INTRODUCTION

In the most recent meta-analysis on assessment centers, it was estimated that there were more than 2000 organizations using some type of assessment center method (Gaugler, Rosenthal, Thornton, & Bentson, 1987). Due to the increasing popularity and effectiveness of the method, it is likely that the number of operating centers has increased well beyond this number in the last 11 years. During the 40 years assessment centers have been in operation, hundreds of thousands of people have participated in them and they have been the subject of over 100 research and review articles as well as several books (Howard, 1997). Assessment centers are conducted for a variety of reasons including employee selection, development planning, identification of training needs, management succession, and promotion. A commonly accepted description of an assessment center is "a variety of testing techniques designed to allow candidates to demonstrate, under standardized conditions, the skills and abilities most essential for success in a given job" (Joiner, 1984, p. 437).

It has been noted that there may not be one "typical" assessment center because assessment centers vary widely in purpose, exercises, and procedures; however, the latest assessment center guidelines established by the 17th International Congress on the Assessment Center Method in 1989 help to specify what is and is not considered an assessment center. This document was designed to establish professional guidelines and ethical considerations for users of the Assessment Center method (Task Force on Assessment

Center Guidelines, 1989 as cited in Thornton, 1992). Generally, these guidelines specify that assessment centers must consist of categories (e.g., attributes, characteristics, aptitudes, qualities, skills, abilities, knowledge, tasks) derived from a job analysis; techniques or exercises designed to provide information on the categories; and assessors that must use a systematic procedure to record specific behavioral observations as they occur (Task Force on Assessment Center Guidelines, 1989 as cited in Thornton, 1992).

The 17th International Congress on the Assessment Center Method also specified what types of activities do not constitute an assessment center. These activities include panel interviews, reliance on a single technique, single assessor assessments, the use of multiple simulations where there is no pooling of data, and a physical location called an “assessment center” if it does not conform to the above requirements.

Validity of Assessment Centers

The first use of multiple assessment procedures is attributed to German military psychologists. Both Britain and the United States then adopted this procedure. By late 1969, multiple assessment programs were being used for identifying management potential in several industrial firms and government agencies including AT&T, Caterpillar Tractor Co., General Electric Co., Sears, Roebuck & Co., the Internal Revenue Service, etc. The first study to systematically assess managerial talent was led by Bray and his associates (Bray & Grant, 1966) for the AT&T Company's Management Progress study, which began in 1956. Douglas Bray generated twenty-six dimensions for the original AT&T

study. These dimensions were developed through a review of the psychology and business literature, through interviews with job incumbents, and through consultation with Bell System executives. In the study, 422 managers were rated during the four years between 1956 and 1960. The results of this assessment center were viewed so useful by managers that the assessment center method spread widely through the Bell System. The results of the Management Progress study show that 51% of those who were predicted to make middle management did, in fact, make it. Only 14% of those predicted not to make middle management actually achieved the level (Bray & Grant, 1966). The predictions after eight years and sixteen years show even more accuracy (Mayes, 1997). In this study, the assessment center results were never revealed to anyone in the organization and they were not used for promotion decisions. Follow up studies to the Management Progress Study demonstrated considerable predictive power for both non-college graduates and college recruits (Bray & Campbell, 1968).

Subsequent studies have supported the predictive validity of assessment centers across a variety of assessment center purposes, designs, and participants (Cohen, Moses & Byham, 1974; Thornton & Byham, 1982; Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984; Gaugler et al., 1987). While the range of validity coefficients have ranged from -.25 to + .78, true validity is estimated at .37 with even higher validities found in studies in which potential ratings are used as the criterion (Gaugler, Rosenthal, Thornton, & Bentson, 1987). Because of this high validity estimate, assessment centers remain one of the best predictors of performance in industrial

psychology (Gaugler et al., 1987). They also appear to work equally well regardless of educational level, gender, or prior experience of the participants (Klimoski and Brickner, 1987)

Although the popularity of assessment centers continues to increase, the controversy surrounding their use continues. Assessment centers have been termed “the modern enigma in human resource practice” (Klimoski & Brickner, 1987, p. 243). Despite the growing literature supporting their overall effectiveness (e.g., Cascio and Silbey, 1979; Sackett & Dreher, 1982; Ritchie and Moses, 1983; Klimoski and Brickner, 1987; Thornton and Cleveland, 1990; Thornton, 1992), little is known about why assessment centers consistently yield predictive validity.

Assessment center design is based on the trait theory of personality that suggests individuals have relatively stable characteristics that influence behavior across many situations and that these traits can be reliably measured (Turnage and Muchinsky, 1982). The classic explanation for why assessment centers work is that assessors observe behavior displayed in exercises, classify those behaviors into categories of human attributes or dimensions, and use those categories to make meaningful predictions about job performance (Howard, 1997). Critics have challenged this explanation because research has consistently shown that dimensions are not rated consistently across exercises, demonstrating a lack of construct validity (Sackett & Dreher, 1982; Bycio, Hahn, & Alvares, 1987; Fleenor, 1996).

In the early 1980's Sackett & Dreher were the first to investigate whether the constructs underlying assessment centers were being reliably measured. Up until this time it

had been assumed that assessment centers were measuring what they purported to measure. They performed a principal components factor analysis and found that the correlations between various dimensions within one exercise were high, whereas the intercorrelations of one dimension across various exercises were low. The factors reflected exercises rather than dimensions. These results have been replicated in many other studies (Archambeau, 1979; Bycio et al., 1987; Joyce, Thayer, & Pond, 1994; Russell, 1987), providing further evidence that what is being measured in assessment centers remains a mystery.

Howard (1993) warned that unless more attention is paid to the constructs being measured, construct validity may end up being the "Big Bad Wolf" that is threatening to blow the house down on assessment centers. She writes, "..It hasn't blown the house down yet, but it is out there huffing and puffing" (p. 15). Thus, research aimed at increasing the construct validity of functioning assessment centers will significantly improve our understanding and acceptance of the assessment center process as a whole.

Improvements in Construct Validity

Small improvements in assessment center construct validity have resulted from changes in rating procedures, rater characteristics, exercise design and dimension characteristics (Gaugler & Thornton, 1989; Shore, Shore, & Thornton, 1992; Kleinmann, 1993; Sagie & Magnezy, 1997). In particular, changes in the selection and definition of assessment center dimensions appears to be a promising area for increasing the construct validity of assessment centers. Many authors have argued that specificity and observability of dimensions should be considered in the search for

construct validity (Thornton, 1992; Hampson, John, & Goldberg, 1986; Turnage & Muchinsky, 1982; Reilly, Henry, & Smither, 1990; Bycio et al., 1987). However, only one study has empirically examined the influence of observability of dimensions on the validity of an assessment center (Shore, Shore, & Thornton, 1992). The increase in construct validity with the use of more 'observable' dimensions in this study reveals that characteristics of dimensions are a useful and important research area for understanding and increasing the construct validity of assessment centers. Unfortunately, a review of references to characteristics of dimensions (e.g., observability, specificity) in the Shore et al. (1992) study as well as the assessment center literature reveals inadequate definitions of observability and related constructs. Even more disturbing is the fact that often times the constructs are not defined at all. As a result of the lack of clarity in defining and measuring these constructs as well as a lack of theory to guide the research, the study of dimension characteristics has not been accompanied by empirical or theoretical progress in the search for increased construct validity. One dimension characteristic that has been completely overlooked in the assessment center literature is the diagnosticity of the behaviors that represent a particular dimension. It is expected that a dimension with highly diagnostic behaviors may be easier for assessors to evaluate than a dimension with few or no highly diagnostic behaviors.

Diagnosticity of Behaviors

The existing literature on assessment centers has completely disregarded the fact that not all behaviors are equally informative (Jones & Davis, 1965). For example,

Shore et al. (1992) found that more ‘observable’ dimensions had greater construct validity than did less ‘observable’ dimensions. In the discussion of these results, the researchers suggest that it is the amount of behavioral information for a dimension that is important to increasing construct validity. However, regardless of the number of behaviors a dimension has representing it, it is more likely that it is the diagnosticity of those behaviors that is most important. The lower the prior probability of a behavior, the more informative or diagnostic it is (Jones & Davis, 1965). According to Trope's model of dispositional judgment (1975), the diagnostic value of a behavior depends on a) the probability that a situation would produce such behavior given that the target person has the hypothesized disposition and b) the probability that the situation would produce the behavior given that the target person does not have the hypothesized disposition. Diagnosticity determines the certainty with which a given disposition is inferred from an identified behavior (Trope & Liberman, 1993). The predictability of dispositions increases with situational information. Assessment centers are designed to have explicit contrived situations and thus, the contextual information provided maximizes the chances for a correct inference of a participant’s standing on a particular dimension.

With practice, the process of inferring dispositions becomes automatized and people are able to observe behavior and almost reflexively make accurate dispositional inferences (Trope & Bassok, 1982). Providing support that this process may be occurring in assessment centers, research has shown that psychologists acting as assessors

often yield higher predictive validities than managers (Gaugler et al., 1987). Because psychologists have more experience observing and classifying behaviors, the process of inferring dispositions from diagnostic information is presumably easier for this group. Regardless of experience inferring dispositions, diagnostic behaviors should serve to increase the accuracy of the inference from behaviors to placement on a particular dimension for all assessors. Thus, the greater the proportion of highly diagnostic behaviors a dimension has representing it, the easier the dimension will be for the assessors to measure, resulting in a greater likelihood that the assessment center is measuring what it purports to measure (i.e., greater construct validity).

The purpose of the current study, therefore, is to attempt to clarify the role of ‘ease of evaluation' of dimensions in an assessment center. The study will examine the relationship of ‘ease of evaluation' to assessment center construct and criterion-related validity.

Literature Review

CONSTRUCT VALIDITY OF ASSESSMENT CENTERS

One of the most widespread criticisms of assessment centers involves their lack of construct validity (Sackett & Dreher, 1982; Fleenor, 1996). Researchers have often found an "exercise" factor that overwhelms the dimensional effect. This finding suggests that assessors are capturing exercise performance in their ratings and not stable personal characteristics, as intended.

Construct validity is established by examining a wide variety of evidence about the internal structure of an assessment procedure and its relationship to other tests and measures. One type of construct validity evidence is the relationship among various parts of the assessment center. The two main statistical procedures that have been used to observe these relationships of within-exercise dimension ratings are multitrait-multimethod validity analysis (Campbell & Fiske, 1959) and factor analysis (Gorsuch, 1983).

Multitrait Multimethod Validity analyses

In order to evaluate an MTMM matrix, convergent and discriminant validity correlations must be computed. To test for convergent validity, one must examine the correlations of the same trait measured by different methods (i.e., monotrait- heteromethod). To test for discriminant validity, one must compare the correlations of different dimensions measured by different methods (i.e., heterotrait-heteromethod) to the correlations of the same dimension measured by multiple methods (i.e., monotrait-heteromethod). A second, and more stringent way of evaluating discriminant validity is to calculate the correlations among the

different dimensions within each of the exercises (mean heterotrait-monomethod).

A lack of convergent and discriminant validity has been shown to exist through multitrait-multimethod studies. It has been shown that ratings on the same dimension across different exercises do not correlate as highly as do ratings on different dimensions in a single exercise (i.e., lack of convergent validity) (Bycio, Alvares, & Hahn, 1987; Russell, 1987, Sackett & Dreher, 1982; Kudisch, Ladd, and Dobbins, 1997). In addition, the multitrait-multimethod analysis of assessment center dimension ratings has consistently shown higher within-exercise correlations of different dimensions than cross exercise correlations of the same dimensions (i.e., lack of divergent validity) (Bycio et al, 1987; Robertson et al., 1987; Sackett & Dreher, 1982; Turnage & Muchinsky, 1982).

In the assessment center studied by Sackett and Dreher (1982), they found ratings of the same ability were uncorrelated across exercises, demonstrating a complete lack of convergent validity. These researchers concluded that there was “virtually no support for the view that the assessment center technique generated dimension scores that can be interpreted as representing complex constructs” (p. 409).

Bycio et al. (1987) attempted to test the cross-situational specificity of the dimensions (i.e., convergent validity) in an assessment center. Similar to Sackett and Dreher (1982) they found almost no evidence of either convergent or discriminant validity. In fact, the discriminant validity coefficients found were perfect or near-perfect correlations. The authors concluded, “.. the

preponderance of evidence suggests that the method does not measure large sets of job-related abilities" (p. 470).

In a similar study, Neidig and Neidig (1984) found reasonably large convergent validity coefficients but low discriminant validity for their assessment center. A number of studies have had similar findings (Archambeau, 1997, Konz, 1988; Outcalt, 1988; Robertson, Gratten, & Sharpley, 1987; Schneider & Schmitt, 1992; Hinrichs & Haanpera, 1976, Turnage & Muchinsky, 1982). Neidig and Neidig (1984) were not particularly surprised by these findings because they argued that "properly designed situational exercises purposely place assessees in a variety of job-related contexts and therefore, stable performance across exercises by all participants is not necessarily expected" (p.184). It is agreed by many researchers that assessment center performance is at least partially situationally determined, however, one basis for using dimensions to begin with is the belief that the dimensions or traits are, at a minimum, moderately stable and can be reliably measured across exercises.

Turnage and Muchinsky (1982) also found high convergent validity and low discriminant validity in their study of assessment centers. They identified person, situation, and trait components via the multitrait-multimethod matrix analysis. Their results revealed that raters generally agreed to a large extent on the ordering of individuals at least on a global basis (i.e., high convergent validity). The authors warned that the high convergent validity coefficients must be interpreted with caution because it is possible for halo error to cause spurious validity coefficients for assessment ratings. Findings demonstrated

that overall ability or trait measures did not separate into specific factors, indicating a lack of discriminant validity.

Kavanagh, MacKinney, & Wollins (1971) also found very little differentiation among traits (discriminant validity) over ratings on 20 dimensions of managerial job performance using a multitrait-multimethod matrix analysis. These authors concluded that the number of dimensions rated should be reduced and suggested using only those dimensions that define effective job performance and are most meaningful.

As described above, while convergent validity has been shown to be weak in some assessment centers, discriminant validity seems to be the larger problem for the construct validity of assessment centers. It is important to note that there is some controversy as to what is considered adequate convergent and discriminant validity coefficients. The same correlation may be interpreted as high or adequate convergent validity in one study or as low in a second study. Because of this ambiguity, validity coefficients must be compared to the results found in other studies in order to facilitate interpretation (Sackett and Dreher, 1982). Unfortunately, these comparisons are rarely seen in the literature.

