over other predictors such as cognitive ability tests and personality inventories, and even single dimensions positively relate to job criteria (e.g. Arthur et al., 2003; Meriac, et al., 2008). For example, Arthur et al. (2003) found validity coefficients for individual dimensions ranging from .25 to .39 with their taxonomy of dimensions, while Meriac and colleagues (2008) demonstrated that the dimensions offer unique variance over cognitive ability and FFM personality traits. Taken together these findings provide support for the continued use

of Assessment Centers for selection. Yet despite these positives, Assessments Centers have continuously shown poor construct validity. In short, the internal structure of AC ratings does not support the dimension constructs they purportedly measure (Lance et al., 2007).

The Construct Paradox

Assessment Centers may resemble jobs and predict performance well, but it is unclear exactly what it is they are measuring (Lance et al., 2000). This concept was sparked by Sackett and Dreher (1982) when they discovered that the same dimensions measured across different exercises had lower correlations than different dimensions within exercises. According to Campbell and Fiske (1959), construct validity is established through demonstrating convergent and discriminate validity. Known as the multitrait-multimethod (MTMM) approach, here AC exercises are treated as specific measures and the dimensions as traits. Like items on a test, each exercise takes estimations of candidates' levels on given dimensions. Under this rationale, a dimension rating in one exercise should correlate highly with that dimension's ratings in other exercises because they are observations of the same trait. However, this is not the case.

What Sackett and Dreher (1982) found was that AC ratings clustered within exercises and lacked convergence across exercises. In more technical terms, the heterotrait-monomethod correlations were larger than the monotrait-heteromethod correlations. This result is not an isolated occurrence. Time and time again others replicated Sackett and Dreher's (1982) findings that dimensions show poor consistency from exercise to exercise and different dimensions are rated similarly in the same exercises (e.g. Bycio, Alvares & Hahn, 1987; Lance et al., 2000; Jackson et al., 2005; Bowler & Woehr, 2006). As these results became accepted as commonplace, attention became increasingly aimed at explaining why these inconsistencies

occur. Initially, the explanation was that poor dimension structure was caused by method variance (Sackett & Dreher, 1982). Here, assessment is contaminated by measurement error. Some possible reasons for this error include: poor Assessment Center design, rater error and inadequate assessor training (e.g. Sackett & Dreher, 1982; Gaugler & Thornton, 1989; Reilly, Henry & Smither, 1990; Lievens, 1998; Lievens, 2008). However, it has been found that although improving these factors does lead to gains in construct validity, exercise effects still dominate (e.g. Lievens, 2001a; Kolk, Born, & van der Dlier, 2002; Lievens, 2002; Lievens, 2008).

Another explanation for exercise effects is that they are not representative of measurement error but instead reflect the situational specificity of behavior (e.g. Neidig & Neidig, 1984; Lance et al., 2000). This view takes exercises not as methods but as work samples, and considers differences in performance from one exercise to another as representing true performance variance across situations. With each exercise contributing specific job-like demands, performance within an exercise is then dependent upon how well one meets each set of situational requirements. As each exercise imposes unique psychological presses upon a participant (Bycio et al. 1987; Schneider & Schmitt, 1992), individual consistency from exercise to exercise is then not expected (Gibbons & Rupp, 2009). Thus, the small same-dimension different-exercise correlations and the strong inter-correlations between AC dimensions within an exercise may not be so problematic (Neidig & Neidig, 1984).

This notion was effectively tested by Lance and colleagues (2000 & 2004), who showed that latent exercise factors were not solely due to measurement error as some had previously thought. If exercise factors are only unwanted variance then they should have zero correlations

with meaningful criteria (Lance et al. 2000). However, Lance et al. (2000 & 2004) showed that exercise factors were positively related to external job criteria such as performance and job knowledge. Thus, exercise effects demonstrate meaningful rating variance.

While it was found that exercise effects are not completely composed of unwanted variance, it was still unclear how to view them. Three important studies were conducted to measure the direct influence of dimensions and exercises on Assessment Center ratings (Lievens & Conway, 2001; Lance, Lambert, Gwin, Lievens & Conway, 2004; Bowler & Woehr, 2006). These endeavors examined numerous MTMM matrices to determine the internal structure of Assessment Center ratings. In short, it was discovered that exercise factors account for a large proportion of rating variance but the amount was different for every study. Using a correlated dimension-correlated uniqueness approach (CDCU), Lievens and Conway (2001) found exercise and dimension factors to contribute equally, with each accounting for 34% of the rating variance. Following this, Lance and colleagues (2004) noted several problems of the correlated uniqueness method when MTMM data is tested. Most notably, this method assumes that exercise factors do not correlate. If this assumption is not met, dimension loadings can be upwardly biased. In a reexamination of the dataset from Leivens and Conway (2001), Lance et al. (2004) found that exercise factors clearly dominated dimension factors. However, perhaps the best estimate was by Bowler and Woehr (2006), who by using meta-analytical techniques were able to combine 34 MTMM matrices. Their findings revealed that exercise factors account for 34% of rating variance, while dimension factors account for 22%.

These findings, although showing that exercise factors account for the highest proportion of rating variance, also demonstrate a split in Assessment Center theory. Mainly, in light of

strong exercise effects do dimension factors maintain utility, and if so, can they still demonstrate some sort of construct validity? Some have suggested abandoning traditional Assessment Center design all together by removing dimensions in favor of task-based rating systems (Lowry, 1995; Jackson et al., 2005; Jackson et al., 2007). Although Jackson et al. (2005) demonstrated that for the most part task-based and dimension-based rating schemes do yield similar psychometric qualities, the fact still remains that current rating schemes are doing a good job of predicting performance. Because of this, it may be important to further consider how dimension constructs are related to job criteria before deciding to throw them out completely.

Situational Behavior

As years of Social Psychology have revealed, behavior is a function of two things: individual traits and situational demands (Funder, 2006). In Assessment Center terminology these are dimensions and exercises (Lievens, 2002). Quite simply, behavior occurs as traits and contextual demands interact. For instance, although a manager may be caring and supportive to his coworkers most of the time, when faced with a rebellious subordinate that same manager may exhibit stern and forceful behaviors in response to the situation. Here, even though the person is typically a calm and agreeable individual, behavioral differentiation is required to meet situational demands and successfully perform.

What should be taken from this is that individual differences do exist (e.g. Costa & McCrae, 1992) and they are contextually dependent (e.g. Mischel, 1968). As Wright and Mischel (1987) explain, traits can be seen as conditional probabilities occurring under certain stimuli. For instance, they showed that children's aggression varied considerably across different situations, yet could be predicted when situational demands were taken into account.

While situational characteristics can include the obvious concrete and physical descriptors such as number of people and temperature of a room, the actual causes of behavior are the psychological perceptions of these stimuli (Saucier, Bel-Bahar & Fernandez, 2006; Funder, 2009). This is a person's perception of situational characteristics and is similar to what Murray (1938) described as beta press. The effect of situational demands should not be taken lightly. Sparking much controversy, Mischel (1968) estimated a behavioral consistency of only .30 from situation to situation. Even when the demands of situations are highly similar, consistency correlations typically only reach .50, thus demonstrating the magnitude of situational effects on behavior (Mischel, 1985). However, this isn't saying that individual tendencies do not play a role in behavioral expression. To the contrary, "even though situations profoundly affect what people do, people can still manage to preserve their distinctive behavioral styles across situations" (Funder & Colvin, 1991, pp. 791).

The point is that both the person and the situation influence behavior. In terms of selection, while Wright and Mischel's (1987) viewpoint helps demonstrate how trait expression is largely an interaction between traits and situational demands, it would be of little use to limit behavioral predictions to only specific contexts. If the goal is to obtain a broad estimate of a trait, or one that can be applied to many situations, then that trait must be measured under a wide variety of contexts (Funder, 2006).

Assessment Centers were designed in such a way. Each exercise presents its own set of demands, and only candidates who possess the appropriate trait levels to meet situational requirements will perform efficiently across all exercises (Gibbons & Rupp, 2009). For example, a person may possess adequate leadership ability for one-on-one meetings and thus do well in a

subordinate role play. However, that same person may have inadequate group leadership skills and perform poorly in unstructured group discussions (Hoeft & Schuler, 2001). As situational characteristics become drastically different, the behavioral requirements placed upon candidates expand (Gibbons & Rupp, 2009). This should in turn accentuate individual differences more so than if behavior was measured in only one class of situations (Hoeft & Schuler, 2001), as not everyone possesses the skill levels to meet the demands of every situation. Here, behavioral inconsistency is actually required for performance consistency, as different sets of behaviors equate to effective performance depending upon situational requirements (Gibbons & Rupp, 2009).

This notion has been demonstrated in Assessment Centers. For instance, Kuptsch, Kleinmann and Köller (1998) found that individuals who rated themselves as high self monitors and therefore as behaving in socially desirable manners received similar dimension scores across exercises, even though those exercises required different behaviors. Here, high self monitors adapted their strategy according to the situational demands, and only because they did this were they able to score consistency. On the other hand, individuals who cannot adapt their behavioral styles across exercises with very different demands are likely to receive inconsistent ratings. Thus, the reliance on a simple MTMM matrix would then be a misleading establishment of construct validity as performance consistency is not expected for all individuals.

Although observation in different types of exercises increases the range of observed behaviors, this increase in assessment information occurs at the expense of performance consistency. This effect of situational features on Assessment Center behavior has been demonstrated by Highhouse and Harris (1993), Haaland and Christiansen (2002), and Lievens,

Chasteen, Day & Christiansen (2006). For example, Highhouse and Harris (1993) defined AC exercise characteristics by measuring the likelihood of occurring behaviors using the Behavioral Q-Sort method (Bem & Funder, 1978). They found that candidates behaved more consistently in situations that were rated as more similar. In those exercises rated as different, monotrait-heteromethod convergence suffered. Haaland and Christiansen (2002) had comparable findings when exercises were described using Trait Activation Potential (TAP), a method of situational description formed by Tett and Burnett (2003). Basically, candidate ratings were more similar when situational demands were more alike and less consistent with demands were unique. Lievens et al. (2006) later expanded this research by applying TAP to numerous MTMM matrices.

These findings demonstrate a split between two perspectives. For one, a greater variety of candidate behaviors are elicited when exercise demands differ (Hoeft & Schuler, 2001). If one wished to observe trait expressions in response to many types of demands, then behaviors should be observed under psychologically varied situations. As a result, more information about candidate tendencies could be gained. On the other hand, the increase in situational specificity corresponds to a decrease in rating consistency, as behavior varies as a function of situational characteristics. According to psychometric theory, this would constitute as error of measurement and be characterized by low reliability (Cronbach, 1960). The two perspectives then work against each other, with one hinging itself upon observing the broadest set of candidate behaviors and the other upon measuring those behaviors with precision. In a way, this tradeoff between information coverage and information accuracy is reminiscent of bandwidth-fidelity theory.

Bandwidth Fidelity Tradeoffs Applied to Assessment Centers

Predictor-criteria standards established by Shannon and Weaver (1949) and later expanded upon by Cronbach (1960) distinguish between two separate predictor-criteria characteristics: bandwidth and fidelity. Cronbach (1960) defines bandwidth as the amount of information captured by a given measure. Also referred to as broad or wideband measures (Cronbach & Gleser, 1965), high bandwidth tests measure traits that encompass a wide variety of behaviors and thus cover a spectrum of complex trait information (Cronbach, 1960). For example, Extraversion is a broad personality trait that is largely related to social behaviors and is composed of many singular traits such as assertiveness and gregariousness (Costa & McCrae, 1992). The measure gives a holistic description of individual tendencies and thus captures a good amount of information.

Fidelity is a separate characteristic that refers to accuracy (Cronbach, 1960), or the quality of information (Shannon & Weaver, 1949). Hogan and Roberts (1996) use the analogy of a microscope for fidelity, as measures high in this characteristic focus on a smaller spectrum of trait information but measure it with great precision. In this sense high fidelity instruments adhere to the properties of psychometric theory (Cronbach & Gleser, 1965). Psychometric instruments are composed of highly correlated items. The measure as a whole is reliable, and error is limited.

One common understanding about fidelity and bandwidth is that while they are on separate continuums, they are inversely related to each other (Shannon & Weaver, 1949; Cronbach, 1960; Hogan & Roberts, 1996). So, a shift in one results in an opposite shift in the other. The reason for this is simple. As the trait domain increases the behaviors that make up that

construct become more varied and thus more loosely correlated. This in turn makes the measure more unreliable (Cronbach, 1960). As reliability is a gauge of a test's accuracy of trait measurement, the wider the spectrum of assessed behaviors the worse the instrument's fidelity. Likewise, as the domain narrows to a smaller and more similar set of behavior, items will be more closely related. Therefore, high fidelity measurements have high reliabilities but cover less information.

A general approach to test selection is that if one wishes to predict a broad set of behaviors, or a wideband criteria, then a predictor with greater trait width should be used (Cronbach, 1960). In contrast, when a criterion is narrow the chosen predictors should also be. The idea is that the characteristics of the predictor ought to be driven by characteristics of the criteria (Hogan & Roberts, 1996), and this includes not only the bandwidth of measures but also the content (Tett, Guterman, Bleier, & Murphy, 2000). Neither approach is necessarily better across all contexts, calling for circumstantial characteristics to be taken into account (Cronbach & Gleser, 1965).

If Assessment Center ratings are considered an instrument, then according to this theory for dimension fidelity to be high the ratings would have to demonstrate consistency. When each exercise is simply considered a measure of given traits, then higher consistency would coincide with a greater proportion of error free variance in measurement and so would have a greater potential to be related to other variables such as job performance. When the dimensions demonstrate inconsistency, under a psychometric approach prediction would suffer due to a greater degree of error in the measurement.

On the other hand, some support can be made for having high bandwidth measures in Assessment Centers. For example, Arthur et al. (2003) suggests that using dimensions with lower inter-correlations results in better criterion validity. They based their conclusion on multiple regression analyses (forward and backward) where dimensions that had a high overlap with others eventually offered no unique variance. This seems common sense, as the more variance two predictors share the less unique variance can be left for other variables. If Assessment Center exercise ratings were treated as single predictors, would the more unrelated exercises assess a broader range of job relevant information then?

Increased bandwidth occurs when exercise demands differ, as trait tendencies are exposed in the face of numerous situational stimuli. Observations then lead to better generalizations of performance tendencies across situations, which help uncover important inter-individual differences during the selection process. Specifically, it gives a better estimate of cross-situational performance, as not everyone possesses adequate trait configurations to successfully perform in each exercise (Gibbons & Rupp, 2009). So, assessment bandwidth increases as contextual demands vary, albeit at the expense of consistency, or fidelity. The question then becomes which is more important?

Job Performance and Assessment Centers

The tug-pull between inconsistent measurement and coverage of measurement should be directly related to how well an instrument predicts certain criteria. In terms of Assessment Center selection, that criterion is job performance. Job performance can be defined as the "observable things people do that are relevant for the goals of the organization" (Cascio & Aguinis, 2005 as quoted from Campbell et al., 1990, pp. 60). Like nearly all selection tests, Assessment Centers

typically use a single, unidimensional measure of job performance for validation. As a whole though, this construct is very complex (Tett et al., 2000; Campbell, 1990). While a general performance factor may be parsimonious and easy to manage (Tett et al., 2000), job performance is clearly a multidimensional construct (e.g. Campbell, 1990; Wise, McHenry & Campbell, 1990; Tett et al., 2000). Campbell (1990) put it nicely by saying "It is axiomatic that job performance is not one thing. A job, any job, is a very complex activity; and, for any job, there are a number of major performance components." (Campbell, 1990, pp. 704).

The domain of managerial performance is an especially varied concept (Tett et al., 2000) and is more relevant to Assessment Centers, as AC's are often used to select managers. At a broad level performance dimensions can be broken into factors that represent basic managerial behaviors. For instance, managerial behaviors include decision making and organizing and planning (Tett et al., 2000). Others have included organizational citizenship behaviors as part of the performance domain (e.g. Scullen, Mount, & Judge, 2003; Borman & Motowidlo, 1997). Bartram (2005) found support for an eight factor model that covers all jobs broadly. At more specified levels the competencies and behaviors required for managerial performance can be broken down even further, as Tett et al. (2000) combined 12 taxonomies of performance into 53 competencies that cover this broad area.

These models demonstrate the complexity of job performance and highlight the need to adequately assess all of its functions when selecting predictors (Tett et al., 2000). More broad, or numerous narrow predictors are essential for the prediction of job performance at this encompassing level (e.g. Wise et al., 1990). Even the seemingly obvious characteristics of a job are represented by deeper complexities (Cascio, & Aguinis, 2005), and job behaviors as a whole

are situation specific (Tett et al., 2000). Because of the great breadth of the job performance domain, one characteristic that allows Assessment Centers predictive validity may then be their design to assess traits under varying conditions. The ability to meet a multitude of assessment requirements should correspond to the ability to meet a multitude of job demands. In other words, Assessment Centers may overcome potential measurement errors and behavioral inconsistencies by matching predictor bandwidth to criterion bandwidth.

The Present Study

The current study will test why, despite poor dimension consistency across exercises, Assessment Center ratings still predict managerial performance. Specifically, does an increase in situational specificity result in a broader assessment of individual tendencies, albeit at the expense of dimension consistency, and in turn allow for a better prediction of job performance? Or, are more consistent ratings less error-ridden and therefore more accurate predictors of performance?

There are several ways to test why Assessment Center ratings predict job performance. As discussed, the degree of situational similarity should affect the breadth of trait measurement and the behavioral consistency across exercises. When contextual demands are the same behavior should be consistent, and likewise when demands are unique behavior should be more situation specific and varied. Therefore, one way to gauge assessment bandwidth would be to explicitly measure the situational features of assessment center exercises. Despite behavioral inconsistency, measures taken in psychologically unique situations should assess a broader trait spectrum and thus capture more unique variance assuming that the content of the predictor matches the content of the criteria. Because Assessment Center exercises are created from job

analysis, this match should exist (Sackett, 1987). In opposition to the concept that behavioral inconsistency is error and under the aforementioned situational specificity/bandwidth hypothesis,

Hypothesis (1)

Exercises with varied psychological demands should capture a wider spectrum of candidate behaviors and lead to better predictions of job performance criteria than ratings from similar exercises.

Measuring situational characteristics allows for an estimate of one of the causes of behavioral variation (Saucier et al., 2007). However, actual Assessment Center behavior should directly be assessed to determine if situational consistency or specificity better translates to predictive validity. Inspecting AC ratings, which are based solely from candidate behaviors, gives such a measure. Specifically, the ratings for every dimension can be examined across exercises to establish behaviorally consistent and inconsistent pairings of PEDR's. Drawing on Hypothesis (1), behavior is still expected to be affected by situational characteristics so that situational specificity should increase trait bandwidth and thus capture more information relevant to incumbent behaviors. However, this measures actual dimension consistency from exercise to exercise.

Hypothesis (2)

Inconsistent dimension ratings reflect measures assessing a broader spectrum of candidate behaviors and so should lead to better predictions of job performance criteria than behaviorally consistent dimension ratings.

CHAPTER III

METHODOLOGY

Participants

Data from two separate Assessment Centers were used for this study. The Assessment Centers were held continuously from 2000 to 2004 and were conducted by the same consulting company. Over 30 organizations across the United States were run through the two assessments for the purposes of managerial development and promotion. Sample 1 consisted of 342 managers who underwent four exercises. Sample 2 was composed of 99 managers who underwent the same exercises as the first sample but also completed a presentation exercise.

Sample 1 was composed of primarily mid level managers from large companies, with over 80% having been managers for over five years. The managers came from a range of industries spanning from banking and insurance companies to heavy manufacturing. The average candidate age was 42 years old. 75% were male and 87% were Caucasian. 72% of the managers held at least a bachelors degree. Sample 2 possessed a very similar composition. All AC candidates were managers from a range of companies that covered numerous industries. The sample was primarily male (79%) and Caucasian (85%) with an average age of 41 years old. 82% had managed for five or more years.

Measures

Assessment Center Exercises

Sample 1 participated in four exercises: a behavioral interview, an in-basket, a group discussion, and a supervisory role-play. Sample 2 participated in the same four exercises but also completed a presentation exercise.