To provide a framework for the variety of convergent and discriminant validity coefficients that have been found in assessment center studies, some examples will be given. Sackett and Dreher (1984) found a mean convergent validity of .09. They did not report discriminant validities for their study. Hinrichs and Haanpera (1976) found that convergent validity coefficients averaged .49 for 14 dimensions and they also did not report discriminant

validities. Archambeau (1979) found average convergent validity of .61 and average discriminant validity equal to .89. Russell (1987), found low convergent validity, .25, and a higher discriminant validity .52, although this discriminant validity coefficient may be considered low compared to many other studies. Similarly, Baker (1986, as cited in Thornton, 1992) found a mean convergent validity of .26 and a mean discriminant validity of .58. Bycio et al. (1987) found a convergent validity of .36 and discriminant validity of .75. Adler and Margolin (1989 as cited in Thornton, 1992) found similar results to Bycio et al. (1987) with a mean of .32 for the convergent validity and a mean of .82 for the discriminant validity. An examination of these coefficients demonstrates the wide range of convergent and discriminant validity coefficients that have been found in assessment center studies.

In a study examining the criterion and construct validity of an assessment center, Chan (1996) examined a comprehensive set of variables including assessor ratings, psychological test measures, and supervisory ratings of job performance as well as actual promotions. He examined the variables through the use of both multitrait-multimethod analyses, factor analysis and external construct validity when placed in a nomological network of constructs and found a lack of both internal and external construct validity for the center. The assessment center was predictive of subsequent promotion (r=.59, p < .01) but not of concurrent supervisory ratings of performance (r=.06, n.s.) The findings of a lack of construct validity in this study is significant because of the rigorous design of the assessment center, providing evidence that the prevalent lack of

construct validity found in assessment center's is not simply due to poorly constructed assessment centers as some claim (Russell, 1994). Chan concluded that priority must be given to what constructs are being tapped in assessment centers because “high criterion-related validity implies that there must be construct validity in assessment centers but we have not yet identified the constructs” (p. 176).

A few studies have even examined the construct validity of individual exercises and the results have not been positive. Brannick, Michaels, & Baker (1989) investigated the convergent and discriminant validity of in-basket scores, a common assessment center exercise. They found little convergent validity between alternate in-basket forms and that training improved in-basket performance on two of the five dimensions as well as overall performance. Both of these findings indicate a lack of construct validity for the use of popular In-basket scores.

Factor Analysis

Factor analysis has also been used to investigate the construct validity of assessment center ratings (Gorsuch, 1983). Factor analysis has been used to examine the correlations among all within-exercise dimension ratings across all the exercises to identify groups of ratings that cluster together (Thornton, 1992). If the ratings are measuring dispositional characteristics such as abilities, then the resulting factors should represent dimensions rather than exercises (Fleenor, 1996).

In one of the most often cited studies on assessment center construct validity, Sackett and Dreher (1982) examined the interrelationships among dimensional ratings between and within exercises in three assessment centers. A

principal-axis factor analysis demonstrated that the factor pattern for all three organizations represented exercises rather than dimensions. Up until this time it had been assumed that assessment centers were based on valid and reliable constructs and that factors would represent the predetermined dimensions. This study was one of the first to show that this was not the case.

Similarly, Bycio et al (1987) performed a series of confirmatory factor analyses on eight abilities from each of five situational exercises and found that the ratings were almost completely situation specific. Assessors were unable to distinguish among the eight abilities. The authors speculated that the demonstration of an exercise effect may imply that the exercises differed on some important parameters that may have limited the extent to which cross-situational consistency could be demonstrated.

In a test of this hypothesis, Schneider and Schmitt (1992) attempted to determine what it is about exercises that overwhelms the dimensional ratings. They investigated the effect of exercise form (e.g., leaderless group exercise, role-play) and exercise content (e.g., competitive, cooperative) on assessment center ratings. As expected, factor analysis revealed that most of the variance in the ratings was explained by exercises and not dimensions. Results indicated that the exercise main effect was due primarily to the form of the exercises and not to the content. They found greater evidence of convergent validity across exercises that were similar in form. Exercise form accounted for 16% of method variance. Exercise content accounted for almost no variance in the ratings, providing little (or no) evidence for an effect of

exercise content (i.e., task design) on convergent validity. The authors interpreted the form effect as potentially reflecting a true exercise effect. Individuals may perform differently in situations that differ (e.g., one-on-one situation versus group situations). Based on these findings, Schneider & Schmitt suggest that in order to understand what happens in the assessment process, considerably more attention must be paid to various aspects of assessment centers such as types of exercises, abilities required, number of participants, etc.

Fleenor (1988), as cited in Thornton (1992) investigated construct validity of dimension ratings from a developmental assessment center. He investigated the relationship between assessment center ratings, job performance ratings, and personality measures through the use of both multitrait-multimethod analysis and factor analysis. He found that the assessment center ratings failed to demonstrate construct validity and the factors underlying the ratings were the assessment center exercises and not the managerial dimensions. Fleenor hypothesized that these findings were a result of the assessors' inability to meet the cognitive demands of the assessment procedure. The assessors are required to use a high level of inference to rate performance on dimensions from observing behaviors in exercises. He suggests that assessment center architects may need to lower the cognitive demands on assessors by using fewer, less-complex dimensions.

Some authors have argued that the consistent findings that assessment centers reflect exercise factors should lead to a change in the scoring of assessment centers such that

they be organized around exercises and not dimensions. In a study designed to compare an assessment center organized around dimensions versus an assessment center organized around the functional structure of managerial work, Joyce, Thayer, and Pond III (1994) investigated the construct validity of two developmental assessment centers. One developmental assessment center measured performance in terms of traditional attribute dimensions and the other in terms of functions performed in managerial work. Results show evidence for construct validity was weak for both sets of constructs. The authors hypothesize that it may be that the construct validity of their assessment center (as well as others) may be affected by the specificity that was included in order to improve reliability. It is possible that a behavior focus may contribute to the inability of assessment centers to demonstrate construct validity. It may be that the focus on behavioral specificity results in the redefinition of the dimension from one exercise to another resulting in new dimensions that may or may not correlate with each other.

In a related study designed to compare the merits of scoring assessment centers by dimension versus scoring them by exercise, Bobrow (1996) explored the data from two validation centers. He found no difference in predicting job performance scoring the assessment center based upon exercises as compared to a scoring method based on dimension. Because no differences were found, Bobrow argues that which method is used should be based on the approach that best fits with the organization's goals and the purpose of the center.

A few studies have not had such dismal findings regarding the construct validity of assessment centers. Louiselle (1986) performed a factor analysis on an assessment center and found that the best explanations of correlations among a set of within-exercise dimension ratings was provided by both exercises and dimensions. Similarly, Kudisch, Ladd, and Dobbins (1997) performed a confirmatory factor analysis that revealed both exercise and dimension factors, suggesting that assessors appear to organize information in terms of both dimensions and exercises. However, similar to past research, their MTMM analysis revealed a lack of both convergent and discriminant validity.

Why Do We Need Construct Validity?

As discussed above, the overwhelming evidence for a lack of construct validity found in studies of within-exercise dimension ratings have raised serious questions about the ability of assessors to make meaningful judgments about dimensions of performance. While some researchers argue that the lack of evidence for the construct validity of assessment centers is “troubling” (Sackett and Dreher, 1982), others insist that such evidence is not necessary to support the use of assessment centers. These researchers argue that assessment centers should be considered valid solely on the basis of accumulated content and criterion related evidence (Neidig and Neidig, 1984). In direct opposition to this claim, Sackett and Dreher (1982) contend that content validity is only one form of evidence of construct validity and therefore it is unwise to depend solely on this evidence. In addition, validity is a unitary concept and overreliance on any one of the three validation

approaches “would be a hindrance to understanding the exact nature of assessment center constructs and explaining their ability to predict performance” (Donahue, Truxillo, Cornwell, & Gerrity, 1997, p.87).

Some critics have even gone as far as to recommend that human attributes be abandoned as the organizing theory for the design of assessment centers and instead exercises or tasks become the focus of the assessment (Russell, 1987; Sackett & Dreher, 1982). Assessment centers are legally justifiable and they have repeatedly shown their ability to predict job performance. Perhaps assessment centers are displaying a congruency in roles between the exercises and the job rather than measuring dimensions (Sackett and Dreher, 1982). It may be easier to accurately evaluate the tasks that are performed in jobs than it is to break down the jobs into the separate components required for construct validity. In addition, developmental feedback is said to be more meaningful when based on behaviors rather than on general constructs (Howard, 1993).

The majority of researchers believe that the suggestion to abandon dimensions as the organizing framework for assessment centers is premature. Role congruency does not advance theory and “if assessment center's are to make a significant contribution to the understanding of managerial performance and organizational processes, much work needs to be done to strengthen their “constructural foundations” (Howard, 1993, p.15). All validation is ultimately construct validation and it is imperative that construct validity should remain an important objective of assessment center development and research (Tenopyr, 1977).

The use of valid and meaningful constructs are especially important to the design and conduct of developmental assessment centers. In developmental assessment centers, in particular, participants are given feedback on dimensional performance. If assessment centers are not accurately measuring these constructs, assessment center feedback may be erroneous and any developmental plans derived from the feedback could be damaging.

Reasons for the Lack of Construct Validity

Howard (1993) suggests that the lack of construct validity found in assessment centers is partially attributable to the practice of rating multiple dimensions within each exercise, which tends to equate the exercises as measures of all dimensions. She points out that the use of assessment centers in this way "does not accord with the original design of assessment centers nor with the logic of exercise development" (p. 14). The original exercises were not selected with the idea that each would adequately measure every dimension. Certain exercises were included to highlight certain dimensions. Alternative explanations for why assessment centers work have been offered including actual criterion contamination, subtle criterion contamination, self-fulfilling prophecy, performance consistency, and intelligence (Klimoski and Brickner, 1987).

Actual criterion contamination is said to exist when assessment center ratings are used in making personnel decisions such as promotions and salary increases. This practice may create an artificially high relationship between assessment center ratings and performance criteria. While there is little question that actual criterion contamination exists in some assessment centers, research by

Gaugler et al. (1987) provides evidence that there is no difference in the accuracy of predictions between assessment centers that used assessment center ratings in their judgments and those that do not. In fact Gaugler et al. (1987) found that all types of research designs (e.g., pure research studies, concurrent validation designs, studies with no feedback of ratings) give about the same estimate of predictive validity. Actual criterion contamination is not a likely explanation for why assessment centers work.

The argument for subtle criterion contamination as the reason for why assessment centers consistently yield predictive accuracy is as follows. Because assessors are often managers in an organization, they are likely to hold many of the same biases and stereotypes about what constitutes a good manager in the organization as do the actual managers who will be making the criterion decisions. These common stereotypes are hypothesized to “contaminate” the ratings, resulting in spuriously high correlations between assessment center ratings and performance criteria. An example of this type of subtle criterion contamination was provided by Guion (1987) who pointed out that “being tall” is a shared stereotype of a good police officer. In centers used to assess police officers, he found that this characteristic tended to bias both assessment center ratings and performance ratings despite the fact that being tall is completely unrelated to job performance. Although these biases obviously exist, there are several weaknesses in this argument as an explanation of assessment center predictive validity. First, assessment centers have been found to work using a wide variety of performance criteria including subordinate ratings (McEvoy & Beatty, 1989), sales

performance (Squires, Torkel, Smither, & Ingate, 1988), and judgments by third party observers (Bray & Campbell, 1968). It is highly unlikely that these findings were based on shared biases of what constitutes a good manager. A second weakness of the subtle criterion contamination argument was discussed by Thornton (1992). Often assessors are psychologists and not managers from an organization. Psychologists are unlikely to hold the same stereotypes of what constitutes a “good manager” in a particular organization. In fact, there is some evidence to support the greater predictive validity of assessment centers when psychologists are used in conjunction with managers (Gaugler et. al., 1987).

Self-fulfilling prophecy is said to exist when expectations held by supervisors about employees' performance influences employees' self-confidence and job performance. Due to the high cost of assessment centers, managers are often aware that only up-and-comers are invited to attend the centers and therefore the managers may infer that they are on the fast track of the organization. They may then perform at a higher level to meet these expectations. This effect, also termed the ‘Pygmalion Effect’ has been supported by research found in both psychology and education literatures (Eden, 1984). However, this theory does not hold as an explanation for why assessment centers work because evidence has shown equivalent predictive accuracy of assessment centers even when individuals were not told the purpose of the center and when they were not given any feedback on their performance (Gaugler et al., 1987).

In psychology it is often stated that the “best predictor of future performance is past performance”. This is the assumption underlying the performance consistency explanation of why assessment centers work. The argument is that assessors are often provided with background information on assessees that they then use to make accurate judgments of future performance. In addition, the argument suggests that even assessors who are not provided with background information are able to make accurate judgments of future behavior based on relevant exercise performance such as a work sample. Because of this, proponents of the performance consistency explanation argue that it is unnecessary to have dimensions to organize the assessment center information. To refute the first argument, there are many assessment centers where assessors were not provided with any information on the background of participants prior to the assessment. These centers have shown similar predictive validities to assessment centers where that information was provided (Thornton, 1992). The second argument is harder to refute. Because of the consistent finding that assessment center ratings tend to cluster around exercises instead of dimensions, suggestions have been made that dimensions be abandoned as the organizing framework for assessment centers. However, there is evidence that the combination of exercise ratings and dimension ratings correlates more highly with managerial success than either set of ratings alone (Wollowick & McNamara, 1969). Also, the original design of assessment centers is based on the assumption that human attributes (i.e., dimensions) can and should be measured consistently across exercises. Abandoning this framework would violate

the original purpose of the assessment center. Performance consistency may partially explain the predictive accuracy of the assessment center; however, there is little evidence suggest that it is the primary explanation (Thornton, 1992).

Finally, managerial intelligence has been proposed as the reason why assessment centers work. Intelligence tests have been shown to be predictive of management success and these tests have been correlated with assessment center ratings (Klimoski and Brickner; 1987; Thornton, 1992). While it is clear that assessment centers are related to intelligence, it is also clear that assessment centers are measuring something beyond intelligence. Supporting this claim is the Wollowick and McNamara (1969) study that demonstrated the combination of intelligence test scores and ratings of dimensions predicted progress better than either the tests or the ratings alone.