Assessment Center Dimensions

In both samples the same set of core behavioral dimensions were assessed. These dimensions were (1) Analyze Issues, (2) Sound Judgment, (3) Establish Plans, (4) Manage Execution, (5) Lead Courageously, (6) Influence Others, (7) Coaching & Development, (8) Fostering Teamwork, (9) Build Relationships, (10) Manage Disagreements, (11) Fostering Open Communication, and (12) Customer Focus. Not all dimensions were rated in every exercise. Table 1 shows which dimensions were rated in which exercises.

Assessment Center Ratings

For each exercise a single assessor made ratings for every relevant dimension following a candidate's completion of that exercise. Assessors rotated across exercises and had doctoral training in Industrial and Organizational Psychology and multiple years of experience conducting Assessment Centers.

Job Performance

Single item measures of global performance and potential for career advancement were used to create a composite measure of job performance for Sample 1. These measures were

completed by the candidate's supervisors around the time they completed the Assessment Center. The two single item measures were combined to represent overall job proficiency (r = .36). For Sample 2, as part of a developmental feedback process, performance ratings of the candidates were made by supervisors using the consulting company's proprietary multidimensional performance measure. This 20 item measure assessed a range of performance competencies. These competencies were summed to create an overall performance composite of managerial job performance, with the mean correlation between items being (.60) and the reliability of the composite of all 20 items estimated at (.97).

Table 1. *Assessment Center Dimensions and Exercises*

	Behavioral Interview	In-Basket	Supervisor Role Play	Group Discussion	Presentation
Analyze Issues	x	x	x	x	x
Sound Judgment	x	x	x	x	
Establish Plans	x	x			x
Manage Execution	x	x	x	x	
Lead Courageously	x	x	x	x	x
Influence Others	x	x	x	x	x
Coaching & Development	x	x	x		
Fostering Teamwork	x	x		x	
Build Relationships	x	x	x	x	
Manage Disagreements	x		x	x	
Open Communication	x	x	x	x	
Customer Focus	x	x		x	

Note: An "x" signifies that the dimension was rated in the given exercise.

Exercise Similarity

Characteristics of the Assessment Center exercises were assessed to test whether variation in situational demands increases assessment bandwidth and thus leads to better predictions of job performance. As discussed, it is important to capture information about the

shared psychological meaning of situations (Saucier et al., 2007). Trait Activation Potential (Tett & Burnett, 2003) is a method that indirectly gauges the psychological demands of a situation and has been used to describe AC exercises in the past (Haaland & Christiansen, 2002; Lievens et al., 2006). Therefore, it was chosen as the method of situation description for the current study.

Trait Activation Potential (TAP) assumes that behaviors are trait-related expressions aroused by situational stimuli (Tett & Guterman, 2000). TAP assesses the relevance of trait expression in a given situation and the strength of cues for expressions of that trait. By describing situations by the type and intensity of expected behavioral expressions, TAP assesses a situation in terms of important contextual cues and features. Like past research with this method, the current study assessed exercises using a Five Factor Model of personality (Costa & McCrae, 1992). Also included though was the measurement of Trait Activation Potential for cognitive ability as a way to gauge whether intellectual requirements differed between exercises.

Table 2. *Trait Activation Potential Ratings by Exercise*

	E	A	C	ES	O	G
Interview	21.00	16.75	15.17	17.00	21.75	21.33
In-Basket	9.92	16.17	26.50	10.58	17.50	24.50
Role Play	20.83	20.83	17.83	21.17	17.58	21.42
Group Discussion	23.83	22.67	18.42	22.25	19.92	21.75
Presentation	18.58	13.83	23.17	21.83	18.25	22.17

Note. Trait Activation Potential ratings were on a 1-5 scale, with 6 items measuring every dimension for each exercise. E = Extraversion, A = Agreeableness, C = Conscientiousness, ES = Emotional Stability, O = Openness to Experience, G = Cognitive Ability.

A TAP questionnaire was filled out by four Assessment Center administrators who were familiar with the Assessment Center exercises. The questionnaire involved rating the degree to which there were situational cues for each FFM trait and cognitive ability, how easily observable

expressions of those traits were, and if behavioral trait expressions were expected to lead to successful exercise performance. For each trait per exercise, six items were rated on five-point Likert scales. Table 2 displays the TAP ratings for each dimension in each exercise, and the Trait Activation Potential survey items can be seen in Appendix A. The estimated inter-rater reliability of a composite of the four raters averaged across all trait ratings and exercises was (.70).

Table 3. *Exercise Similarity Profile for In-Basket and Group Discussion Exercises*

Trait	TAP In-Basket	TAP Group Discussion	Absolute Difference
Extraversion	9.92	23.83	13.91
Agreeableness	16.17	22.67	6.50
Conscientiousness	26.50	18.42	8.08
Emotional Stability	10.58	22.25	11.67
Openness	17.50	19.92	2.42
Cognitive Ability	24.50	21.75	2.75
Profile Similarity			45.34

Note: TAP = Trait Activation Potential rating, which are based on 6 items per trait per exercise. A TAP score for an exercise can range from 6-30.

Using the TAP ratings, exercises could be described according to situational similarity and uniqueness. Funder & Colvin (1997) discuss the process of creating situational profiles. Here, by examining the absolute mean agreement between a set of items different situations can be compared. If the trait ratings (e.g. Extraversion TAP, Agreeableness TAP) are treated as items, then absolute differences can be taken for each paired trait rating for each set of exercises. Then, these values are summed for a comparison of exercise similarity. Smaller values denote similar exercises and larger values exercises with greater cue variance. An example of this process can be seen in Tables 3 and 4. Table 3 represents two exercises that have very different

demands, while Table 4 contains exercises that are more similar in terms of demands and situational cues.

Table 4. *Exercise Similarity Profile for Role Play and Group Discussion Exercises*

Trait	TAP Role Play	TAP Group Discussion	Absolute Difference
Extraversion	20.83	23.83	3.00
Agreeableness	20.83	22.67	1.84
Conscientiousness	17.83	18.42	.59
Emotional Stability	21.17	22.25	1.08
Openness	17.58	19.92	2.34
Cognitive Ability	21.42	21.75	.33
Profile Similarity			9.18

Note: TAP = Trait Activation Potential rating, which are based on 6 items per trait per exercise. A TAP score for an exercise can range from 6-30.

Using the absolute deviation scores, the exercise pairings were ordered by exercise similarity. Thus, the ratings from the two most similar exercises were combined to form similar exercise composite 1. The ratings from the next two most similar exercises were combined to form similar exercise composite 2. Likewise, the ratings from the two most dissimilar exercises were combined to form dissimilar exercise composite 1 and so forth. From this a rank ordering of exercise combinations was made according to situational similarity and can be seen in Table 5.

Behavioral Consistency

While the above method assesses the situational features as a cause of behavior, an examination of actual Assessment Center ratings acts as an overt measure of consistency. Thus, this study explicitly examined performance consistency through the creation of two optimally consistent and inconsistent predictor composites. Here, a simple examination of the MTMM

matrix determined for each dimension, in what exercises that dimension was rated most similarly and in what exercises it was the most situation specific. In turn two predictor composites were made to represent consistent AC performance and inconsistent AC performance. By doing this every dimension was included in the two composites– that is, as long as it was rated in at least three exercises. This is so that there was a similar and dissimilar pair. In Sample 1 there were 11 dimensions that were measured across three or more exercises, and in Sample 2 there were 12.

Table 5. *Exercise Pairing Similarity Scores*

Exercise Pairing	Absolute Difference of TAP Scores	Similarity Rank
RP - GD	9.18	1
INT - RP	15.34	2
RP - PR	16.67	3
INT - GD	19.50	4
GD - PR	21.35	5
INT - PR	22.51	6
IB - PR	26.67	7
INT - IB	36.84	8
IB - RP	37.99	9
IB - GD	45.34	10

Note: TAP = Trait Activation Potential rating, which are based on 6 items per trait per exercise and can range from 6-30. For each exercise pairing, the absolute differences were taken from TAP scores from matched dimensions and then summed for the overall Absolute Difference score. Scores for this can range from 0 to 144. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion, PR = Presentation.

Note that when examining correlations between PEDR's, some correlations were extremely close or even identical to each other for the same dimension. In response to this, an arbitrary value of .003 was set beforehand as the cutoff in determining whether two correlations were unique from one another. For instance, when examining which Oral Communication pairing to include in the consistent composite, if PEDR's for this dimension from the in-basket

and presentation correlated (.253), and the correlation from the role play and group discussion was (.258), then these combinations were treated as separate. Therefore, Oral Communication scores in the role play and group discussion exercises would be included in the consistent composite, while those for in-basket and presentation would not. However, if the difference between these correlations was less than .003 (say .253 and .255), then these values would be treated as similar. Here, because these are treated as the same and are the highest inter-correlations for the Oral Communication dimension, they would then both included in the consistent composite by weighting each pairing and summing.

In Sample 1, of a possible 54 same-dimension correlations from the 11 relevant dimensions assessed in at least three exercises, five (9%) were within the .003 value. However, only two (4%) of these were actually the most consistent or inconsistent pairing for their given dimensions and thus used in creation of the composites. In Sample 2, of a possible 69 same-dimension PEDR correlations for the 12 relevant dimensions assessed in at least three exercises, seven (10%) were within this value. Like Sample 1 though, only three (4%) of these met the criteria to be included into the consistent or inconsistent composites.

In Sample 1 the mean monotrait-heteromethod correlation in the consistent composite was .15. For the inconsistent composite it was .02. Thus, the two composites show a clear difference in terms of relative consistency. However, these same-dimension correlations are each extremely low. This causes reasons to worry about the generalizability of the current AC sample. Traditionally, monotrait-heteromethod correlations range around .25 (Bowler & Woehr, 2006). In the current sample the range of MTMM correlations was (-.03 to .26), and so even the highest value barely reached typical means. The same trend occurred in Sample 2. The mean monotrait-

heteromethod correlation in the consistent composite was .23, and for the inconsistent composite it was .12. These values are higher than Sample 1, but still well below the overall average usually seen. It may be important to point out that the heterotrait-monomethod correlations were also much smaller than typically reported ($\bar{r}$ = .30 in Sample 1, $\bar{r}$ = .37 in Sample 2). Thus, all ratings seem to be more independent of each other than is typical. So, although dimension consistency in the current study is lower than what is usually seen, within exercise effects are as well. These concerns will be addressed in more detail in the General Discussion. In terms of assessing the predictive differences of consistent and inconsistent composites, the true concern is whether or not the two composites differ in terms of consistency, which they do.