As discussed above, the majority of evidence refutes these alternative theories as viable explanations for why assessment centers work. It appears that the traditional explanation is most tenable: Assessors observe behaviors displayed in exercises, classify the behaviors into meaningful categories, make judgments of overall performance, and accurately predict measures of job performance. Thus, the problem remains that there is a lack of construct validity evidence to support this explanation.

Sackett and Dreher (1982) have proposed that the limited construct validity may be due to a halo effect. This hypothesis has been elaborated on by Thornton (1992). Assessors may form a general impression that a participant is “doing well” or “doing poorly” and this general assessment is successfully predicting job performance or

promotability. Sackett and Hakel (1979) performed a policy-capturing study on individual decision processes in assessment centers and found that assessors did not differentiate dimensions. In fact, they found assessors' dimensional ratings to be dominated by a single, common underlying factor.

One potential cause for halo effects is that assessors may have a limited capacity to process information and that the greater the complexity of the judgment task, the more prone it will be to cognitive biases (Bycio et al., 1987). Assessors are often asked to rate too many dimensions and the dimensions are typically complex, resulting in more cognitive complexity than the average person can handle. Assessors are frequently asked to rate multiple dimensions within each exercise, often resulting in a dimension x exercise grid where each dimension is measured in each exercise. Limiting the cognitive demands placed on assessors should help to reduce these biases.

A second and related reason for halo effects may be an insufficiently large sample of observable behavior (Cooper, 1981; Kleinmann & Koller, 1993). Many exercises do not sample enough behaviors to provide evidence for multiple dimensions, and assessors end up basing their judgments on only a few indicators (Bycio et al., 1987). Also, the same behavior may be used to infer more than one dimension. Lack of observable behavior will have consequences for both convergent and discriminant validity. More observable behaviors can enhance discriminant validity because lower correlation coefficients for various dimensions within an exercise are more likely when the indicators of dimensions can be observed sufficiently. As Gaugler and Thornton

(1989) point out, “In order for convergent coefficients to reach a meaningful level, it is necessary for dimension-relevant behavior to be observable in various exercises” (p. 616).

True exercise effects have also been hypothesized as contributing to the findings that ratings tend to cluster around exercises instead of dimensions. Ability may overlap with the exercise such that certain individuals may perform better in one-on-one situations, group discussions, or on various individual exercises. In addition, some people may perform better in competitive situations while others perform better in cooperative situations. (Neidig and Neidig, 1984). In these cases, halo error may, in part, represent the true performance differences between exercises.

Halo effects would also be evident if dimensions had high intercorrelations. Dimensions that are tapping similar constructs will result in individuals that perform similarly on multiple dimensions, resulting in a lack of discriminant validity.

As discussed above, the most frequently given reasons for weak construct validity findings are that assessors are cognitively overburdened, insufficiently large samples of observable behavior for dimensions, true exercise effects, and dimensions with high intercorrelations. In all likelihood the lack of construct validity is due to a combination of all the above reasons. One method for isolating the reasons for low construct validity in assessment centers is to investigate variables that are hypothesized to effect construct validity in functioning

assessment centers. In recent years, small improvements in construct validity can be seen using this method.

Gains in Construct Validity

Several types of methodological improvements have boosted construct validity. The use of retranslated behavioral checklists helped standardize assessor observations and resulted in increased construct validity (Reilly et al., 1990). The retranslation procedure was as follows: Assessors were trained using conventional assessment center training and rating scales. After each exercise, assessors rated the two candidates assigned to her or him on each dimension using a 1-5 scale. At the end of the day, raters stated their ratings to the other assessors for each dimension and each exercise but they did not engage in consensus discussions. Assessors were asked to provide behaviors “that, when they occurred, caused them to judge an assessee as being higher or lower” (p. 74) in the particular rating category (i.e., dimension) for each exercise. Behaviors were then retranslated into their relevant dimensions. For each exercise, a criterion of 80% agreement was used to select behaviors for each dimension. The resulting behaviors were organized into final lists by dimension within each exercise. A second assessment was then conducted with assessors using the retranslated checklists to code behaviors on a new sample of assessees. Reilly et al. (1990) hypothesized that by focusing assessors on retranslated behavior, three problems would be reduced: 1) cognitive demands on the assessor, 2) the problem of operational definitions and, 3) the degree to which exercises elicit dimension-relevant behavior would be clear.

Assessor use of behavior checklists in the Reilly et al. (1990) study increased the average convergent validity from .24 to .43 while simultaneously decreasing the average heterotrait-monomethod correlations (i.e., discriminant validity) from .47 to .41. After using behavioral checklists convergent validity was slightly higher than the discriminant validity, a relationship not found in many studies. In fact, in most of the studies previously discussed the level of discriminant validity was almost twice as high as the convergent validity (Bycio et al., 1987; Sackett & Dreher, 1982). Reilly et al. (1990) hypothesized that the increase in construct validity was most likely due to a reduction in the cognitive demands of the assessor resulting from the use of the retranslated behavioral checklists. Reilly et al.(1990) concluded that assessors are able to make more construct valid judgments when they have the opportunity to observe adequate amounts of behavior relevant to the dimensions.

In a related study, Donahue et al. (1997) compared the use of untranslated behavioral checklists (i.e., developed without the use of a retranslation procedure) and graphic rating scales on the construct validity of assessment centers. They found the mean heterotrait-monomethod correlations were .40 for the behavioral checklists and the mean heterotrait-monomethod correlations were .61 for the graphic ratings scales, indicating an increase in discriminant validity with the use of the behavioral checklist. However, they did find slightly poorer convergent validity for the assessment centers using the behavioral checklists. Similar to previous research, they found greater evidence for exercise factors rather than

dimension factors for both methods. The greater convergent validity found in the Reilly et al. (1990) study can be partly attributed to the finding that there is greater evidence of convergent validity across exercises that are similar in form (Schneider and Schmitt, 1992). In the Reilly study, they investigated the convergent validity of two group exercises, both of which involved an assembly problem.

Thornton, Tziner, Dahan, Clevenger, & Meir (1997) also found considerable evidence of construct validity in their assessment center using the Behavioral Reporting Method (also known as the AT&T method). In this method, no dimension ratings are given at the exercise level. In the integration discussion, assessors report only behaviors relevant to each dimension, but not ratings. Assessors then make comparisons of behaviors relevant to each dimension across exercises. Assessors give dimension ratings only after behaviors are reported across all exercises. Results indicated that ratings on 16 dimensions correlated with independent measures of comparable constructs measured by ability tests and by psychologists' assessments based on interviews and personality test scores. In addition, a factor analysis of their data showed meaningful groupings that conformed to some degree to the a priori categories (i.e., dimensions). The authors suggest that the significant construct validity found in the study may be partially due to the problem of halo error when assessors are asked to make ratings of dimensions at the exercise level. This scenario may lead assessors to think in terms of overall exercise performance, causing observers to generalize from some overall impression of doing well or doing poorly on

each exercise in the evaluation of all the dimensions. This problem is avoided with the use of the Behavioral Reporting Method.

Advances in the construct validity of assessment centers have also been made in relation to assessment center dimensions. It helps to focus consensus discussions on dimensions rather than exercises (Silverman et al., 1986). In a study of the number of assessment center dimensions as a determinant of assessor accuracy, Gaugler and Thornton (1992) found greater evidence of convergent validity when three instead of six or nine dimensions were evaluated. However, number of dimensions did not affect the accuracy of assessors' observations or the discriminant validity of their dimension ratings. They concluded that the cognitive complexity of the rating task is reduced when a smaller number of dimensions are assessed, resulting in greater convergent validity.

In another study dealing with the dimensions used in assessment centers, Kleinmann (1993) investigated whether the extent to which participants recognize rating dimensions in assessment centers has an effect on performance. He found that assessment center dimensions lack transparency for the participants. However, he did find an increase in convergent validity of dimension ratings when participants accurately perceived that the same dimension was being evaluated in 2 exercises.

In a study looking at both the influence of type of assessor and characteristics of dimensions on the construct validity of assessment centers, Shore, Shore, and Thornton (1992) examined the construct validity of self and peer evaluations by investigating their relationships to

conceptually similar and dissimilar constructs derived from cognitive ability and personality measures. They found that peer evaluations predicted job advancement better than self-evaluations and that evidence for construct validity was stronger for peer than for self-evaluations. For both evaluation sources, stronger support was shown for the more observable assessment dimensions. The authors concluded that peer and self evaluations will be most useful when they focus on dimensions for which participants have greater amounts of behavioral information on which to base their judgments.

Sagie and Magnezy (1997) also investigated the effect of assessor type (i.e., psychologist or manager) on assessment center construct validity. They found that assessor type did influence construct validity. Using confirmatory factor analysis, they found psychologists were able to distinguish all five predetermined dimension categories, whereas managers were only able to distinguish two dimension categories. This finding lends support to the most recent assessment center meta-analysis (Gaugler et al., 1987) which found that criterion related validity was higher when both psychologists and managers were used as assessors rather than managers only.

From the above discussion, many of the increases seen in construct validity were due to changes in rating procedures, exercise design, rater characteristics, and dimension characteristics. One of the most promising and unexplored ways to increase construct validity is through a closer examination of dimensions and dimension characteristics. In her paper, A Reassessment of Assessment Centers: Challenges for the 21st Century (1997), Ann Howard

makes several recommendations for how to increase construct validity through the selection and use of dimensions. First, she suggests selecting fewer but more observable dimensions. Second, cover the important domains but don't expect sharp differentiation of dimensions that are subtle variations on the same theme. Third, rate dimensions only after you have enough behavioral evidence to do so, and fourth, define dimensions clearly and unambiguously and add key behaviors for specificity and structure. Similarly, Thornton (1992) suggests in order to maximize the usefulness of an assessment center, designers should select between five and seven of the most ‘observable’ dimensions to be evaluated.

While these recommendations may seem useful, they are too vague to result in increased construct validity. For example, neither Howard nor Thornton define what they meant by ‘more observable'. As I will discuss in a moment, the definition of observability is subject to numerous interpretations. In Howard's (1997) third recommendation, it is not clear what constitutes enough behavioral evidence. In the fourth suggestion, she recommends clear dimension definitions and the addition of key behaviors for specificity and structure. In the same article, she suggests that a future research question in the assessment center literature should be: “Can key behaviors enhance the accuracy of dimension ratings and boost convergent and discriminant validity?” (p. 43). She does not specify the definition of key behaviors; however, it is possible that she is referring to those behaviors that are particularly diagnostic for a dimension. Dimensions that are represented by these key (i.e., highly diagnostic) behaviors are likely

to be easier for assessors to evaluate accurately, resulting in increased construct validity of those dimensions.

Observability of Dimensions

As discussed above, some researchers have argued that characteristics of dimensions are partially responsible for the lack of construct validity found in assessment centers. One discussed, but relatively unexamined characteristic is the observability of dimensions. A review of the literature reveals that, when discussed, the concept is rarely defined and has little basis in theory. Only one study was found in which observability of dimensions was examined empirically (Shore et al., 1992).

One potential reason for the lack of research in the area has to do with the confusion regarding the definition of 'observability'. Shore et al. (1992) defined observable as more behavioral and less abstract. In their discussion of the importance of observability of dimensions, Kudisch et al. (1997) defined observable dimensions as being 'more overt'. Howard (1997) stresses the importance of having a few, highly observable dimensions in an assessment center but she does not offer a definition of the concept. Similarly, a number of authors have discussed the observability of dimensions without bothering to offer a definition (Turnage and Muchinsky, 1982; Reilly, 1990; Kleinmann, 1993). Some researchers allude to observability issues when they are discussing a related concept, the specificity of dimensions.

Thornton (1992) discusses the specificity of dimensions and suggests that there are four categories of dimensions-moving from general to specific, they are 1) classes of dimensions (e.g., communication, decision making), 2)

dimensions (e.g., oral communication, problem analysis), 3) situational dimensions (e.g., quantitative analysis, staffing analysis), and 4) subdimensions (e.g., technical translation, fact finding). While he does not imply that specificity of dimensions is the same concept as observability of dimensions, there does appear to be some overlap. For example one could argue that the subdimensions are the most behavioral dimensions and that the classes of dimensions are the most abstract.

A different organizational framework for the specificity of dimensions was suggested by Hampson et al. (1986). They suggest a 3-level model of dimensional abstraction. They found that for most purposes the top level, involving broad traits such as extroversion is considered too abstract, whereas the bottom level, involving traits such as musical or stingy is too specific. The middle level involving concepts such as energy level, sociability, reliability and assertiveness seems to be most useful for assessment centers. Again, these descriptions do not seem independent of what has been referred to as the ‘observability' of dimensions. If ‘observability' is defined as more or less abstract (as defined by Shore et al.), there appears to be little difference between the specificity and observability of dimensions, adding even more ambiguity to the literature on the definition of observability.

Discussion of the importance of the observability of dimensions can be found in the discussion section of many articles. For example, Turnage and Muchinsky (1982) claim that the use of more basic behavioral dimensions may help to enhance the power of the assessment dimensions. Reilly et

al. (1990) claim that definitions of dimensions are typically written in general terms and are not always clearly related to the operational definitions (i.e., behaviors elicited by the exercises). Without clear operational definitions, it is a much more cognitively challenging task to decide which behaviors belong to which dimensions. Kleinmann (1993) concluded his article by stating that researchers who want to take account of the halo effect should construct assessment centers in which dimension relevant behavior is highly observable in exercises. Bycio et al. suggest that cross-situational consistency may be easier to attain in an assessment center when highly publicly observable dimensions are used. They suggest that this reasoning may explain why oral communication (they believe most observable in their study) had larger ability variances than other dimensions. They conclude the article by emphasizing the importance of obtaining multiple ability-related observations within each exercise.

Kudisch et al. (1997) came to a similar conclusion in their study of the construct validity of diagnostic assessment centers. They found that cross-situational consistency may be easier to attain when skill dimensions are more overt, or easier to observe. In the study, they found that those dimensions that can be considered most easily observed (e.g., oral communication and written communication) produced the greatest convergent validity in the multitrait-multimethod matrix. These dimensions also produced some of the best CFA factor loadings-- another measure of convergent validity (Widaman, 1985). Those dimensions that were less observable (i.e., less overt) such

as analysis/judgment produced some of the weakest construct evidence. There was no difference in discriminant validity between the observable and less observable dimensions, although the researchers did find the correlations of different dimensions within an exercise were relatively lower than most prior studies (i.e., discriminant validity). Despite this finding, the discriminant validity for this assessment center was weak. While the observability findings are compelling, these distinctions were not hypothesized a priori and observability was not rated by independent judges; therefore, results must be interpreted with caution and the phenomenon deserves further investigation.