CHAPTER IV

SAMPLE 1 RESULTS

Dimension Convergence by Exercise Similarity

Before examining the hypothesized relationships, the Trait Activation Potential ratings were first examined to determine whether they accurately predicted behavioral consistency across exercises. As with previous research, exercises with more varied demands should result in less consistent dimension ratings (e.g. Highhouse and Harris, 1993). This relationship serves as the foundation for Hypotheses 1 and 2. Thus, the monotrait-heteromethod correlations were examined according to the exercise similarity scores.

Table 6. *Monotrait-Heteromethod Correlations by Exercise Similarity (Sample 1)*

Exercise Pairing	Exercise Similarity Rank based on TAP	Mean MTHM Correlation	Median MTHM Correlation
RP - GD	1	.13	.15
INT - RP	2	.15	.16
INT - GD	3	.09	.06
INT - IB	4	.09	.08
IB - RP	5	.08	.07
IB - GD	6	.07	.06
Average of 3 Similar Exercise Pairings		.12	
Average of 3 Dissimilar Exercise Pairings		.08	

Note: N = 342. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion. TAP = Trait Activation Potential. MTHM = monotrait-heteromethod.

Table 6 displays the six exercise pairings in Sample 1, their order according to exercise similarity, and the mean and median monotrait-heteromethod correlations between each exercise pairing. As can be seen, there is a clear pattern between convergence and exercise similarity. The three similar exercises had the highest monotrait-heteromethod correlations (mean = .12), and the

three dissimilar exercises had the lowest (mean = .08). To test whether this difference was significant, a non-parametric test of the medians was employed due to the non-normal distribution and the dependence of the observed correlations. A sign test was used where within each dimension, pairwise comparisons were made between monotrait-heteromethod correlations in similar exercises and monotrait-heteromethod correlations in dissimilar exercises. This procedure was used in a similar fashion in Haaland and Christiansen (2002). The sign test uses a binomial distribution, where if there is not a difference for dimension consistency in similar and dissimilar exercises, then an equal number of high values should be observed for each group. If a difference exists between the groups, then a greater number of the larger MTHM correlations should fall towards one direction.

To perform this test, monotrait-heteromethod correlations of similar exercises were compared to the corresponding same-dimension correlations from dissimilar exercises. Table 7 displays the individual dimension comparisons across similar and dissimilar exercises. As seen, the majority of dimensions maintained greater convergence across pairs of similar exercises. However, to test the *overall* convergence across similar and dissimilar exercise pairs, all of the paired monotrait-heteromethod comparisons were examined simultaneously, resulting in 66 pairwise[1] comparisons between similar and dissimilar exercises.

Doing this across all dimensions simultaneously results in a test of overall convergence and 66 pairwise comparisons (link to that footnote). Of these, 65% of the higher monotrait-heteromethod correlations occurred in similar exercises. Employing a chi-square

[1] *With sign test, if pairwise observation are of the same value then the pair is discarded from the analysis as neutral. Two monotrait-heteromethod correlations were treated as the same if they were within .003 of each other. Out of a possible 69 pair-wise comparisons between same-dimension correlations from similar and dissimilar exercises, three were within this range and were discarded from the sign test.*

Table 7. *Monotrait-Heteromethod Comparisons by Exercise Similarity (Sample 1)*

	Monotrait-Heteromethod Correlations						Monotrait-Heteromethod Comparison			
Dimension	RP-GD	INT-RP	INT-GD	INT-IB	IB-RP	IB-GD	# Pairwise Comparisons	% Greater Similar	% Greater Dissimilar	χ^2
Analyze	.14	.15	.22	.22	.10	.15	9_n	50	50	.00
Judgment	.17	.17	.14	.15	-.03	.06	9	89	11	5.44*
Execution	.09	.06	.04	.06	.03	.05	9_n	75	25	2.00
Lead Cour.	.21	.18	.20	.01	.08	.06	9	100	0	9.00**
Influence	.18	.26	.06	.10	.06	.11	9	67	33	1.00
Coaching	-	.03	-	-.02	.04	-	2_n	100	0	-
Teamwork	-	-	-.02	.12	-	.08	2	0	100	-
Build Rel.	.08	.16	.07	.12	.22	.04	9	44	56	.11
Manage Disa.	.16	.17	.05	-	-	-	0	-	-	-
Open Comm.	.04	.14	.01	.05	.13	.05	9	33	67	1.00
Est. Plans	-	-	-	-	-	-	-	-	-	-
Cust. Focus	-	-	.10	.03	-	.06	2	100	0	-

Note: N = 342. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion. # Pairwise Comparisons refers to the number of same-dimension comparisons across similar and dissimilar exercise pairs. This refers to the number of comparisons prior to removing any pairs that were of the same magnitude, of which there were three within .003. These are marked by the subscript "n" and were not used in computing the chi square statistic. Chi-square was computed using sign test at 1 df and one-tailed tests. Because tests with less than four pairwise comparisons could not possibly reach significance, chi-square statistics were not computed unless there were at least four comparisons for a given dimension. * $p < .05$. ** $p < .01$.

test with one degree of freedom indicated that similar exercises had significantly larger monotrait-heteromethod correlations (median = .14) than dissimilar exercises (median = .06, χ^2 = 6.06, $p < .05$). Thus, in line with previous research (Haaland & Christiansen, 2002; Highhouse and Harris, 1993; Lievens et al., 2006) convergence was higher when situational demands were more similar.

Validity by Exercise Similarity

Hypothesis 1 posited that dissimilar exercises would assess a wider spectrum of candidate behaviors and lead to better predictions of job performance than ratings from similar exercises. However, before combining exercises into composites and examining validities, the predictive nature of the individual overall exercise ratings (OER) were first examined. Table 8 displays the means, standard deviations and correlations for the OER's, the overall dimensions ratings (ODR's), the overall assessment center rating (OAR) and job performance. The correlations between job performance and the OER's for the interview (r = .16), in-basket (r = .16), and group discussion (r = .12) were all relatively similar. However, the role play's validity (r = .05) was noticeably lower than any other exercise, making any exercise pairing including it potentially disadvantaged. This concern will be elaborated upon further in the General Discussion. A combination of all OER's into an overall assessment rating (OAR) resulted in an overall validity of (r = .20) for the Assessment Center.

Table 8. *Descriptive Statistics and Correlations amongst OER's, ODR's, OAR and Job Performance (Sample 1)*

	Mean	1	2	3	4	5	6	7	8	9	10	11
1. Interview	3.26	(.23)										
2. In-Basket	2.82	.17	(.42)									
3. Role Play	2.83	.25	.10	(.37)								
4. Group Discussion	2.89	.14	.13	.21	(.43)							
5. Analyze Issues	3.13	.30	.45	.36	.55	(.36)						
6. Judgment	2.96	.40	.41	.48	.55	.70	(.32)					
7. Establish Plans	3.02	.39	.59	.11	.15	.36	.40	(.42)				
8. Manages Execution	2.73	.39	.39	.32	.38	.40	.57	.44	(.32)			
9. Leads Courageously	3.05	.36	.26	.39	.38	.31	.53	.23	.46	(.39)		
10. Influence	2.77	.39	.50	.49	.51	.54	.53	.33	.34	.38	(.34)	
11. Coaching	2.79	.35	.39	.45	.06	.10	.21	.26	.26	.21	.40	(.33)
12. Teamwork	3.01	.25	.53	.09	.43	.37	.30	.31	.19	.09	.44	.22
13. Building Relationships	3.09	.29	.34	.45	.28	.21	.17	.14	.07	-.02	.48	.42
14. Managing Disagreements	2.70	.32	.14	.51	.57	.38	.42	.08	.23	.34	.50	.26
15. Open Communication	3.09	.33	.19	.48	.47	.35	.31	.17	.14	.05	.46	.29
16. Customer Orientation	3.21	.29	.43	.17	.43	.33	.41	.30	.42	.45	.33	.16
17. OAR	2.98	.57	.63	.58	.64	.70	.76	.52	.61	.56	.78	.50
18. Job Performance	10.65	.16	.16	.05	.12	.20	.22	.14	.11	.16	.16	.05

Note: N = 342. OER = Overall Exercise Rating. ODR = Overall Dimension Rating. OAR = Overall Assessment Center Rating. Numbers in parentheses are standard deviations.

Table 8 *Continued…*

	12	13	14	15	16	17	18
1.							
2.							
3.							
4.							
5.							
6.							
7.							
8.							
9.							
10							
11.							
12.	(.37)						
13.	.36	(.37)					
14.	.32	.38	(.34)				
15.	.40	.61	.48	(.37)			
16.	.22	.16	.23	.15	(.35)		
17.	.57	.55	.62	.59	.56	(.22)	
18.	.03	.10	.10	.02	.09	.20	(1.59)

Table 9. *Exercise Composite Validity by Exercise Similarity (Sample 1)*

Exercise Pairing	Exercise Similarity Rank based on TAP	Exercise Composite Validity
RP - GD	1	.12
INT - RP	2	.13
INT - GD	3	.18
INT - IB	4	.20
IB - RP	5	.15
IB - GD	6	.19
Average of 3 Similar Exercise Pairings		.14
Average of 3 Dissimilar Exercise Pairings		.18

Note: N = 342. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion. TAP = Trait Activation Potential.

To test Hypothesis 1 the six paired exercise composites were arranged by similarity and correlated with job performance. Table 9 shows each paired exercise composite and their validities. As seen, the lowest validity (r = .12) belonged to the composite containing the two most similar exercises (role play – group discussion). The next most similar pairing (interview – role play) had the second lowest validity (r = .13). On the other hand, the composites with the two highest validities belonged to the most dissimilar pair of exercises (r = .19, in basket – group discussion) and the third most dissimilar pair (r = .20, interview – in basket).