Shore, Shore, and Thornton (1992) were the only researchers to empirically test the observability of dimensions on construct validity of an assessment center. The authors hypothesized that the information available to assessment center participants may affect their ability to make valid performance judgments. They predicted that peer and self-evaluations of more easily observable dimensions would be more strongly related to criterion measures than would dimensions requiring a greater degree of inferential judgment. Observability of dimensions was rated by two independent groups of judges who were given only dimension definitions. The dimensions classified as most observable were ‘most active in business discussions’, ‘most persuasive’, and ‘expresses ideas most clearly’. The least observable dimensions were classified as ‘most original’, ‘most likeable’, and ‘can work with least direction’.

Shore et al. (1992) found strong support for greater construct validity for the more observable dimensions and

hypothesized three potential explanations for the results: 1) the more easily observable dimensions were less cognitively demanding of raters than were the less observable dimensions, 2) because observable dimensions are more behavioral, participants may view them as more easily verifiable by raters. Participants may represent themselves more accurately when they know the raters will also be evaluating them on the behavioral dimensions, and 3) the observable dimensions may appear to assessees to be more critical to being a manager than the more abstract dimensions and therefore, participants are likely to focus on the behaviors representing these dimensions. The researchers concluded that "peer and self-evaluations will be most useful (i.e., more valid) when they focus on dimensions for which participants have greater amounts of behavioral information on which to base their judgments" (p.52).

A review of the references to characteristics of dimensions (e.g., observability, specificity) in the assessment center literature reveals inadequate definitions of observability and related constructs. It is often unclear what specifically researchers are attempting to measure. While Shore et al. appear to be defining observability as the amount of behavioral information available for a dimension, this is not the only interpretation of observability. Observability of a dimension could also be interpreted as dimensions that are more or less inferential, the diagnosticity of the behaviors that represent the dimension, some combination of both, etc. As discussed above, few researchers even attempt to define the construct of observability. In order to gain any

useful information from this line of research, it is imperative that we clarify what the construct or constructs are that we are attempting to measure.

DIAGNOSTICITY

An alternative construct to observability of dimensions that may be more directly related to the validity of assessment center's can be termed ‘ease of evaluation’. In other words, it is not the amount of behavioral information that is important, rather, the diagnosticity of that behavioral information that is most meaningful. In order to be diagnostic, a behavior should be probable for a person with the hypothesized disposition but improbable for someone without the disposition. For example, one dimension may have numerous behaviors generated under it; however, each of those behaviors may only be of limited diagnostic value for that dimension. On the other hand, we may find a dimension that has few behaviors that are relevant to the dimension. However, each relevant behavior has a high degree of diagnosticity for an individual's placement along the dimension. As discussed earlier, Howard's (1997) references to key behaviors representing a dimension can be interpreted as referring to the diagnosticity of those behaviors. The more diagnostic the behaviors are for a particular dimension, the easier that dimension will be for an assessor to evaluate it accurately.

Support for the importance of the differential diagnosticity of behaviors comes from the attribution literature. Trope (1986) developed a two-stage model of dispositional judgment that is directly applicable to assessment center research. He suggested that the first stage is an identification stage in which behaviors are

observed and categorized in dispositional or dimension relevant terms. The second stage is an inductive inference stage in which previously identified behavior is used to determine whether the target person has the corresponding disposition. In the first stage, individuals will take into account the characteristics and constraints of the situation. The second stage of the process involves the assessment of the diagnostic value of the identified behavior. According to Trope and Liberman (1993), the diagnosticity of a behavior depends on two kinds of behavior probabilities- a) the probability that a situation would produce such behavior given that a target has the hypothesized disposition and b) the probability that the situation would produce the behavior given that the target does not have the hypothesized disposition.

A direct relationship can be drawn between Trope's model and the process of assessment center ratings. For example, if we are assessing the aggressiveness dimension, we would a) examine whether in a particular situation (i.e., exercise) we would expect an aggressive person to react aggressively and b) whether a non-aggressive person would react aggressively in the same situation. If a is yes and b is no, and in the absence of previous disconfirming evidence, we will infer that the person is high on aggressiveness. If a is yes and b is yes, the behavior is not very diagnostic for the dimension and we will not make an inference about the individual's placement on the aggressiveness dimension. Again, in order for a behavior to be considered diagnostic, it should be probable for a person with the hypothesized disposition but improbable for someone without the disposition.

The Assessment Center Context

The unique design of assessment centers may be an ideal place for accurately inferring dispositions from behavior. In the assessment center context assessors are attempting to infer an individuals standing on a number of dimensions or traits by observing behavior under a variety of conditions.

Because assessment centers are usually made up of multiple situational exercises in a constrained setting, it has been suggested that assessment centers represent “strong” rather than “weak” situations (Highhouse & Harris, 1993). Assessment centers are designed to present targets and observers with a high fidelity simulation of the stimuli experienced in jobs for which assessees are being judged. Some researchers have argued that strong situational inducements will attenuate the diagnostic value of a behavior by making the behavior seem probable regardless of whether the target person possesses the hypothesized disposition (Trope & Burnstein, 1975). However, these researchers admit that in the presence of strong intrinsic inducements, the occurrence of behavior may still retain some diagnostic value regarding the actor's dispositions. For example, a popular assessment center exercise is the role-play. A typical scenario might involve the participant playing the role of a supervisor who is asked to have a meeting with a disgruntled employee. The employee, a confederate, closely follows a script in which he or she is not calmed by anything the participant says and the employee remains confrontational. In this situation, it should be clear to most participants that one purpose of this particular exercise is to test the patience of the assessee and to determine whether the participant can remain calm in

the face of conflict. In this scenario, when participants become verbally or physically hostile toward the employee, the behavior would be considered diagnostic.

More recently, Trope and colleagues (Trope & Cohen, 1989; Trope, 1986; Trope, 1989; Trope & Liberman, 1993) have consistently demonstrated that strong situations will make the process of accurately inferring dispositions simpler. They argue that in the presence of either a strong inhibitor or a strong inducement, a behavior will be highly diagnostic because the disposition is necessary for the behavior to occur. Situational biases in behavior identification and failure to correct for such biases may offset inferential adjustment. Perceivers may know that person X was provoked and may properly adjust their inferences regarding X's dispositional hostility. However, the provocation may bias perceivers toward identifying X's behavior as hostile even when it is not. Reliance on this identification will lead perceivers to potentially attribute more hostility to person X when he is provoked than when he is not provoked.

In strong situations, people will also obtain diagnostic information when the behavior runs counter to what is being induced by the situation. For example, in a group discussion, a typical assessment center exercise, individuals may be explicitly told to take their time, make sure they weigh all the evidence and told not to rush into a decision. Given these instructions, when an individual proceeds to make swift decisions and spends little time contemplating the issues, we can reasonably infer that this person may be high on the dimension of decisiveness. There was nothing in the situation that should have induced the individual to behave this way and therefore we can

reasonably conclude that the behavior was due to something within the person. In support of this argument, Trope, Cohen, & Alfieri (1991) found that situational ambiguity will attenuate the effect of the behavior on dispositional attribution. It could be argued that the situations seen in assessment centers are not typically ambiguous.

Further support for this hypothesis was found by Shoda, Mischel, & Wright (1989). In a study of the effects of situation-behavior relations on dispositional judgments, they found that predictability of dispositions increases with explicit situational information. They go on to argue that if an assessment tool is to be useful, it must incorporate such information. Because of the power of constrained situations on the ease of inferring traits, it is not wonder that assessment centers have been so successful in predicting job performance and promotion.

Further, assessment centers have a high degree of experimental realism. Experimental realism is defined as the degree to which an experiment absorbs and involves its participants (Myers, 1998). Dispositional judgments are likely to be more accurate when they are made in a naturally occurring situation in which people are actively involved (Shoda et al., 1989). The use of exercises such as role-plays, group discussions, and in-baskets help to ensure that the assessment center will elicit active participation from assessees.

The diagnosticity of a particular behavior will be reduced if there is uncertainty regarding the identification of that behavior. Referring to Trope's model, the first stage is identification of the behavior. If it is not clear what behavior was performed, it will not be possible to

accurately infer an individual's position on the relevant trait. In addition, the strongest inferences will be made regarding an actor's dispositions when neither extrinsic nor intrinsic inducements can account for his or her behavior. For example, an assessment center is attempting to measure initiative. A participant demonstrates no evidence of initiative in the assessment center across multiple exercises even though the individual is instructed to develop new ideas, actively participate in discussions, etc. Assuming this individual is motivated to perform well in the assessment center (e.g., it is for selection into a job), the lack of initiative behaviors that are shown can reasonably be inferred as demonstrating that this person is low on the initiative dimension.

Behavior-Trait Inference

Is it possible for people to make correct and accurate dispositional inferences of others in a limited time frame? The overwhelming evidence suggests that it is possible (Trope, 1989; McArthur & Barons, 1983; Trope et al., 1991). Trope suggests that personal dispositions are often directly observable and that human perceptual processes are sufficient to explain dispositional attribution. In fact, with practice, dispositional inference operations may become automatized (Trope & Bassok, 1982). Consequently, the attribution of a disposition to another person may be experienced as a direct, immediate perception. Basically, people are remarkably good at inferring dispositions based on behavioral information. Relating to assessment centers, this implies that once assessors have spent significant amounts of time inferring traits or dispositions, they may get more accurate at inferring an individual's standing on a

particular dimension. Assessors become expert at this judgment task. In fact, as discussed earlier, this hypothesis may explain meta-analytic findings that psychologists often yield higher predictive validities than managers (Gaugler et al., 1987).

Impact of Prior Behavioral Evidence

Diagnosticity literature is useful for describing how an individual is able to make dispositional inferences based on only a few diagnostic behaviors. However, in assessment centers, assessors are asked to make judgments for a dimension based on multiple behaviors observed across multiple exercises. Are assessors able to combine multiple diagnostic behaviors to form accurate impressions? Again, the research shows that the answer is yes (Trope, 1989; Trope & Burnstein, 1979; Shoda et al., 1989)

When an immediate behavior is the only source of information about a target, inferences regarding the hypothesized disposition derive entirely from the diagnostic value of that behavior. However, prior information about the assessee may take the form of social stereotypes, the target person's appearance, the opinions of others, and prior behavioral evidence. Prior beliefs based on any of these sources are integrated with the diagnostic value of the immediate behavior at the inference stage.

The impact of prior behavioral evidence will increase to the extent that the number of prior behavioral observations is large and the relative frequency of the relevant behavior is high. In addition, if the behavioral observations are nonredundant (i.e., the behaviors are different and are observed under varied circumstances) and each behavior is diagnostic as defined by Trope's two stage

model (Trope & Cohen, 1989) the behaviors will have the greatest degree of impact.

The influence of prior behavioral evidence increases with the number of past behaviors and with the relative frequency with which those behaviors were diagnostic of the hypothesized trait (Trope & Liberman, 1993). The influence of prior behavioral evidence decreases with the redundancies among past behaviors. Redundancy reflects the similarity among the behaviors and the circumstances in which they occur. For example repeating the same kind of friendly behavior (e.g., smiling) toward the same person is more redundant than expressing different kinds of friendly behavior (e.g., smiling, hugging, starting a friendly conversation) toward different people (Trope & Liberman, 1993).

Supporting the argument that redundant behaviors should not be as diagnostic as nonredundant behaviors is the interactionist approach to dispositions (Trope & Cohen, 1989). Proponents of this view argue that it is not simply the number of behaviors representing a dimension or a disposition, it is the correspondence of the behavior to the situation that is important. This theory assumes that different situations are not psychologically interchangeable and therefore behaviors can not be (and are not) simply aggregated. In assessment centers, this translates to the belief that the same behavior exhibited in two different exercises may have different diagnostic value.

Further, some researchers (Zedeck, 1986) have argued that the common finding of an exercise factor solution could be due to halo error and the practice of making ratings on several dimensions based on the same behavior.

Justification for this hypothesis comes from a study by Brannick et al. (1989). They found that poor performance on a "red hot" item resulted in negative ratings on five separate dimensions. Vivid information may have a disproportionate influence on human judgment because it evokes a rich associative network. These associations influence the manner in which information is processed (Isen & Diamond, 1989)

Thus, diagnosticity of behaviors must be taken into account in the search for greater construct validity of assessment centers. Diagnostic behaviors will serve to increase the accuracy of the behavior-trait inference for a particular dimension. Thus, it may be assumed, the more highly diagnostic behaviors a dimension has representing it, the easier the evaluation of the dimension will be for the assessors, resulting in greater construct validity for the assessment center.

Researchers such as Shore et al., (1987) that claim the observability of a dimension is equivalent to the "amount of behavioral information" available for a particular dimension are ignoring the importance of differential diagnosticity of the behaviors observed. A re-examination of the Shore et al. study shows that the 'ease of evaluation' of the dimensions could have been confounded with their measures of observability, which they defined as 'more behavioral'. Because of the imprecise measurement of observability in the study (i.e., the use of untrained raters who were unfamiliar with the assessment center exercises), I would argue that a potential reason the observable dimensions resulted in greater construct validity is because the dimensions rated as observable were simply easier to evaluate more accurately

(i.e., had a greater proportion of highly diagnostic behaviors) rather than the explanations given by the authors. As Trope's two-stage attribution model suggests, there is more cognitive effort involved when a perceiver is forced to resolve inconsistencies in the information than when inferences can be easily and routinely made. Following this logic, one can assume that cognitive demands will be reduced when an assessor is faced with highly diagnostic behaviors for a dimension, resulting in an easier and more accurate evaluation of that dimension.

Assessment Center Literature on Diagnosticity

Although the topic of diagnosticity of behaviors or 'ease of evaluation' of dimensions has not directly been addressed in the assessment center literature, a few studies have investigated similar concepts without using the same terminology.