When the six exercise pairings were split into three similar and three dissimilar, the mean validities were (r = .14) for similar pairings and (r = .18) for dissimilar pairings. Examination of pairwise dimension composite validities for the similar and dissimilar exercise pairings can be seen in Table 10. To test whether ratings from dissimilar exercises significantly

Table 10. *Dimension Validity Comparisons by Exercise Similarity (Sample 1)*

	Dimension Composite Validity						Validity Comparison			
	Similar Exercises			Dissimilar Exercises						
Dimension	RP-GD	INT-RP	INT-GD	INT-IB	IB-RP	IB-GD	# Pairwise Comparisons	% Greater Similar	% Greater Dissimilar	χ^2
Analyze	.13	.15	.16	.20	.16	.17	9_n	0	100	8.00**
Judgment	.18	.14	.18	.15	.16	.19	9	44	56	0.11
Execution	.06	.06	.06	.10	.10	.10	9	0	100	9.00**
Lead Cour.	.08	.11	.15	.18	.10	.14	9	33	67	1.00
Influence	.08	.06	.15	.18	.11	.18	9	11	89	5.44*
Coaching	-	.00	-	.08	.04	-	2	0	100	-
Teamwork	-	-	.08	.01	-	.00	2	100	0	-
Build Rel.	.03	.04	.08	.14	.08	.12	9_n	0	100	8.00**
Manage Dis.	.05	.09	.10	-	-	-	0	-	-	-
Open Comm.	.02	.00	.02	.08	.07	.09	9	0	100	9.00**
Est. Plans	-	-	-	.14	-	-	0	-	-	-
Cust. Focus	-	-	.05	.04	-	.13	2	50	50	-

Note: N = 342. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion. # Pairwise Comparisons refers to the number of same-dimension comparisons across similar and dissimilar exercise pairs. This refers to the number of comparisons prior to removing any pairs that were of the same magnitude, of which there were two within .003. These are marked by the subscript "n" and were not used in computing the chi square statistic. Chi-square was computed using sign test at 1 df and one-tailed tests. Because tests with less than four pairwise comparisons could not possibly reach significance, chi-square statistics were not computed unless there were at least four comparisons for a given dimension. * $p < .05$. ** $p < .01$.

predict job performance better than ratings from similar exercises, a sign tests was once again employed using a chi-square statistic at 1 degree of freedom. As before, this was used due to the non-normal distribution and non-independence of the measures.

Of the seven dimensions in which chi-square statistics could be calculated, *all* were in the hypothesized direction and five were statistically significant (see Table 10)[2]. Thus, individual dimensions more accurately predicted job performance when they were measured across situations with varying demands. In terms of examining the exercise pairings as a whole, testing all the pairwise dimension validities simultaneously resulted in a total of 67 comparisons[3], with 84% of them having higher validity coefficients in dissimilar exercise pairs. Overall, dissimilar exercises had higher dimension composite validities (median = .11) than similar exercises (median = .08, $\chi^2 = 30.22$, $p < .001$). Thus, support was found for Hypothesis 1. When predicting job performance, a composite of two psychologically dissimilar exercises will result in better prediction of job performance than two very similar exercises.

It is worthy to note that it is common in Assessment Centers for different exercises to measure different sets of dimensions, and in Sample 1 this is no different. Because some exercises contain more dimension measurements than others and this could potentially lead to a more comprehensive measurement, exercises with less dimension ratings may then be

---

[2] *When conducting sign tests at the dimension level, validities were only compared when a given dimension was assessed in both exercises. For example, in Table 10 there is no value for the coaching dimension in the RP-GD pairing, as it was measured in the role-play but not in the group discussion. Although this dimension measure was included when forming the actual exercise-exercise composites that were correlated with job performance (composites shown in Table 9), when comparing at the dimension level it was not analyzed because potential bias could occur if compared against dimension composites formed from two exercises.*
-A dimension needed at least four comparisons to be tested for significance using sign test.

[3] *With sign test, if pairwise observations are of the same value then the pair is discarded from the analysis as neutral. Two correlations were treated as the same if they were within .003 of each other. Out of a possible 69 pair-wise comparisons, two were within this range and were discarded from the sign test.*

disadvantaged in any validity analyses. To control for this, or to put all exercise pairings on an equal playing field, a separate set of analyses were conducted where only the dimensions that were shared across all exercises were used for the creation of the exercise composites. Therefore, each exercise measured the exact same set of dimensions. When this was done and the exercise pairs were correlated with job performance and arranged by similarity, the rank order of validities was identical to the results found when not controlling for the number of dimensions. Thus, the results from these analyses are not reported.

Validity by Behavioral Consistency

While it was shown that Assessment Center validity was highest when using two optimally dissimilar exercises, an examination of actual dimension consistency could help provide a better understanding of how exactly Assessment Centers predict performance. Hypothesis 2 posited that composites of inconsistent ratings would reflect a broader spectrum of candidate behaviors which in turn would result in better predictions of job performance. As already shown, convergence was lower in dissimilar exercises, and so one of the causes of inconsistency is established. However, it has yet to be seen whether this inconsistency actually results in better predictions of job performance.

As described, the PEDR's for each dimension were examined to create consistent and inconsistent AC dimension composites[4]. These individual dimension composites were then aggregated to create optimally consistent and inconsistent sets of Assessment Center ratings. Table 11 displays which PEDR's were used to create the composites for every dimension. The

[4] Because "Establish Plans" was only assessed in two exercises in Sample 1, it was not included in either the consistent or inconsistent composites.

mean monotrait-heteromethod correlation for the consistent composite was .15 (ranging from .04 to .26), while for the inconsistent composite it was only .02 (ranging from -.03 to .10).

Table 11. *Validities of Consistent and Inconsistent PEDR pairings (Sample 1)*

Dimension	Consistent Pairing	Consistent Validity	Inconsistent Pairing	Inconsistent Validity
Analyze	INT-IB	.20	IB-RP	.16
Judgment	RP-GD, & INT-RP	.19	IB-RP	.16
Execution	RP-GD	.06	INT-RP	.06
Lead Cour.	RP-GD	.08	INT-IB	.18
Influence	INT-RP	.06	INT-GD	.15
Coaching	IB-RP, & INT-RP	.03	INT-IB	.08
Teamwork	INT-IB	.01	INT-GD	.09
Build Rel.	IB-RP	.08	INT-GD	.12
Manage Dis.	INT-RP	.10	INT-GD	.10
Open Comm.	INT-RP	.00	INT-GD	.02
Customer Focus	INT-GD	.05	INT-IB	.04
Composite		.15		.19

Note: N = 342. INT = Interview, IB = In-basket, RP = Role-Play, GD = Group Discussion.

To begin, composites of the individual dimensions were correlated with job performance. As seen in Table 11, inconsistent composites had higher validities for 7 out of the 10 allowed dimension comparisons[5]. However, using the same non-parametric median test previously employed, the dimension validities of the inconsistent pairings were not significantly greater (median = .10) than those of the consistent pairings (median = .06, χ^2 = 1.60, p = ns). This is not surprising, given there were only a total 10 pairwise comparisons. To reach significance with this amount, at least nine would have to favor one side of the median. Thus, sampling error cannot be ruled out.

[5] *Note: If pairwise observations are of the same value then the pair is discarded from the analysis as neutral. Two correlations were treated as the same if they were within .003 of each other. Out of a possible 11 pair-wise comparisons, the two validities for Managing Execution were within the .003 range and thus treated as neutral.*

Next, the overall composites were correlated with job performance. The inconsistent composite was composed of the inconsistent pairings for all dimensions, while the consistent composite was composed of the consistent pairings for all dimensions. As seen in Table 11, the inconsistent composite had a greater relationship with job performance (r = .19) than the consistent composite (r = .15). It is worthy to notice that the validity of the inconsistent composite was very similar to the OAR's (r = .20).

Because of the dual dependence of the variables, a Hotelling-Williams t-test was used to examine whether the overall inconsistent composite was a significantly better predictor. The Hotelling-Williams t-test takes into account the correlation between predictors and the within person nature of the analysis. It was found that the inconsistent composite did not significantly predict job performance better than the consistent composite (t = 1.03, df = 339, p = ns). Therefore, while a composite of inconsistent Assessment Center ratings had a higher correlation with job performance, the possibility that this difference was due to sampling error cannot be ruled out. Thus, limited support was found for Hypothesis 2.

CHAPTER V

SAMPLE 1 DISCUSSION

In line with research from social psychology (e.g. Wright & Mischel, 1987) and previous Assessment Center findings (e.g. Highhouse & Harris, 1993), behavioral ratings were most consistent across two similar situations. This relationship formed the foundation for Hypothesis 1. Specifically, two dissimilar exercises were expected to predict job performance better than two very similar exercises because a wider range of behaviors will be observed in psychologically unique situations. Solid support was found here, as composites of AC ratings from two dissimilar exercises had higher criterion validities. While exercise characteristics were theorized to affect behavioral consistency, Hypothesis 2 examined the direct result of this inconsistency as it related to predicting job performance. Although inconsistent ratings in general had higher predictive validities than consistent ratings, the difference was small and non-significant. To further test the generalizability of these results, Sample 2 was looked at according to the same methods and hypotheses.

CHAPTER VI

SAMPLE 2 RESULTS

Dimension Convergence by Exercise Similarity

As with Sample 1, the monotrait-heteromethod correlations were examined to determine whether they were larger across similar exercises. Table 12 displays the ten exercise pairings from Sample 2, their order according to exercise similarity, and the mean and median monotrait-heteromethod correlations between each exercise pairing. As with Sample 1, there was a strong pattern between dimension consistency and exercise similarity. The three most similar exercise parings had a mean monotrait-heteromethod correlation of .20. This was noticeably higher than the three most dissimilar pairings, which showed a mean convergence of only .09. When considering the possible 10 exercise pairings as five similar and five dissimilar, the differences still remained prominent, as the mean monotrait-heteromethod correlations were .21 for similar and .11 for dissimilar pairings.

Individual dimensions were examined according to their cross-exercise convergence and can be seen in Table 13. Of the seven dimensions for which chi-square statistics could be calculated, six had higher monotrait-heteromethod correlations across the similar exercise pairs and five of these differences were statistically significant (Table 13). Overall, across the five similar and five dissimilar exercise pairings there were a total of 111 pairwise monotrait-heteromethod comparisons[6]. Using these pairings, a sign test of the median correlations was once again employed to test whether monotrait-heteromethod correlations were significantly different

[6] Note: *With sign test, if pairwise observations are of the same value then the pair is discarded from the analysis as neutral. Two correlations were treated as the same if they were within .003 of each other. Of a possible 117 pairwise comparisons, six were within a value .003 and so were treated as neutral.*

for pairs of similar and dissimilar exercises. Of these, similar exercises (median = .19) had higher dimension consistency than dissimilar exercises 82% of the time (median = .11, χ^2 = 45.41, p < .001). Thus, both samples showed a positive trend between exercise similarity and behavioral convergence. Assessment Center ratings demonstrated better convergence across psychologically similar exercises.

Table 12. *Monotrait-Heteromethod Correlations by Exercise Similarity (Sample 2)*

Exercise Pairing	Exercise Similarity Rank based on TAP	Mean MTHM Correlation	Median MTHM Correlation
RP - GD	1	.20	.22
INT - RP	2	.16	.15
RP - PR	3	.24	.22
INT - GD	4	.14	.14
GD - PR	5	.30	.31
INT - PR	6	.14	.13
IB - PR	7	.14	.13
INT - IB	8	.10	.13
IB - RP	9	.06	.04
IB - GD	10	.09	.11
Average of 3 Similar Exercise Pairings		.20	
Average of 3 Dissimilar Exercise Pairings		.09	

Note: N = 99. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion, PR = Presentation. TAP = Trait Activation Potential. MTHM = monotrait-heteromethod

Table 13. *Monotrait-Heteromethod Comparisons by Exercise Similarity (Sample 2)*

	Monotrait-Heteromethod Correlations										Monotrait-Heteromethod Comparison			
	Similar Exercises					Dissimilar Exercises								
Dimension	RP-GD	INT-RP	RP-PR	INT-GD	GD-PR	INT-PR	IB-PR	INT-IB	IB-RP	IB-GD	# Pairwise Comp.	% Greater Sim.	% Greater Dis.	χ^2
Analyze	.24	.15	.22	.26	.31	.24	.11	.00	.00	.15	25_{nn}	91	9	15.70**
Judgment	.25	.27	-	.17	-	-	-	.16	.07	.13	9	100	0	9.00**
Execution	.10	-.09	-	.14	-	-	-	.16	-.06	.19	9	22	78	2.77
Lead Cour.	.37	.13	.21	.15	.33	.10	.02	.01	.00	.13	25_{n}	100	0	24.00**
Influence	.26	.24	.30	.01	.26	.16	.30	.02	-.03	.11	25_{n}	71	29	4.16*
Coaching	-	.14	-	-	-	-	-	.01	.13	-	2	100	0	-
Teamwork	-	-	-	.22	-	-	-	.22	-	.02	2_{n}	100	0	-
Build Rel.	.20	.23	-	.14	-	-	-	.14	.21	.03	9_{n}	75	25	2.00
Manage. Dis.	.05	.15	-	.07	-	-	-	-	-	-	0	-	-	-
Open Comm.	.18	.19	-	.16	-	-	-	.10	.17	.04	9	89	11	5.44*
Est. Plans	-	-	-	-	-	.07	.15	.13	-	-	0	-	-	-
Cust. Focus	-	-	-	.13	-	-	-	.18	-	.05	2	50	50	-

Note: N = 99. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion, PR = Presentation. # Pairwise Comp. refers to the number of same-dimension comparisons across similar and dissimilar exercise pairs. This refers to the number of comparisons prior to removing any pairs that were of the same magnitude, of which there were six within .003. These are marked by the subscript "n" and were not used in computing the chi square statistic. Chi-square was computed using sign test at 1 df and one-tailed tests. Because tests with less than four pairwise comparisons could not possibly reach significance, chi-square statistics were not computed unless there were at least four comparisons for a given dimension. * $p < .05$. ** $p < .01$.

Validity by Exercise Similarity

Once again, the validities of individual OER's were first examined before testing Hypothesis 1 in Sample 2. Table 14 displays the means, standard deviations and correlations amongst OER's, ODR's, the OAR and job performance in Sample 2. As was the trend in Sample 1, there was variation in terms of how predictive each OER was. The interview (r = .17), in-basket (r = .19), and group discussion (r = .16) had similar validities. The presentation exercise was the best predictor of job performance (r = .23). On the other hand, ratings from the role play exercise correlated only .04 with job performance, thus inhibiting the validity of any exercise pairing including this exercise. This concern will be elaborated upon in more detail in the General Discussion section. The OAR for Sample 2 had a correlation of (.27) with job performance.

Table 15 displays each exercise composite and their validities. Like Sample 1, pairings of dissimilar exercises for the most part resulted in higher predictive validities. When the top three similar and top three dissimilar pairings were examined the mean validities were (r =.15) and (r = .21) respectively. When the 10 pairings were split into five and five, the similar exercises had a mean validity of (r = .18), and the dissimilar exercises (r = .23). As in Sample 1, to test whether ratings from dissimilar exercises significantly predict job performance better than ratings from similar exercises, a sign test was employed using a chi-square statistic at 1 degree of freedom. Table 16 displays the individual dimension comparisons across similar and dissimilar exercise pairs.

Table 14. *Descriptive Statistics and Correlations amongst OER's, ODR's, OAR and Job Performance (Sample 2)*

	Mean	1	2	3	4	5	6	7	8	9	10	11
1. Interview	3.24	(.29)										
2. In-Basket	2.77	.20	(.43)									
3. Role Play	2.88	.24	.08	(.43)								
4. Group Discussion	2.93	.13	.16	.22	(.48)							
5. Presentation	3.04	.21	.19	.26	.22	(.54)						
6. Analyze Issues	3.13	.36	.36	.39	.52	.56	(.36)					
7. Judgment	3.02	.36	.39	.48	.65	.25	.69	(.35)				
8. Establish Plans	2.89	.38	.51	.24	.26	.72	.59	.44	(.37)			
9. Manages Execution	2.76	.44	.31	.36	.56	.25	.44	.66	.42	(.34)		
10. Leads Courageously	3.01	.37	.29	.42	.41	.54	.49	.54	.57	.54	(.39)	
11. Influence	2.91	.45	.51	.46	.53	.61	.55	.57	.66	.54	.64	(.38)
12. Coaching	2.75	.39	.47	.51	.08	.06	.18	.24	.18	.20	.18	.36
13. Teamwork	2.98	.27	.57	.17	.39	.12	.18	.29	.33	.26	.10	.47
14. Building Relationships	3.08	.19	.28	.42	.26	.06	.03	.09	.03	.06	-.01	.29
15. Managing Disagreements	2.69	.41	.22	.45	.56	.15	.45	.46	.23	.43	.33	.51
16. Open Communication	3.10	.29	.21	.52	.39	.12	.38	.31	.13	.16	-.03	.35
17. Customer Orientation	3.19	.49	.40	.25	.45	.26	.33	.49	.40	.40	.48	.49
18. OAR	3.00	.61	.58	.57	.60	.60	.73	.73	.70	.66	.67	.86
19. Job Performance	103.00	.17	.19	.04	.16	.23	.19	.14	.26	.20	.17	.21

Note: N = 99. OER = Overall Exercise Rating. ODR = Overall Dimension Rating. OAR = Overall Assessment Center Rating. Numbers in parentheses are standard deviations.

Table 14 Continued…

	12	13	14	15	16	17	18	19
12.	(.38)							
13.	.48	(.44)						
14.	.51	.44	(.4)					
15.	.32	.27	.31	(.32)				
16.	.42	.38	.66	.54	(.4)			
17.	.26	.34	.26	.29	.24	(.39)		
18.	.51	.54	.41	.61	.51	.64	(.24)	
19.	.12	.17	.05	.10	.05	.17	.27	(11.63)

Table 15. *Exercise Composite Validity by Exercise Similarity (Sample 2)*

Exercise Pairing	Exercise Similarity Rank based on TAP	Exercise Composite Validity
RP - GD	1	.14
INT - RP	2	.14
RP - PR	3	.17
INT - GD	4	.22
GD - PR	5	.25
INT - PR	6	.25
IB - PR	7	.27
INT - IB	8	.23
IB - RP	9	.17
IB - GD	10	.23
Average of 3 Similar Exercise Pairings		.15
Average of 3 Dissimilar Exercise Pairings		.21

Note: N = 99. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion, PR = Presentation. TAP = Trait Activation Potential.

Of the seven dimensions in which chi-square statistics could be calculated, six were in the hypothesized direction and four of these differences were statistically significant. In terms of examining the exercise pairings as a whole, testing all the pairwise validities simultaneously resulted in a total of 111[7] pairwise dimension composite comparisons, with 71% of the higher validities occurring in dissimilar exercise pairs. Overall, dissimilar exercises had higher dimension composite validities (median = .16) than similar exercises (median =.10, χ^2 = 19.90, $p < .001$), supporting Hypothesis 1. As was seen in Sample 1, when predicting job performance a composite of two psychologically dissimilar exercises will result in better validity than a composite of two very similar exercises.

[7] Note: *With sign test, if pairwise observations are of the same value then the pair is discarded from the analysis as neutral. Of a possible 117 pair-wise comparisons, six were within the value of .003 and so were treated as neutral.*

Table 16. *Dimension Validity Comparisons by Exercise Similarity (Sample 2)*

	Dimension Composite Validity										Validity Comparisons			
	Similar Exercises					Dissimilar Exercises								
Dimension	RP-GD	INT-RP	RP-PR	INT-GD	GD-PR	INT-PR	IB-PR	INT-IB	IB-RP	IB-GD	# Pairwise Comp.	% Greater Sim.	% Greater Dis.	χ^2
Analyze	.12	.14	.08	.19	.13	.15	.14	.22	.13	.18	25_n	21	79	8.17**
Judgment	.11	.09	-	.12	-	-	-	.12	.11	.13	9_{nn}	14	86	3.57
Execution	.13	.03	-	.14	-	-	-	.17	.16	.25	9	0	100	9.00**
Lead Cour.	.05	.05	.10	.16	.20	.24	.21	.15	.02	.13	25	36	64	1.96
Influence	.16	.04	.16	.23	.32	.23	.17	.07	.02	.18	25_n	58	42	0.67
Coaching	-	.11	-	-	-	-	-	.08	.10	-	2	100	0	-
Teamwork	-	-	-	.08	-	-	-	.16	-	.19	2	0	100	-
Build Rel.	.04	-.01	-	.04	-	-	-	.05	.04	.10	9_n	0	100	8.00**
Manage Dis.	.09	.08	-	.08	-	-	-	-	-	-	0	-	-	
Open Comm.	.01	.04	-	.08	-	-	-	.27	.18	.23	9	0	100	9.00**
Est. Plans	-	-	-	-	-	.23	.20	.26	-	-	0	-	-	-
Cust. Focus	-	.13		.18	-	-	-	.13	-	-	2_n	100	0	-

Note: N = 99. INT = Interview, IB = In-Basket, RP = Role-Play, GD = Group Discussion, PR = Presentation. # Pairwise Comp. refers to the number of same-dimension comparisons across similar and dissimilar exercise pairs. This refers to the number of comparisons prior to removing any pairs that were of the same magnitude, of which there were six within .003. These are marked by the subscript "n" and were not used in computing the chi square statistic. Chi-square was computed using sign test at 1 df and one-tailed tests. Because tests with less than four pairwise comparisons could not possibly reach significance, chi-square statistics were not computed unless there were at least four comparisons for a given dimension. * $p < .05$. ** $p < .01$.