The Reilly et al. (1990) study was discussed above. In this study, assessors were asked to provide behaviors, "that, when they occurred, caused them to judge an assessee as being higher or lower" in the particular dimension for that exercise. These behaviors were retranslated into their relevant dimensions. A criterion of 80% agreement was then used to select behaviors for each dimension within each exercise. The behaviors that passed this criterion were used to form a checklist of behaviors for each dimension within each exercise for future assessors. This procedure demonstrates the use of diagnostic behaviors for increasing the construct validity of the assessment center although that was not explicitly stated. In fact, the increase in validity that was found with the use of this retranslation procedure provides clear support for the hypothesis that

making assessments with diagnostic behaviors will make the evaluation of a dimension easier and will result in increased construct validity. Although the number of behaviors generated under one dimension ranged from 4 to 32, Reilly et al. (1990) did not make any hypotheses concerning the relative construct validity of the different dimensions.

In a recent article titled, A Person Perception Explanation for Validation Evidence on Assessment Centers, Jones (1997) included person perception theories in the search for answers to the assessment center construct validity dilemma. While he does not deal with the diagnosticity of behaviors directly, he does argue that assessment center's predictive accuracy stems from the use of motive-based traits which, while not often formally included in assessment centers, provide rich information about assessees. He cites Howard's (1993) finding that a "need for advancement" motive predicted advancement and various life satisfaction measures and Whitmore's (1995) findings that "career motivation" judgments predicted advancement and developmental activity when all other ratings were accounted for. He claims that assessors are able to observe a wide range of possible behaviors within a relatively constant setting and thus, they are able to make determinations as to the type of motives at work when a behavior is observed.

While it may not always be obvious to assessors what a participant's goals are for a particular behavior, the relatively constrained situation of the assessment center allows assessors to make reasonably accurate judgments of target motive traits. Similar to his two stage attribution model, Trope is suggesting that for some behaviors assessors

are able to determine the cause or motivation of the behavior, providing crucial information about the assessee's true performance on a particular dimension. The lack of explicit inclusion of assessee motive traits in assessment center dimensions may not discourage experienced assessors from basing their judgments on the motive information. This practice may lead assessors to provide accurate predictions of future organizational performance and/or advancement in the organization, resulting in assessment centers that have high predictive validity. This practice may simultaneously introduce noise into the ratings because some assessors may be more skilled at inferring motive-based traits than others, potentially resulting in an increase in common method variance and a lack of construct validity evidence for the method.

In assessment center exercises, some behaviors may be performed by all or most participants, whereas other behaviors may be performed by only a few individuals. In addition, some behaviors may be performed frequently while others only occur once. Most importantly, some behaviors tell us a great deal about an individual's standing on a particular dimension, whereas other behaviors tell us little, if anything about an individual's performance on the dimension. It is likely that assessors integrate this diagnosticity information for the behaviors into their ratings in addition to the frequency of such behaviors. The assessment center literature has failed to investigate this possibility as a reason for the lack of construct validity that has consistently been found in assessment center research.

SUMMARY AND HYPOTHESES

Because of the potential for improving construct validity, ease of evaluation of dimensions warrants further investigation. The current study will attempt to clarify the role of ease of evaluation of dimensions in an assessment center and the relationship to assessment center construct and criterion related validity.

The present study will add to the literature on assessment centers in the following ways. First, the study will clarify the operationalization of ease of evaluation of dimensions. When observability of dimensions has been defined in the literature, it has been a broad definition such as ‘overt’ or ‘behavioral’. Ease of evaluation is a more specific and more accurate reflection of the construct that is likely to be influencing the validity of assessment centers. In the current study, ease of evaluation is operationalized as the proportion of highly diagnostic behaviors representing a dimension.

Second, the current study will use a pre-study with the exercises used in the assessment center to evaluate the ease of evaluation of each of the dimensions. Shore et al. (1987) used I/O graduate students who had little or no assessment center experience as raters nor did they have any familiarity with the exercises used in the center. The determination of whether a dimension was more or less observable was made by these judges who only had access to the general dimension definitions. The present study will improve upon this methodology by using trained assessment center raters who are familiar with the assessment center exercises to code behaviors and to determine the diagnosticity of those behaviors.

Third, the current assessment center was used for selection of production workers, resulting in dimensions in the current assessment center (e.g., quality, problem assessment, influence) that are considerably different than the typical managerial dimensions seen in most assessment centers (e.g., decision making, leadership, planning and organizing). The dimensions that were evaluated in the current study tend to be at a more specific level than managerial dimensions. Construct validity of this type of assessment center has yet to be examined in the literature.

Fourth, the current assessment center was rigorously designed to conform to most or all recommendations set forth by the 17th International Congress on the Assessment Center Method. One common complaint in the assessment center literature is that often dimensions are measured in exercises where there is little or no opportunity for dimension-relevant behavior to be displayed. In the past, all dimensions were measured in all exercises. Recommendations have been made to only measure dimensions in exercises where there will be adequate opportunities to observe relevant behavior; however, this has not become common practice. The current assessment center measured each dimension in at least two (and at most three) of only the most relevant exercises.

Fifth, dimension specific performance data are available which match up with 5 of the 7 assessment center dimensions. In his review of the literature on assessment centers, Thornton (1992) came to the conclusion that "No studies to date have examined the ability of within-exercise dimension ratings to predict measures of the same attributes in work settings" (p. 119). Dimension specific performance

data will allow a close examination of the predictive validity of individual dimensions.

Sixth, with the exception of the study by Chan (1996), few studies have looked at both criterion and construct validity of an assessment center simultaneously. The present study will explore both types of validity to gain a more complete picture of how ease of evaluation of dimensions influences validity. Finally, within exercise dimensions ratings of performance will be used to assess the existence of an exercise versus dimension effect and to examine both convergent and discriminant validity.

Hypotheses

Based on the literature on assessment centers, the existing literature on the relationship of observability of dimensions to construct validity and the person perception literature on diagnosticity, the following 5 specific hypotheses are proposed with regard to the ‘ease of evaluation’ construct:

1) High ease of evaluation dimensions will show higher convergent validity than low ease of evaluation dimensions

2) High ease of evaluation dimensions will show greater discriminant validity than low ease of evaluation dimensions.

3) High ease of evaluation dimensions will show a greater average criterion-related validity for dimensional performance than will low ease of evaluation dimensions

4) High ease of evaluation dimensions will demonstrate greater criterion-related validity for performance than will low ease of evaluation dimensions

5) The subset of low ease of evaluation dimensions will not demonstrate incremental validity above the validity found when using the high ease of evaluation dimensions.

METHOD

Participants

Assessees (N=1788) were selected for one of two jobs (i.e., production associate or equipment services associate) in a large car manufacturing company in the Southeastern United States. The sample included 1166 males (65%), 395 females (22%), 227 unknown (13%); 1243 whites (80%) and 318 minorities (18%). A second sample was made up of a random subset of 280 of the above assessees that were hired and were later evaluated by their supervisors on their performance. This sample included 216 males (77%), 64 females (23%); 220 whites (79%), and 60 minorities (21%). There were no significant differences between the two samples on any of the independent variables (e.g., dimension scores, exercise scores).

Data Description

Archival data were obtained from a large Industrial/Organizational Consulting firm specializing in assessment centers. In addition to the predictor dimensions listed below, both datasets contained information including an employee ID number, gender, race, veteran status, assessment center date, and date of hire. Each dataset contained both within exercise dimension rating scores and exercise scores. Only the smaller of the two datasets contained data on performance dimensions. The data were anonymous and individuals could not be identified individually. All analyses were done at the aggregate level.

Dimensions and Exercises

Dimensions. 1) *Quality orientation* was defined as accomplishing tasks through concern for all areas involved,

no matter how small; showing concern for all aspects of the job; accurately checking processes and tasks; maintaining watchfulness over a period of time. Sample key behaviors included attending to all details of the job and accurately checking processes or work outputs.

2) *Work Pace* was defined as performing work at a specific pace without unnecessary expenditures of time or waste of supplies and materials; demonstrating a consistent rate of speed for accomplishing activities in a specific order. Sample key behaviors included performing at a consistent and appropriate speed and performing work with high accuracy.

3) *Problem Assessment* was defined as securing relevant information and identifying key issues and relationships from a base of information; relating and comparing data from different sources; identifying cause-effect relationships. Sample key behaviors included gathering information, organizing information, and anticipating potential problems.

4) *Problem Solution* was defined as committing to an action after developing alternative courses of action that are based on logical assumptions and factual information and that take into consideration resources, constraints, and organizational values. Sample key behaviors included developing and considering alternatives, selecting a course of action, and being decisive.

5) *Influence* was defined as using appropriate interpersonal styles and methods to inspire and guide others toward goal achievement; modifying behavior to accommodate tasks, situations, and individuals involved. Key behaviors included making recommendations that are sound and have impact, and developing and considering alternatives.

6) *Meeting participation* was defined as using appropriate interpersonal styles and methods to motivate and guide a meeting toward its objectives; modifying behavior according to the tasks and individuals; being aware of the needs and potential contributions of others. Sample key behaviors included making procedural suggestions, summarizing information, and soliciting the ideas of others.

7) *Teamwork* was defined as active participation in, and facilitation of, team effectiveness; taking actions that demonstrate consideration for the feelings and needs of others; being aware of the effect of one's behaviors on others. Sample key behaviors included acknowledging others' concerns and contributions and clearly communicating relevant ideas.

Exercises. Six simulations- two production exercises, two group discussion exercises, two problem-solving (analysis) exercises- were included in the assessment center.

1) *Production Exercises*- In these simulations the participant was asked to perform a series of production-focused exercises. The first production exercise involved mounting tire rims on axles using a specific set of procedures. The second production exercise involved inspecting parts for specific defects using a specific set of procedures. The dimensions assessed in these exercises are Work Pace and Quality Orientation.

2) *Group Discussion Exercises*- In these simulations four to six participants were asked to discuss issues and situations, reaching consensus on recommendations or solutions. Both the group discussion exercises were designed to place candidates in situations in which they are

a member of a team put together to provide recommendations about a particular issue within a fictitious organization. Next, candidates gathered to hold a discussion about the issue and come to agreement about solutions and recommended actions. The dimensions assessed in these exercises are Influence, Meeting Participation, Teamwork, and Problem Solution (second group discussion exercise only).

3) *Problem Solving-Analysis exercises*- In the first problem solving exercise, participants were asked to look into a fictitious organization's operations. The participant was given information about the organization and asked to study the data in an attempt to gather pertinent information that will allow him or her to make sound recommendations to resolve operational difficulties. In the second analysis exercise, participants were asked to develop process improvement ideas in a production situation. The dimensions assessed in these exercises are Problem Assessment and Problem Solution. A dimension/exercise coverage grid can be seen in Table 1.

Predictor variables

Ease of evaluation of dimensions was the independent variable. It was operationalized as the proportion of high diagnostic behaviors representing a dimension. Diagnosticity ratings were generated by 6 Industrial/Organizational graduate students who had previous assessment center experience. Diagnosticity ratings were made for lists of key behaviors that were given to raters prior to rating the assessment center candidates in 1993. For each behavior, the 6 raters were asked, “When a behavior occurs in the assessment center context, would the performance of that behavior cause you to judge an assessee

as being higher or lower on the particular dimension". Ratings were made on a scale of 1 – 'not diagnostic at all' to 7 – 'completely diagnostic'. A secondary measure of ease of evaluation was a global judgment made by trained assessors as to which of the dimensions were high ease of evaluation or low ease of evaluation.

Criterion variables

The dependent variables consist of supervisory ratings of 1) work pace, 2) quality orientation, 3) teamwork, 4) problem assessment, 5) problem solution, 6) applied learning, 7) technical/professional knowledge and skills, 8) job fit/ motivation and 9) an overall rating. In addition, there is a dichotomous promotion variable available only in the larger of the two datasets that will serve as the final dependent variable.

Most of the dimensions were defined above. The definitions for the remaining dimensions are 1) technical/ professional knowledge and skills was defined as having achieved a satisfactory level of technical and professional skill or knowledge in position-related areas; keeping up with current developments and trends in areas of expertise, 2) job fit/ motivation was defined as the extent to which activities and responsibilities available in the job are consistent with the activities and responsibilities that result in personal satisfaction; the degree to which the work itself is personally satisfying, and 3) applied learning was defined as assimilating and applying in a timely manner new job-related information that may vary in complexity.

Each incumbent's job performance was rated by his/her supervisor on the above 8 dimensions and one overall job

performance rating. The questionnaire contained 41 items, 5 to assess each dimension and the one overall rating.

Operationalizing ease of evaluation

Key behaviors were used during the assessment center process in 1993 to help focus assessors on important behaviors. These key behaviors were provided by the consulting organization to the assessment center raters prior to participation in the center. Assessors were trained in the use of these behavioral standards and were told to use them as an aid in evaluating performance.

To begin the process of operationalizing ease of evaluation, the key behaviors were evaluated to make sure that they 1) were representative of a particular dimension and 2) were, in fact, behaviors. Any listings that did not meet both of the above criteria were removed from the list. Once the lists were finalized, trained assessment center raters (not the same individuals that participated in the assessment center) rated the diagnosticity of each of the behaviors listed under a dimension.

Raters consisted of 6 Industrial/Organizational psychology students who had experience working as assessors in assessment centers. Each rater had been involved in formal assessment center training prior to participation in the present study. While training differed for each individual, all individuals were familiar with the assessment center process in general, the distinction between behaviors and inferences, observing and classifying behaviors, classifying behaviors into dimensions, and the variety of exercises used in assessment centers.

The 6 assessors participated in the diagnosticity training that involved two meetings, totalling 7 hours. The

training session consisted of 1) an overview of the assessment center 2) a brief overview of the study, 3) discussion of the concept of diagnosticity of behaviors, 4) explanation of the operationalization of the dimensions, 5) description of the exercises, 6) discussion of rating errors, 7) practice rating session, and 8) the actual diagnosticity ratings. During the practice session, raters were given a sample exercise (i.e., group discussion exercise) and behaviors for two dimensions (i.e., analysis and judgment) from an unrelated selection assessment center conducted in 1997. The same diagnosticity rating scale was used.

Ratings of key behaviors were done individually for approximately 5 behaviors at a time and then discussed. After each set of ratings, the raters discussed each rating until a consensus was reached. Prior to the rating session, it was decided that practice ratings would continue until there was at least 80% agreement on all ratings. After rating the first dimension (11 behaviors), there was little disagreement about the ratings for all raters and agreement reached close to 100%. Because of the almost perfect agreement, the group completed 9 of the 12 behaviors for the second dimension. Agreement stayed above 80% for all behaviors.

Once the practice session was completed, the 6 trained raters were asked to rate how diagnostic each behavior was of a particular dimension on a scale of 1 to 7, 1 representing ‘not diagnostic at all’ and 7 representing ‘extremely diagnostic’ (see Appendix A). The diagnosticity behavior ratings were done for each of the 7 dimensions. Each of the 7 dimensions had a list of between 3 and 10

unique behaviors representing it (see Appendix B). After the rating of each dimension, the raters used a group discussion format to resolve discrepancies and determine consensus on each rating. Finally, as a secondary measure of ease of evaluation, raters were asked to rank order the seven dimensions with regard to their "Ease of Evaluation" with 1 representing 'easiest to evaluate' and 7 representing 'hardest to evaluate'.

Interrater reliability was done on individual rater scores prior to consensus. Intraclass correlation coefficients ($\underline{\rho}$) were used because "…the intraclass correlation is the most appropriate measure of interrater reliability for interval scale data" (Tinsley & Weiss,1974; Kozlowski & Hattrup, 1992). The intraclass correlation coefficient can be interpreted as the proportion of the total variance in the ratings due to variance in the dimensions being rated. Values approaching 1 indicate a high degree of reliability whereas values approaching 0 indicate a complete lack of reliability.

As recommended by Tinsley & Weiss (1974) for situations when between judge differences do not lead to corresponding differences in the ultimate classification, between rater variance was not included as part of the error term. In this case, the between judge mean differences are unimportant because final ratings for each behavior were obtained by consensus.

Across the 6 raters, the reliability for Influence was $\underline{\rho}$ = .96 (n=7), Meeting participation $\underline{\rho}$ = .92 (n= 10), Problem assessment $\underline{\rho}$ = .91 (n=7), Quality $\underline{\rho}$ = .89 (n=6), Problem solution $\underline{\rho}$ = .86 (n=8), Teamwork $\underline{\rho}$= .81 (n=10), and Workpace

$\underline{\rho}$ = .80 (n=3). In this analysis, n is equal to the number of behaviors rated. The overall reliability of all of the ratings across all dimensions was $\underline{\rho}$ = .91 (n=51).

A second measure of reliability was assessed for each individual rater. Individual ratings prior to consensus were correlated with final consensus ratings. Because the final consensus ratings were used to determine the classification of high versus low ease of evaluation dimensions, it was important to assess the degree of agreement between raters' pre-consensus and post-consensus ratings. This analysis would demonstrate if one or a couple of the raters had an inordinately large influence on the rest of the group during consensus discussions. Rater 1 had an average rater reliability of $\underline{r}$=.70, rater 2 $\underline{r}$=.65, rater 3 $\underline{r}$=.72, rater 4 $\underline{r}$=.60, rater 5 $\underline{r}$=.71, and rater 6 $\underline{r}$=.75.

Once the diagnosticity ratings were finalized (see Appendix B), the proportion of high diagnosticity behaviors were assessed for each dimension. A behavior was rated as highly diagnostic if it was given a 5 or greater on the 7 point likert-type scale. The resulting proportions represent each dimension's 'ease of evaluation' and can be seen in Table 2. The proportion of 'high diagnosticity' behaviors was used to measure ease of evaluation because it is expected that a dimension with few, highly diagnostic behaviors will be easier to evaluate than a dimension with many behaviors, a few highly diagnostic and many with only limited diagnostic value. The proportion of highly diagnostic behaviors was .83 for Quality orientation, .70 for Meeting participation, .66 for Workpace, .60 for

Teamwork, .43 for Influence, .28 for Problem assessment, and .13 for Problem Solution.

As a secondary measure of ease of evaluation, the rank ordering of high to low ease of evaluation was averaged across all raters. The resulting rank ordering (from high ease to low ease) was Work pace, Quality orientation, Meeting participation, Teamwork, Problem assessment, Influence, and Problem Solution (see Table 3).

Both the proportion of high diagnosticity behaviors for a dimension and the rank orderings of ease of evaluation were compared to determine which dimensions would be labeled high ease of evaluation and which dimensions would be classified as low ease of evaluation. The same four dimensions: Quality, Workpace, Meeting participation, and Teamwork were determined to be high ease of evaluation for both operationalizations and thus these dimensions will be labeled as “high ease of evaluation dimensions”. The same three dimensions: Influence, Problem assessment, and Problem Solution were determined to be low ease of evaluation for both operationalizations and thus these three dimensions will be labeled as “low ease of evaluation”.

STATISTICAL ANALYSES AND RESULTS

Descriptive analyses

The analyses for hypotheses 1 and 2 were done using data from the larger of the two datasets in order to enhance the power of the tests. Listwise deletion was used for this dataset, resulting in N=1555. All data were converted to z scores, and with Pearson product-moment correlations, a correlation matrix was generated (see Table 4). The table includes means and standard deviations of all within exercise dimension scores.

There were 41 items on the performance appraisal questionnaire, representing 8 scales that made up the dependent variables: workpace, quality orientation, teamwork, problem assessment, problem solution, learning, jobfit, and knowledge (see Appendix C). The final item on the questionnaire asked about an individual's overall performance. The alpha reliability of the entire instrument was r=.98. Coefficient alpha reliabilities for each scale were r=.96 for workpace, r=.95 for quality orientation, r=.92 for teamwork, r=.92 for problem assessment, r=.95 for problem solution, r=.94 for learning, r=.91 for jobfit, and r=.93 for knowledge. Intercorrelations of the dependent variables can be seen in Table 5.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was used in addition to multitrait-multimethod analysis (MTMM) to assess the convergent and discriminant validity of the assessment center. The analysis of the MTMM matrix is based on correlations among observable variables containing measurement errors, while the subsequent interpretation contains conclusions about latent constructs. Confirmatory

Factor Analysis allows a researcher to investigate the relationships of the latent variables as well as modeling correlations among the factors. The application of CFA to the MTMM matrix will more accurately assess the contribution of trait, method, and error variance for each variable. These analyses presented unique challenges because unlike most assessment centers, each dimension was measured in at least two and at most three of the seven exercises.

Confirmatory factor analysis was used to compare the "fit" of 10 measurement models to the data using LISREL VIII (Joreskog & Sorbom, 1993). For all analyses in LISREL, the covariance matrix was analyzed as opposed to the correlation matrix. Models 2-10 represent submodels of model 1. A description of the competing models follows:

Null Model: This model is the most restricted model. It consisted of 15 unique factors associated with the 15 measured variables.

Model 1: 7 oblique dimensions and 6 oblique exercises. Similar to Bycio et al. (1987) and Kudisch et al.(1993), this model reflects the traditional design of the assessment center. In this assessment center 6 exercises were used to assess the 7 dimensions. All dimensions were allowed to correlate with all other dimensions. Dimension intercorrelations will provide evidence as to the degree of discriminant validity between dimensions. All exercises were allowed to correlate with all other exercises (see Figure 1).

Model 2: 7 orthogonal traits and 6 orthogonal exercises. Model 2 is the same as model 1 however, all of the variables were constrained to be orthogonal. Model 2 is a more parsimonious model than model 1 (see Figure 1).

Model 3: 7 oblique dimensions and 6 orthogonal exercises. Model 3 is the same as Model 1; however, this model constrains the exercise factors to be orthogonal while permitting the dimension factors to intercorrelate (see Figure 1).

Model 4: 7 oblique dimensions and 3 orthogonal exercises. Model 4 is based on the observation that the 6 exercises fall into 3 types of methods (i.e., production, group discussion, and problem solving). Similar to what Schneider and Schmitt (1992) found in their study of the form versus content of assessment center exercises, a good fit of model 4 to the data would imply that it is the form of the exercise that may be driving the exercise effects consistently found in the literature (see Figure 2).

Model 5: 1 dimension and 6 orthogonal exercises. Model 5 is based on past exploratory and confirmatory factor analytic research that found assessors were unable to distinguish among traits, thus, indicating no or little discriminant validity (Schneider & Schmitt; Bycio et al., 1987). This lack of discrimination has been termed a halo effect in which assessors form a general impression of whether an assessee is performing well or performing poorly in the assessment center. In model 5, exercises were not allowed to intercorrelate (see Figure 3).

Model 6: 1 Dimension and 6 oblique exercises. Model 6 is based on the same rationale as Model 5; however, exercises were allowed to intercorrelate (see Figure 3).

Model 7: 1 Dimension and 3 exercises. Model 6 also represents research findings that there is one global trait or halo factor that exists, rather than the 7 intended dimensions. This model is a highly parsimonious model,

although it significantly departs from the intended design of the center. In model 7, the 6 exercises are collapsed into exercises of similar form as in model 4 (see Figure 4).
Model 8: 7 oblique dimensions. Model 8 is based on the traditional assessment center assumption that assessee traits can be measured across exercises. This model implies that dimensions are the underlying foundation of the assessment center ratings. Model 8 directly rivals models 9 and 10. Accepting model 8 would provide evidence of high construct validity for the center as a whole (see Figure 5).
Model 9: 6 orthogonal exercises. Model 9 is based on past research that indicates the dimensions within exercises are highly correlated and across exercise dimensions are not correlated, indicating a complete lack of convergent validity (Sackett & Dreher, 1982; Robertson et. al., 1987). Acceptance of model 9 would indicate that participant traits were not being measured at all and it would provide support for arguments suggesting that human attributes be dropped as the organizing framework for assessment centers in favor of exercises or tasks (see Figure 6).
Model 10: 6 oblique exercises. Model 10 is a less parsimonious version of model 9 with exercises allowed to intercorrelate (see Figure 6).

Model Evaluation. As originally specified, model 1 (7 dimensions and 6 exercises) produced some problems. First, the model failed to converge because the matrix was not positive definite. Second, when the model did converge on a solution, an error appeared which indicated that one or more indicators in the Theta-Delta matrix 'may not be identified', indicating that for at least one variable of the matrix, the estimates of error variance were negative

(i.e., Heywood cases). This is considered an improper solution because these are values that are impossible in the population. (Bollen, 1989). Negative variance estimates are a result of a communality (squared correlation between the latent variable and a measurement variable) estimate larger than 1.00. Negative variances have many possible causes including insufficient data, bad starting estimates, and the need to specify a better fitting model (Wothke, 1993).

In a series of Monte Carlo experiments, Boomsma (1982) and Anderson and Gerbing (1984) found that nonconvergent solutions were greatest for confirmatory factor analysis on samples with N's less than 150 and for CFA's with only two indicators per factor (as cited in Bollen, 1989). Specifically, having only two indicators per latent dimension (as in the current case) resulted in negative error variances even when the models were correctly specified and good starting values were given. Similarly, a study by Marsh and Bailey (1991, as cited by Hoyle, 1995), using 435 MTMM matrices based on simulated and real data demonstrated that confirmatory factor analysis resulted in improper solutions 77% of the time. Improper solutions were greatest when the design was small (i.e., 3T x 3M vs. 5T x 5M), when the sample size was small and when the assumption of unidimensional method effects was violated.

After a careful review of the data, it seems most likely that the reason for the negative error variances found in most of the above models is due to the fact that each latent variable only had two indicators. As discussed earlier, each of the latent factors (i.e., dimensions) were measured in only two (or three) of the 6 exercises in an attempt to reduce the cognitive burden of assessors and to

maximize the construct validity of the dimensions. Good starting values were given and the N size of 1555 far exceeded the recommended N of 250.

There are several ways of reacting to improper solutions. First, the suggestion was made to replace the default starting values and to replace the LISREL default maximum-likelihood-based solution with an unweighted least squares solution (ULS), since the former method is particularly prone to produce Heywood cases (Wothke, 1993). In unweighted least squares, the residual matrix consists of the difference between the sample variances and covariances and the corresponding ones predicted by the model. While the most widely used fitting function for structural equation models is maximum likelihood, in practice, there is little difference between the results produced with these two methods. Unlike the ML approach, ULS estimates do not depend on a normality distribution assumption; however, the estimates are not scale free (Bollen, 1989). Because the current analyses were done using the covariance matrix and all of the variables were measured on the same scale, it was decided that using unweighted least squares would be an acceptable alternative to the maximum likelihood estimation procedure.

The recommended changes were made to the LISREL programs; however, some of the models continued to have problems with negative error variances (i.e., Heywood cases). Bollen suggests that there are three ways to deal with Heywood cases. Similar to what is automatically done in Bentler’s EQS program, one could re-estimate the model with an inequality constraint on the diagonal so that none of the error variances are negative, drop the x variable

that has the negative error variance, or constrain the error variance to be zero. While there are problems with all three approaches, constraining the error variances to zero for the offending variables seemed to be the most appropriate solution in the current study. Dropping the x variables would only serve to cause more identification problems for the model and constraining the error variance to remain between 0 and 1 would most likely result in error variances of zero, producing the same result.

There is little written in the literature on the consequences of constraining the error variance to equal zero on the model results and interpretation. The primary caution given is that because it is highly unlikely that the variables were measured without error in the population, there is some fundamental problem with the data, such as an under-identified model or inadequate sample size (Bollen, 1989). Ignoring these problems will not make them go away. However, there is no evidence that imposing these constraints will substantially change the model fit or factor loadings for the non-constrained variables (unlike the situation where the solution does not converge). For the constrained variables, we are still able to assess the relative contribution of trait and method influences.

The practice of setting negative variances to zero is so common in the literature that often researchers do not even acknowledge any caution in interpreting their results (Kudish et al., 1997). For example, Kleinmann & Koller (1997), while re-analyzing data from a study by Bycio et al. (1987) using confirmatory factor analysis noted that 11 of 29 models had problems including solutions not converging, factor correlations greater than 1, and negative error

variances. At no point in this article do they address these concerns although they do stress caution in interpreting their results because of “the relatively small sample size of N=70” (Kleinmann and Koller, 1997). Some authors have suggested that modifying models moves a researcher from confirmatory to exploratory factor analysis (James, Mulaik, & Brett, 1982). For this reason alone, caution should and will be used when interpreting the results.

Table 6 reports the corresponding goodness-of-fit statistics from the LISREL analysis for the different models. Determination of model fit in structural equation modeling is not straightforward. One must keep in mind model fit, model comparisons, and model parsimony.