Validity by Behavioral Consistency

The effect of dimension consistency on Assessment Center validity was examined in Sample 2 using the consistent and inconsistent composites. When inspecting these relationships at the dimension level, inconsistent dimension composites had higher validities for only 7 of the 12 dimensions. Table 17 displays these results.

Table 17. *Validities of consistent and inconsistent PEDR pairings (Sample 2)*

Dimension	Consistent Pairing	Consistent Validity	Inconsistent Pairing	Inconsistent Validity
Analyze	GD-PR	.13	IB-RP, & INT-IB	.19
Judgment	INT-RP	.09	IB-DR	.11
Establish Plans	IB-PR	.20	INT-PR	.23
Execution	IB-GD	.25	INT-RP	.03
Lead Cour.	RP-GD	.05	IB-RP	.03
Influence	RP-PR, & IB-PR	.17	IB-RP	.02
Coaching	INT-RP	.11	INT-IB	.08
Teamwork	INT-IB, & INT-GD	.14	IB-GD	.19
Build Rel.	INT-RP	-.01	IB-GD	.10
Manage Dis.	INT-RP	.08	RP-GD	.09
Open Comm.	INT-RP	.04	IB-GD	.23
Customer Focus	INT-IB	.13	IB-GD	.18
Composite		.21		.21

Note: N = 99. INT = Interview, IB = In-basket, RP = Role-Play, GD = Group Discussion, PR = Presentation.

Using the same non-parametric median test previously employed, the dimension validities of the inconsistent pairings were not significantly different (median = .10) than those of the consistent pairings (median = .12, χ^2 = .33, *p* = ns). When looking at the overall composites formed form all dimensions, the difference in predictive validity was negligible as well. The inconsistent composite had a correlation of (.21) with job performance, while the consistent

composite also correlated (.21).These values are obviously not significantly different from one another. Thus, support was not found for Hypothesis 2 in Sample 2.

CHAPTER VII

SAMPLE 2 DISCUSSION

As in Sample 1, behavioral ratings were most consistent across pairs of similar situations. Furthermore, dissimilar pairs of exercises led to better predictions of job performance, once again lending support to Hypothesis 1. Thus, behavioral observation across sets of unique situations seems to result in a more comprehensive and accurate evaluation of individual tendencies. In terms of Hypothesis 2, which examined the direct effect of inconsistency as it related to predicting job performance, there was no difference between sets of consistent ratings and sets of inconsistent ratings. While Hypothesis 2 was therefore not supported in sample 2, it is noteworthy that inconsistency did not adversely affect predictive validity. Thus, while traditional MTMM theory deems inconsistency as problematic, this lack of consistency did not appear to affect correlations with meaningful external criteria.

CHAPTER VIII

GENERAL DISCUSSION

This study took an in depth look at the predictive nature of Assessment Centers. Amidst much research regarding the internal consistency of AC ratings, a theory was developed to better explain why Assessment Centers predict performance despite poor construct validity. Under an assumption of situational specificity, it was assumed that dimension inconsistency occurs as a result of the varied psychological nature of Assessment Center exercises, and that this inconsistency in itself could reflect a broader assessment of individual tendencies. Due to the variation in behavioral requirements, it was predicted that dissimilar exercises would better predict job performance than pairs of similar exercises. Likewise, the ensuing inconsistency that which occurs was posited to reflect this broader assessment and thus also result in higher validities than more consistent AC ratings.

Several noteworthy findings were uncovered. First, as with previous research (Highhouse & Harris, 1993; Haaland & Christiansen, 2002) support was given that behavioral consistency is lower when situational demands are unique. In both Sample 1 and Sample 2 the monotrait-heteromethod correlations were highest in pairs of similar exercises. When exercises became more dissimilar the dimension consistency from exercise to exercise became lower. This finding strengthens the theory of situational specificity in Assessment Centers. Specifically, the behavioral inconsistency so often witnessed can at least partially be explained by the varied nature of AC demands. Likewise, as research and theory have generally accepted, inconsistency in of itself is not solely measurement error. Instead, due credit should be given to the strength of situational demands and the effect of those demands on behavior.

Table 18. *Summary of Hypothesis Testing*

	Sample 1	**Sample 2**
Hypothesis 1		
Exercises with varied psychological demands will capture a wider spectrum of candidate behaviors and lead to better predictions of job performance criteria than ratings from similar exercises.	Supported	Supported
Hypothesis 2		
Inconsistent dimension ratings reflect measures assessing a broader spectrum of candidate behaviors and so will lead to better predictions of job performance criteria than behaviorally consistent dimension ratings.	Partially Supported	Not Supported

Second, initial support was found for a situational specificity/bandwidth approach to maximize predictive validity in Assessment Centers (Table 18). As posited, combinations of exercises with varied psychological demands were more predictive of job performance than exercises that were rated as more similar. This relationship was found in both samples, highlighting the importance of assessing trait-expressions not only in one class of situations, but instead under a variety of contextual demands. As behavior is a function of both the person and the situation (Funder, 2006) it is no doubt intuitive that a better understanding of individual tendencies can be met when behavior is assessed in different types of job-related scenarios. It is important to note that while traditional construct theory calls for high reliability (Campbell & Fisk, 1959), the current findings demonstrate that dissimilar situations result in both inconsistency and yet higher predictive validity. It may be possible that with high fidelity tests like Assessment Centers, because overt behaviors are more unpredictable than paper and pencil responses, observation across *many* situations is required to truly get an accurate assessment of

one's individual tendencies. In terms of behavioral assessment, bandwidth may then be the belle of the ball.

This finding is important in regards to the use of Assessment Center as selection methods and shows direct support for this method's content validity. By nature, AC's are developed to represent the broad set of performance related behaviors relevant to a given job (e.g. Neidig & Neidig, 1984). By demonstrating that two very different exercises can lead to better predictions of job performance, it is possible that Assessment Center design can be improved by attempting to tap the broadest and most varied contexts of work behavior. From a utility perspective this is especially important, as those exercises that are most dissimilar can be chosen when constraints on testing time or cost require so. A combination of repetitive and overlapping simulations will not be as informative as several simulations that all measure different requirements of a job. For example, instead of incorporating two role plays that each involve manager-customer interactions, one exercise could be changed to reflect a subordinate-manager exchange. As Assessment Centers are simulated scenarios used to predict incumbent performance, the AC as a whole will benefit if the exercises encompass the entire job domain as well as possible.

Lastly, while exercise characteristics were theorized to affect behavioral consistency, this inconsistency itself was examined as it related to predicting job performance. This thesis if the first known study to examine this relationship at the sample level. Findings here were promising, yet inconclusive. Inconsistency was examined with job performance to determine whether the expected consistency across exercises is actually a desirable characteristic in terms of predictive validity. In Sample 1 an optimally inconsistent composite of AC ratings resulted in higher predictive validity than an optimally consistent composite, although the difference between the

correlations was not statistically significant. Considering that each predictor was composed of, to a portion, the same PEDR's (see Tables 8), any difference is somewhat impressive. However, in Sample 2 the validities were essentially the same for both inconsistent ratings and consistent ratings.

It is then unclear what do draw from these relationships. Although it seems likely that inconsistency is reflective of a more varied and broader assessment, it seems just as plausible that some of this inconsistency occurs because of measurement error as well. This error could stem from poorly defined dimensions, not possessing equal opportunities to observe all dimensions across exercises, poor rater training, and so forth. For example, in the current study dissimilar exercises had both more inconsistent ratings and higher predictive validities. So then why wasn't an optimally inconsistent composite a significantly better predictor? It is likely that inconsistency is a function of *both* high bandwidth *and* measurement error. The inconsistent composite was composed of the absolutely most unrelated PEDR's, with some of these values falling as negative correlations. It is unlikely why any measures of the same dimension, even in very different situations, should be negatively related. So, although it may be true that dimension inconsistency across two exercises likely occurs because of inter-individual differences in the ability to handle two sets of behavioral demands, it is also likely that this is coupled with true measurement error. Thus, inconsistency in Assessment Centers may have dual causes and dual effects, although the likes of which may be difficult to separate.

While the general finding that inconsistent composites do not predict performance better than consistent composites may seem discouraging, it does serve to further move us past the Assessment Center construct debate. As recent construct research has generally accepted

dimension inconsistency as an expected occurrence, this study demonstrates that inconsistency in no way hinders the predictive nature of Assessment Centers. In both samples a composite of inconsistent ratings performed as well, if not better than consistent composites in predicting job performance. The belief that consistency is required across exercises once again seems to be misleading. Although finding only small differences in validity coefficients, this study poses tentative support that inconsistency may actually be a desirable aspect of ratings.

Limitations and Areas for Future Research

Despite promising findings supporting a theory of "situational bandwidth" in Assessment Centers, some concerns threaten the generalizability of these results. First and foremost, despite using two samples in an attempt to broaden the findings, both these groups were very similar. Each sample underwent Assessment Centers conducted by the same consulting company and composed of mostly the same exercises. Thus, while treated as separate, it is hard to come to any definite conclusions regarding these results, as the same essential AC design was incorporated in both. Considering the strongest findings involved the predictive nature of AC exercises, one is left to wonder whether a different Assessment Center with completely different exercises would find the same results. What if different types of exercises were used (e.g. business games and case-analyses, or role plays and group discussions with very different demands)? Further, although interviews are commonly used in Assessment Centers it may not be appropriate to treat them as exercises, which was done in both samples here.

On a related note, in both samples one exercise had extremely lower validity coefficients than others. While this exercise (role play) was included in each the similar and dissimilar pairings, it causes concern for the stability of the findings. It is possible that any pairing that

incorporated the role play exercise had lower validities not because the combination involved dissimilar or similar assessments, but rather because not all exercises were on an equal playing field to begin. This could have been a problem not only when testing the validities across exercises, but also when forming the optimally inconsistent and consistent composites, as PEDR's from the role play were more likely to have lower relationships with job performance. Although not every exercise should predict performance equally well, a pure test of the situational bandwidth hypothesis would require so. Thus, examining these relationships with a different set of exercises would increase confidence in these findings.

Future research should expand upon these results and look at the relationship between exercise similarity/bandwidth, the ensuing dimension consistency, and in what ways this affects an Assessment Center's ultimate outcome: prediction or development. While this study provides tentative understanding, a wider variety of exercises should be examined that differ in form, surface level content, deep level content, type of job, type of sample, and type of Assessment Center. For instance, how do the effects of exercise dissimilarity compare between off-the-shelf AC's like those from this study and fully tailored Assessment Centers developed from job analysis? The off-the-shelf nature of the Assessment Centers in this study may have actually attenuated results, as there is likely to be less overlap between job demands and predictor characteristics. Thus, some exercises and some behaviors may not have been necessarily relevant for all jobs. This is likely reflected by the overall low validity coefficients for both AC's, which place themselves at the bottom end of Gaugler et al.'s (1987) 95% confidence interval and well below values reported more recently (Meriac et al., 2008).

In a related vein, the monotrait-heteromethod correlations and heterotrait-monomethod correlations were both much smaller than what is typically found in Assessment Centers. It is unclear exactly why this was so, but conventional thinking might point to flaws in AC design and implementation. Despite expecting dimension inconsistency in Assessment Centers, there is no reason to think that these values would be as low as they were or that within exercise correlations would be so small. Thus, the makeshift nature of the off-the-shelf Assessment Centers used in this study causes legitimate concern.

On a final note, while Trait Activation Potential is a useful method of assessing differences between situations, there are perhaps other or more precise measurement techniques that could also be employed. With that said, at this point there is no agreed upon way on how to measure situations, let alone specific AC exercises. As TAP has revealed meaningful information in this study and in past Assessment Center research (e.g. Haaland & Christiansen, 2002), it may be an appropriate technique for exercise description until a more established and precise method comes along.

Conclusion

Like the person-situation debate of social psychology, behavioral ratings in Assessment Centers are no doubt the function of situational demands and true individual differences. The current study took an in depth examination of how the situational characteristics of AC exercises affect the breadth of observable behaviors, thus affecting the bandwidth and accuracy of individual assessment. Findings provide a lesson of the basic tenets of measurement – that there are gains and losses in choosing the domain of one's measure. Assessment Centers are by nature, a rather high bandwidth assessment. The measured behavioral constructs span commonly

assessed cognitive aptitudes such as problem solving and decision making, interpersonal skills like communication and presentation, and even stability traits like stress tolerance and adaptability; this all while spanning numerous situational contexts and scenarios. While this high bandwidth assessment results in a measure that encompasses a wide body of behaviors, it does so at the expense of consistency. This thesis examined whether this inconsistency is actually such a bad thing in Assessment Centers. Intuitively, broader or two dissimilar situations resulted in lower dimension consistency and poorer psychometric properties according to traditional MTMM theory. However, despite a lack of consistency, it was these dissimilar exercises that better predicted job performance. Likewise, although findings were not conclusive, it seems the actual inconsistency typically observed in Assessment Centers in no way hinders the predictive validity that they display.

In regards to the "construct debate", findings here provide more reason to move past the concern for across-exercise consistency and within-exercise discrimination. We already know that exercise effects account for greater proportions of variance that do behavioral dimensions. This is accepted. In today's age, examinations of shear internal psychometrics in Assessment Centers may be less worthwhile than simultaneous examinations of how the structure of AC constructs relates to meaningful outcomes such as the prediction of job performance. As others have displayed (e.g. Lance et al, 2000), these methods may be working exactly as they were intended to. The next step is to further move towards the creation of a comprehensive theory that links both constructs and outcomes. Examining how things like exercise demands and behavioral consistency affect predictive validity is an applicable way to improve upon the Assessment Center methodology and further our understanding of behavioral assessment.

APPENDIX A

EXAMPLE OF THE TRAIT ACTIVATION POTENTIAL ITEMS: TRAIT ACTIVATION POTENTIAL ITEMS FOR EXTRAVERSION IN THE TASK FORCE EXERCISE

Please answer the following questions in regards to the Assessment Center exercise and FFM Dimension presented below.

TASK FORCE
EXTRAVERSION

Is there opportunity to observe behaviors related to the trait of Extraversion while a candidate completes the Task Force Exercise?

NO, very few behaviors are related to trait		SOME behaviors are related to trait		YES, observed behaviors are primarily related to trait
1	2	3	4	5

Is it possible to make inferences of a person's level of Extraversion from observing the Task Force Exercise?

Behaviors are NOT related to trait		Behaviors allow for SOME inferences of trait level		Behaviors allow for FIRM inferences of trait level
1	2	3	4	5

In the Task Force Exercise, are behaviors related to Extraversion highly valued, expected or required for success?

Trait NOT Important for success		Important, but some traits are more important for success		MOST important trait for success
1	2	3	4	5

How often do these behaviors occur in the Task Force?

	Behaviors NEVER occur		Behaviors SOMETIMES occur		Behaviors FREQUENTLY occur
Taking control in group situations	1	2	3	4	5
Presenting information to persaude others	1	2	3	4	5
Actively participating in conversations	1	2	3	4	5

REFERENCES

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

Bartram, D. (2005). The Great Eight Competencies: A Criterion-Centric Approach to Validation. *Journal of Applied Psychology*, *90*, 1185-1203.

Bem, D.J., & Funder, D.C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review*, *85*, 485-501.

Borman, W.C., & Motowidlo, S.J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-109.

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.

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, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. *Handbook of Industrial and Organizational Psychology*, *1*, 687-732.

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

Cascio, W.F., & Aguinis. (2005). *Applied Psychology in Human Resource Management.* Upper Saddle River, NJ: Pearson Prentice Hall,

Costa, P.T., & McCrae, R.R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*, 5-13.

Cronbach, L.J. (1960). *Essentials of Psychological Testing, (2nd Ed.*). New York, NY, US: Harper & Row.

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological Tests and Personnel Decisions, (2nd Ed.*). Urbana, IL: University of Illinois Press.

Funder, D.C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*, 21-34.

Funder, D.C. (2009). Naive and obvious questions. *Perspectives on Psychological Science*, *4*, 340-344.

Funder, D.C., & Colvin, C.R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60*, 773-794.

Funder, D.C., & Colvin, R.C. (1997). Congruence of others' and self-judgments of personality. In Hogan, R., Johnson, J.A., & Briggs, S.R. (Eds.), *Handbook of Personality Psychology* (617-647). San Diego, CA, US: Academic Press.

Funder, D.C., Furr, M.R., & Colvin, R.C. (2000). The Riverside Behavioral Q-sort: A tool for the description of social behavior. *Journal of Personality*, *68*, 451-489.

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

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

Gibbons, A.M., & Rupp, D.E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*, *35*, 1154-1180.

Guidelines and ethical considerations for assessment center operations: International task force on assessment center guidelines. (2009). *International Journal of Selection and Assessment*, *17*, 243-253.

Haaland, S., & Christiansen, N.D. (2002). Implications of trait-activation theory for *Personnel Psychology*, *55*, 137-163.

Highhouse, S., & Harris, M.M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology*, *23*, 140-155.

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment*, 9(1), 114-123.

Hogan, J., & Roberts, B.W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior*, *17*, 627-637.

Jackson, D.J., 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., Stillman, J.A., & Atkins, S.G. (2005). Rating Tasks Versus Dimensions in Assessment Centers: A Psychometric Comparison. *Human Performance*, *18*, 213-241.

Kolk, N.J., Born, M., & van der Dlier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325-338.

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

Lance, C.E., Foster, M.R., Gentry, W.A., & Thoresen, J.D. (2004). Assessor Cognitive Processes in an Operational Assessment Center. *Journal of Applied Psychology, 89*, 22-35.

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

Lance, C.E., Lambert, T.A., Gwin, A.G., Lievens, F., & Conway, J.M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377-385.

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

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

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

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

Lievens, F. (2008). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18*, 102-121.

Lievens, F., Chasteen, C.S., Day, E.A., & Christiansen, N.D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*, 247-258.

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

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

Meriac, J.P., Hoffman, B.J., Woehr, D.J., & Fleisher, M.S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052.

Mischel, W. (1968). *Personality and assessment*. Hoboken, NJ, US: John Wiley & Sons Inc.

Mischel, W., & Peake, P.K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.

Mischel, W. (1985, October). Diagnosticity of situations. Paper presented at the meeting of the Society for Experimental Social Psychology, Eston, IL.

Murray, H.A. (1938). *Explorations in personality*. New York: Oxford University Press.

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

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

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

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

Saucier, G., Bel-Behar, T., & Fernandez, C. (2007). What modifies the expression of personality tendencies? Defining basic domains of situation variables. *Journal of Personality, 75*, 479-504.

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.

Scullen, S. E., Mount, M. K., & Judge, T. A. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88*, 50–66.

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*. Champaign, IL, US: University of Illinois Press.

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.

Tett, R.P., & Burnett, D.D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500-517.

Tett, R.P., Guterman, H.A., Bleier, A., & Murphy, P.J. (2000). Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance*, *13*, 205-251.

Thornton III, G.C., & Gibbons, A.M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 19*, 169-187.

Wise, L.L, McHenry, J., & Campbell, J.P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. *Personnel Psychology*, *43*, 355-366.

Wright, J.C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology*, *53*, 1159-1177.