Chi-square (X^2) is the most popular way of evaluating model fit. Chi-square assesses the magnitude of the discrepancy between the sample and fitted covariance matrices (Hoyle, 1995). However, there are problems with using only chi-square because it is sensitive to sample size. In large datasets, even trivial differences between the observed and fitted matrices may result in a significant X^2 and a rejection of the model. Therefore, it is important to include multiple fit indices in the search for the best fitting models.

The Normed Fit Index (NFI) is an incremental fit index that is recommended when sample size is large. NFI measures the proportionate improvement in fit by comparing the target model with the null model in which all of the observed variables are uncorrelated (Hoyle, 1995). Values vary from 0 to 1 with valued approaching 1 indicating good fit.

The Comparative Fit Index (CFI) is another recommended fit index when sample size is large although it is often somewhat downward biased (Hoyle, 1995). Values approaching 1 indicate good fit whereas values approaching 0 indicate a poor fit. The cutoff criterion of greater than .90 is required for model selection.

The Adjusted Goodness of Fit Index (AGFI) was used because it adjusts for the bias of fit indexes resulting from model complexity (Gerbing & Anderson, 1993). Often times there is greater goodness of fit for more complex, highly parameterized models because of the loss of degrees of freedom. Because MTMM data is inherently complex, this is an issue in the current data. AGFI is an absolute fit index because it directly assesses how well an a priori model reproduces the sample data. AGFI values vary between 0 and 1. Values approaching 1 represent a better fit and values approaching 0 represent the worst possible fit. Again, a cutoff criterion of .90 is required for model selection.

The Root Mean Square Residual (RMR) indicator was used because it is less likely to be influenced by sample size and model complexity. If the discrepancy between the observed correlations and the model reproduced correlations are very small, the model has a good fit. RMR has a better fit as residuals approach 0. A criterion of <.05 is required for model acceptance.

According to the fit measures, 6 of the models produced adequate fit) (see Table 6). The taxonomy for nested models as described by Widaman(1985) was used and chi-squared difference tests were employed where appropriate to determine which of these good fitting models had the best

fit with the data (see Table 7). Models 1, 3, 6, and 7 demonstrated the best fit based on a series of X^2 significance tests and the fit indexes. As expected, the least restrictive model, model 1 with 7 oblique dimensions and 6 oblique dimensions had a superior fit to any of the more restrictive models based on the fit indices. This was to be expected. In addition, because of the over-parameterization of the model, it resulted in a 'perfect fit'. LISREL was able to 'perfectly' reproduce the data. Thus, no fit indices were produced for the model. A chi-square significant difference test with model 3 indicates that model 1 does significantly improve the fit over model 3. However, since model 3 is the more parsimonious model, there is reason to accept it as the better model. Model 3 reflects 7 oblique dimensions while constraining the exercises to be orthogonal. Because model 3 had good fit with the data and retained 7 distinct dimensions, this model will be examined to assess the consistency in trait, method, error variance, and dimensions intercorrelations.

Using a X^2 comparison, model 6 (1 dimension and 6 oblique exercises) was a better fit than model 3 (7 oblique dimensions and 6 orthogonal exercises). Again, this model reflects findings in past research that assessors are measuring one global trait rather than distinct dimensions. Despite the fact that model 6 will not help test any of the a priori hypotheses, it will be retained to assess the degree to which this 'halo factor' exists in the data through some exploratory analyses at the end of the results section.

Finally, model 7 had a good fit with the data despite the fact that in X^2 comparisons, model 7 did not

significantly improve the fit over any of the other three best-fitting models. However, it is the most parsimonious model that had acceptable fit. Some researchers argue for the parsimony of a model to be used as a decision criteria in evaluating model fit (Hoyle, 1991). A less complex model that accounts for the data equally well may be preferred over more complex models. Model 7 reflected 1 global trait and 3 orthogonal exercises. In this model, the 6 exercises were collapsed into 3 exercises of similar form or type. Again, model 7 will not be used to test any a priori hypotheses, however, it will be used to examine the extent to which the exercise factors accounted for variance in the ratings through some exploratory analyses.

Because there was not one clear model that was superior to the others, results will be included for three models: Model 3, model 6, and model 7. The parameter estimates from completely standardized LISREL solutions for the 3 selected models can be found in Tables 8, 9, and 10 respectively. The proportion of variance explained by trait factors for all 3 models can be found in Table 11. Hypotheses 1 and 2 were evaluated with the results from model 3, the only accepted model that retained the 7 dimensions. Models 6 and 7 were retained to be used for exploratory analyses regarding the existence of a halo factor and exercise effects.

Both traditional multitrait-multimethod (MTMM) analysis and the results of the confirmatory factor analysis were used to examine hypotheses 1 and 2.

Hypothesis 1

Hypothesis 1 stated that high ease of evaluation dimensions would have higher convergent validity than low ease of evaluation dimensions. Similar to what was done by Shore et al. (1992), before testing this and subsequent hypotheses by comparing average correlations, the Pearson correlations were transformed to z scores and differences in mean correlations were then analyzed. It should be noted that because of the large sample size, significant results for trivial differences are more likely. For this reason, unless otherwise specified, all t-tests, mean comparisons, and regression analyses will be assessed using α=.01.

To test for convergent validity using MTMM, one must examine the correlations of the same trait measured by different methods (i.e., monotrait- heteromethod). Convergent validity is achieved when the validity diagonal values are “significantly different from zero and sufficiently large” (Campbell and Fiske, 1959, p. 82). The convergent validity coefficients for high ease of evaluation dimensions ranged from -.03 to .41, with an average monotrait-heteromethod coefficient of .22 (see Table 12). The convergent validity coefficients for low ease of evaluation dimensions ranged from .07 to .39 with an average monotrait-heteromethod coefficient of .19 (see Table 13). High ease of evaluation dimensions did not have a significantly higher average convergent validity (*r*=.22) than low ease of evaluation dimensions (*r*=.19),z=.92, *ns*.

The percentage of trait variance for each of the dimensions found in confirmatory factor analysis can also be interpreted as a measure of convergent validity. Using the

loadings from model 3, the high ease of evaluation dimensions (i.e., workpace, quality, teamwork, and meeting participation) had an average of 39% of variance accounted for by trait factors and an average of 21% accounted for by method factors (see Table 8). Low ease of evaluation dimensions (i.e., influence, problem assessment, problem solution) had an average of only 17% of variance accounted for by trait factors and an average of 17% accounted for by method factors.

In summary, for model 3 low ease of evaluation dimensions had significantly less average variance accounted for by trait variance (M=.17) than did high ease of evaluation dimensions (M=.39), $t(1554)=15.62$, $p < .01$. Also, high ease of evaluation dimensions had more average variance explained by trait factors (M=.39) than method factors (M=.21), $t(1554)=12.90$, $p < .01$, another indicator of convergent validity. In model 3, low ease of evaluation dimensions had the same amount of average variance explained by trait factors (M=.17) as by method factors (M=.17).

A closer examination of both the confirmatory factor analysis and multitrait-multimethod results reveals similarities in the findings. Using MTMM, the highest to lowest convergent validity coefficients were found for meeting participation (.41), influence (.39), teamwork (.37), workpace (.11), problem assessment (.10), problem solution (.07), and quality (-.03). While there is no direct way of measuring Campbell & Fiske's convergent validity with confirmatory factor analysis, the percentage of trait variance accounted for is often used as one indicator of convergent validity in CFA. The average percentage of trait variance accounted for by each of the

dimensions, from highest to lowest, was teamwork (55%), quality (47.5%), meeting participation (43.5%), influence (40%), problem assessment (15%), workpace (13%), and problem solution (4.3%). As one can see, the only true discrepancy in the findings is the quality dimension which has negative convergent validity and relatively high average trait variance accounted for. A review of Table 11 will help explain these findings. Trait variance is assessed at the within exercise dimension level and the quality dimension was assessed in two separate exercises. In production exercise #1, quality had almost all of its variance accounted for by the trait factor. However, quality as measured in production exercise #2 accounted for no trait variance, resulting in a slightly negative convergent validity coefficient. The fact that there was no trait variance explained for quality in production exercise #2 indicates that the exercise was not accurately tapping the targeted dimension.

Taken together, this pattern of results provides mixed support for hypothesis 1. Because of the problems discussed earlier with the use of observable variables in the Campbell and Fiske procedure, there is reason to weigh the confirmatory factor analysis results more heavily. It must be noted that because some of the confirmatory factor analysis results yielded improper solutions, caution must also be used in interpreting these results. However, the CFA results clearly supported hypothesis 1 across multiple, good fitting models, indicating support for hypothesis 1.

Hypothesis 2

Hypothesis 2 stated that high ease of evaluation dimensions would show more discriminant validity than low

ease of evaluation dimensions. There are two ways to test hypothesis 2 using the Campbell and Fiske (1959) multitrait-multimethod procedure. To test for discriminant validity, one must compare the correlations of different dimensions measured by different methods (i.e., heterotrait-heteromethod) to the correlations of the same dimension measured by multiple methods (i.e., monotrait-heteromethod). If discriminant validity exists, the monotrait-heteromethod correlations should be higher than the heterotrait-heteromethod. For high ease of evaluation dimensions, the average heterotrait-heteromethod correlation was .05. Three of the four dimensions (teamwork, workpace, and meeting participation) had monotrait-heteromethod correlations larger than the mean heterotrait-heteromethod correlations, demonstrating some amount of discriminant validity (see Table 12).

For low ease of evaluation dimensions, the average heterotrait-heteromethod correlation was .08. Two of the three dimensions (problem assessment and influence) had monotrait-heteromethod correlations greater than the mean heterotrait-heteromethod correlation, again, demonstrating some discriminant validity (see Table 13). The average heterotrait-heteromethod correlation for high ease of evaluation dimensions (r=.05) was not significantly lower than the correlation (r=.08) for low ease of evaluation dimensions, z=.84, ns.

A second, and more stringent discriminant validity criterion specifies that the monotrait- heteromethod coefficients should be higher than their corresponding heterotrait-monomethod coefficients. For high ease of evaluation dimensions, two of the four dimensions (teamwork

and meeting participation) met this criteria. For low ease of evaluation dimensions, only one of the three dimensions (influence) met this criteria.

A close examination of the discriminant validity correlations indicates that two of the four high ease of evaluation dimensions and one low ease dimension demonstrate discriminant validity. As a group, there is slightly more evidence of discriminant validity for high ease dimensions (.22 compared to .05) compared to low ease dimensions (.19 compared to .08).

An examination of the correlations between dimension factors found using confirmatory factor analysis will provide a more powerful test of the discriminant validity of the high versus low ease of evaluation dimensions. Average dimension correlations were taken for all variables for model 3 (see Table 14). For model 3, the average dimension correlation for high ease of evaluation dimensions (r=.17) was significantly lower than the average dimension correlation for low ease of evaluation dimensions (r=.39), z=3.84, $p < .01$, demonstrating greater discriminant validity for high ease of evaluation dimensions. Taken together, these results provide support for hypothesis 2.

Hypothesis 3

Hypothesis 3 stated that high ease of evaluation dimensions would demonstrate greater average criterion-related validity for dimensional performance than would low ease of evaluation dimensions. One dimension from high ease of evaluation (meeting participation) and one dimension from low ease of evaluation (influence) did not have corresponding dimension-specific performance ratings; therefore, these two dimensions were not included in the

analysis. The within dimension exercise scores for five dimensions (workpace, quality, teamwork, problem assessment, problem solution) were examined for their relationship to the corresponding performance variable. The average performance ratings for each of the five corresponding performance appraisal scales were used as the dependent variables (workpace average, quality orientation average, teamwork average, problem assessment average, problem solution average). Predictive validity coefficients were obtained for each of these five predictors with the corresponding dimension ratings. For example, the across exercise dimension score for workpace was correlated with the performance appraisal workpace average. The across exercise dimension score for quality was correlated with the performance appraisal quality orientation average. The correlations were averaged for the high ease dimensions and the low ease of evaluation dimensions.

Using the validity correlations from Table 15, high ease of evaluation dimensions did not have a significantly higher average criterion related validity r=.032 than low ease of evaluation dimensions r=.0075, z=.68, ns. with their corresponding performance dimension. Hypothesis 3 was not supported.

Hypothesis 4

Hypothesis 4 stated that high ease of evaluation dimensions would demonstrate greater criterion-related validity for job performance than would low ease of evaluation dimensions. This hypothesis was examined through multiple regression analyses for each of the nine dimensional performance ratings and the one overall performance rating. In addition, the dichotomous promotion

variable found in the larger of the two datasets was also used as a dependent variable. Because promotion is a dichotomous variable, both logistic regression and multiple regression were used to examine the relationship between dimensions and promotion. In this case, there were no differences between the results of the two regressions; therefore, the results will be discussed in multiple regression terms for consistency of interpretation.

In each regression equation high ease of evaluation dimensions were entered at step 1 and low ease of evaluation dimensions were entered at step 2. All within exercise dimension scores were averaged for each dimension and these averaged scores were used as inputs into the regression equation. Therefore, four variables were entered at step 1 (teamwork, workpace, quality, meeting participation) and 3 were entered at step 2 (problem solving, problem assessment, influence). The results of these regression analyses can be found in Table 16.

The two dependent variables that were significantly predicted by high ease dimensions were teamwork, $F(4,275)=3.58$, $p < .01$ and promotion, $F(4,1550)=14.81$, $p < .01$. For these two equations, a second regression equation was done with the order of entry reversed to determine which of the two sets of variables was accounting for more unique variance. In both cases, the high ease of evaluation dimensions continued to account for significantly more variance than the low ease of evaluation dimensions, providing partial support for hypothesis 4.

Hypothesis 5

Hypothesis 5 stated that the subset of high ease of evaluation dimensions would demonstrate predictive validity

equivalent to the validity found when using all dimensions. This hypothesis was examined through a review of the regression equations in Table 16. Because of the lack of prediction of 8 of the ten dependent variables, it is difficult to adequately examine hypothesis 5.

For the regression equations predicting teamwork performance, the subset of low ease of evaluation dimensions did not significantly add prediction above and beyond that of the high ease of evaluation dimensions. Beta-weights for the dimensions in this equation can be seen in Table 17. On the contrary, for the regression equation predicting promotion, the subset of low ease of evaluation dimensions did add significant prediction above and beyond that of high ease of evaluation dimensions. Beta-weights for the dimensions in this equation can be seen in Table 18. Overall, hypothesis 5 was not supported.

Exploratory analyses

In order to more closely examine the degree of exercise effects in the data, exercise scores were examined for their relationship to the across exercise dimension scores and the dependent variables. Exercise scores were available for each participant on all six exercises. Exercise scores ranged from 1 to 5 for all exercises.

The observed intercorrelations of the exercise scores can be found in Table 19. Means, standard deviations of exercise scores, and their correlations with across exercise dimension scores can be found in Table 20. All six exercise scores were entered simultaneously in a regression equation with overall performance as the dependent variable. Contrary to what might be expected from past literature

(Bobrow, 1996), the model was not significant, F(6, 268)=1.08, ns.

A review of the parameter estimates from model 7 (1 dimension and 3 orthogonal exercises) supports the existence of an exercise effect for the group discussion exercises only. The loadings on the group discussion method for influence, meeting participation, and teamwork all ranged between .54 and .65. The same method effect was not apparent for problem solution measured in group discussion exercise #2.

In order to examine the evidence for a 'halo factor' or one global trait accounting for the variance in the assessment center, the parameter estimates from models 6 and 7 were reviewed. In both of these models, acceptable fit was obtained with 1 global trait dimension. The parameter estimates loading on the global trait for model 6 (1 dimension and 6 oblique exercises) ranged from -.22 to .42 with an average trait loading of .12. In fact, three of the loadings were negative. The parameter estimates loading on the global trait factor for model 7 (1 dimension and 3 orthogonal exercises) ranged from -.22 to .45 with an average trait loading of .13. The same three within exercise dimensions had negative loadings similar to what was found in model 6 (quality in production exercise #2, teamwork in group discussion exercise #1, and teamwork in group discussion exercise #2). A comparison of the amount of variance accounted for in the 3 accepted models may be found in Table 11. There was far more variance accounted for by trait factors in model 3 (29%) than in models 6 and 7 (5% and 5% respectively). Despite the fact that both models 6 and model 7 produced adequate fit as measured by the chi-

square and fit indices, it does not appear that one ‘global trait’ adequately accounts for the trait variance in the data. In further support that the assessment center was measuring distinct dimensions, the average latent dimension intercorrelations was .23 for model 3 (7 oblique dimensions and 6 orthogonal exercises).

DISCUSSION

This study investigated the relationship between ease of evaluation of dimensions and the construct and criterion related validity of a selection assessment center. Results of the multitrait-multimethod analysis provided minimal support for the hypotheses that high ease dimensions would have greater convergent and discriminant validity than low ease dimensions. Confirmatory factor analysis results provided much stronger support for the greater convergent and discriminant validity of high ease of evaluation dimensions. Results of the hierarchical multiple regression analyses did not provide support for the hypotheses suggesting that high ease dimensions would demonstrate greater criterion related validity for job performance and promotion than low ease of evaluation dimensions. In fact, regression analyses revealed little or no predictive validity for the assessment center as a whole for either exercise scores or dimension scores.

Operationalizing Ease of Evaluation

From analyses of the ratings derived from the rating session, it is clear that the operationalization of ease of evaluation was successful. The reliability of the ratings was high for all dimensions (.80-.96). Individual rater reliabilities were lower (.60-.75); however, these reliability coefficients did not take into account ratings that were off by only one rating point (e.g., 4 and 5), a frequent occurrence during the rating process. Therefore, these estimates may actually underestimate the degree of agreement between individual raters and the final ratings.

The correlation between individual ratings and final consensus ratings also provides information about the relative influence of individual raters on the outcome of the consensus process. The consensus discussions might be called into question if one person's ratings had an exceptionally high correlation with final consensus ratings. This result would imply that one individual might be determining the final ratings, resulting in a situation that maximizes individual biases. A review of the correlations provides evidence that no one single rater was overly influential.

Throughout the process, it was clear that raters could identify what behaviors were more or less diagnostic when asked the question, "When a behavior occurs in the assessment center context, would the performance of that behavior cause you to judge an assessee as being higher or lower on the particular dimension"? In addition, raters were not told the specific definition of the ease of evaluation construct and they were asked to rank order how easy to evaluate the dimensions were. A strong indicator of the success of the operationalization is that the same dimensions fell into the categories of high and low ease of evaluation in each of the two operationalizations.

It should be noted that raters seemed to have an easier time rating the practice dimensions (i.e., analysis, judgment) than the assessment center dimensions (e.g., workpace, problem assessment). It may be the case that certain "types" of dimensions (e.g., higher level dimensions) have more variance in the observed behaviors and therefore the diagnosticity of those behaviors can be rated

more reliability. This hypothesis needs to be examined empirically.

Conversations with raters after the rating process was completed indicated that they felt comfortable and familiar with the dimension definitions and the exercises. In retrospect, it probably would have helped to have more behaviors to rate under each dimension although this was not possible in the current center. In addition, there were no negative key behaviors under any of the dimensions. It is likely that negative behaviors may be weighted more heavily when a dimension is being evaluated. Support for this hypothesis comes from a study by Brannick et al. (1989) that suggests individuals tend to remember negative information more and tend to use it more in evaluations than positive or neutral information. Future researchers should attempt to have full access to the design and administration of the assessment center in order to maximize the measurement of the construct of ease of evaluation.

Hypothesis 1

Hypothesis 1 predicted that high ease of evaluation dimensions would have greater convergent validity than low ease of evaluation dimensions. Mixed support was found for hypothesis 1. Almost equal monotrait-heteromethod correlations were found in the high ease and low ease dimensions in the MTMM analysis (.22 and .19, respectively). For high ease dimensions, the monotrait-heteromethod correlations ranged from -.03 (for quality) to .41 (for meeting participation). Monotrait-heteromethod correlations for the low ease dimensions ranged from .07 (for problem solution) to .39 (for influence). Despite the wide range of convergent validity correlations for both sets of

dimensions, support was found for hypothesis 1 based on the confirmatory factor analysis results. In the model that retained 7 distinct dimensions and 6 exercise factors, high ease of evaluation dimensions had greater average variance accounted for by trait factors (55%) than did the low ease of evaluation dimensions (35%). In addition, for model 3, high ease dimensions had more variance accounted for by the dimension factors than for exercise factors, another measure of convergent validity.

These results are consistent with the findings by Shore et al. (1992) and Kudisch et al. (1997) that more 'observable' dimensions demonstrated greater convergent validity than less 'observable' dimensions. In these cases, observability was defined as requiring a greater or lesser degree of inferential judgment. As discussed earlier, observability is a related construct to ease of evaluation.

Because it is difficult to evaluate the significance of correlations, it is useful to compare the present findings with findings from similar studies. A review of the summary of correlations reveals that while at first glance the convergent validity coefficients found in the present study appear somewhat low, they are similar to the convergent validities found in some other studies (see Table 21).

The present CFA findings that there is moderate convergent validity in the assessment center are inconsistent with previous studies that have found a complete lack of convergent validity (Sackett and Dreher, 1982). However, some recent studies have found greater support for the convergent validity of dimensions utilizing confirmatory factor analysis techniques. Kleinmann & Koller (1997) reanalyzed data from the Bycio et al. (1987) study

and found greater evidence of dimensions accounting for variance in the assessment center ratings than did Bycio et al. The original researchers had incorrectly concluded a complete lack of construct validity for the assessment center. Kleinmann & Koller (1997) argued that perhaps the utilization of strictly observable variables (i.e., correlational analyses) have been systematically underestimating the influence of dimensions on assessment center ratings. However, the majority of studies that have been able to use confirmatory factor analysis have shown that ratings are dominated by exercise factors (Sackett & Harris; Joyce et al., 1994; Silverman et al., 1986; Turnage & Muchinsky, 1982). In other words, the ratings represent the level of performance in the assessment center exercises, rather than performance on the dimensions. Contrary to these studies, the present assessment center did not show evidence of overwhelming exercise effects.

A second reason that has been discussed in the literature for why there is a lack of construct validity found in assessment centers is an overburden on the cognitive resources of raters (Howard, 1997). Typically raters are asked to rate at least 6 or more dimensions in a single exercise, leading to a reduction in the cognitive resources available to observe all relevant behaviors. In this assessment center, raters were asked to rate typically two and no more than four dimensions in a single exercise. This procedure should result in less cognitive burden for raters and may be a possible explanation for the relatively moderate convergent validity coefficients found.

Hypothesis 2

Hypothesis 2 predicted that high ease of evaluation dimensions would demonstrate greater discriminant validity than would low ease of evaluation dimensions. This hypothesis was partially supported by the results of the MTMM analysis and more fully supported by the confirmatory factor analysis.

Some amount of discriminant validity was found for both sets of high and low ease of evaluation dimensions. A comparison of the discriminant validity coefficients (i.e., heterotrait-heteromethod and heterotrait-monomethod correlations) with similar studies demonstrates that the discriminant validity coefficients found in the present study for both high and low ease of evaluation dimensions are relatively low compared to other studies, indicating that distinct traits are being measured. For example, in the present study the average heterotrait-heteromethod correlations were .05 for high ease dimensions and .08 for low ease dimensions. The range of heterotrait-heteromethod coefficients for the 8 studies that reported them was .06 to .45 with an average hthm of .23. Similarly, in the present study the average heterotrait-monomethod correlations was .19 for high ease dimensions and .35 for low ease of evaluation dimensions. The heterotrait-monomethod correlations for the twenty studies that reported them ranged from .25 to .90 with an average htmm of .54. The results of the study demonstrate that not only did high ease dimensions demonstrate greater discriminant validity than low ease dimensions but the entire assessment center exhibited more discriminant validity than has been seen in previous studies.

One potential explanation for the greater discriminant validity found for high ease dimensions in support of hypothesis 2 may be the low correlations of some of the high ease dimensions with all other dimensions including some of the dependent variables. In particular, quality orientation and work pace are either not correlated or are negatively correlated with most of the other variables assessed in the center. It was expected that the high ease dimensions would be more validly rated because of their greater proportion of highly diagnostic behaviors exhibited in the exercises. These dimensions would then be easier to rate accurately, resulting in greater construct and predictive validity for those dimensions. This reasoning is called into question when one closely examines the relationships of the high ease dimensions. As predicted, the data provided initial support for hypothesis 2; however, caution should be used when interpreting the results. It is possible that the differences found in discriminant validity for high and low ease of evaluation dimensions may not be due to the 'ease of evaluation' construct. Future studies will shed light on the true nature of the relationship.

According to Campbell & Fiske (1959), maximally similar methods should be used in order to increase the chances of convergent validity. In order to increase the chances of discriminant validity between two traits, maximally dissimilar methods should be used. In this particular assessment center, the designers tried to maximize both convergent and discriminant validity simultaneously by measuring the same variables in the same "type" of exercise (e.g., workpace measured in two production exercises, influence measured in two group discussion exercises) and

different dimensions in different types of exercises (e.g., workpace measured in production exercises and problem assessment measured in production exercises). However, because there were at least two dimensions measured in the same type of exercise, we would expect to see high convergent validity and low discriminant validity between the dimensions measured in the same exercises (e.g., workpace and quality; problem assessment and problem solution). In general, we did not see this relationship in the data, indicating that exercise effects may be lower in this study than is typically found in the literature. In light of this design, the greater discriminant validity found in this study are to be expected.

Shore et al. hypothesized that they found greater construct validity for their assessment center than was typically seen because they used across exercise dimension rating procedures as opposed to within-exercise dimension ratings. The researchers suggest some reasons why the use of within-exercise ratings may place a greater cognitive burden on assessors. First, the ratings are based on less behavioral evidence than are across exercise dimension ratings. Second, the demand characteristics of particular exercises may reduce convergence across exercises by eliciting different types of behaviors and third, within-exercise ratings are typically made by one rater and therefore, may be susceptible to more bias than are across-exercise ratings made by combining information among multiple raters. In the present assessment center, within exercise dimension ratings were used and moderate construct validity was found. It is possible that the use of across

exercise dimension scores may have even added to the construct validity found in the center.

However, despite the potential for across exercise dimension scores to increase the construct validity of the assessment center, there are benefits to using within-exercise dimension ratings. With these ratings, it is possible to examine the issue of traits versus exercises, a fundamental issue in the search for construct validity.

The preliminary finding of greater convergent and discriminant validity for high ease dimensions needs to be re-assessed in future research. Future research in a variety of types of assessment centers and with different dimensions will help determine whether the support found in the present study for the relationship between high ease of evaluation dimensions and greater construct validity is real or simply a by-product of the idiosyncrasies of this particular assessment center.

Hypothesis 3

Hypothesis 3 predicted that high ease of evaluation dimensions would have greater average criterion related validity for dimensional performance than will low ease of evaluation dimensions. Although high ease of evaluation dimensions had slightly greater average criterion related validity than low ease of evaluation dimensions, the difference was not significant. In fact, the average criterion related validity for dimensional performance for both high and low ease dimensions was unexpectedly low, less than 4%. For both groups the validity coefficients are extremely small. It should be noted that these validity coefficients were not corrected for restriction of range or unreliability in the criterion, a common practice in

validation research. However, even with these corrections, the correlations would remain far lower than has been seen in previous research (Thornton, 1992).

While there are no previous studies in the literature that have looked at the prediction of dimensional performance, it was expected that using the same dimension as both the predictor and criterion would provide a closer examination of the predictive validity of individual dimensions. Unfortunately, the complete lack of prediction found was contrary to the moderate to high predictive validities that have been found in previous studies predicting performance (Gaugler et al., 1987; Hunter and Hunter, 1984). Future studies must investigate the relationship of validity of assessment centers for predicting dimensional performance.

Hypothesis 4

Hypothesis 4 predicted that high ease of evaluation dimensions would demonstrate greater criterion-related validity for performance than would low ease of evaluation dimensions. This hypothesis was unsupported for 8 of the 10 dependent variables. As can be seen in table 16, for these 8 dependent variables (i.e., overall performance, quality, knowledge, problem assessment, problem solution, workpace, learning, and jobfit) there was no significant prediction for either high or low ease of evaluation dimensions.

For the remaining two dependent variables (i.e., predicting teamwork, and predicting promotion) the high ease of evaluation dimensions did account for more unique variance than low ease of evaluation dimensions, indicating, at first glance, modest support for hypothesis 4. However, two things should be noted. First, the levels of prediction

were, for all practical purposes, extremely low (i.e., 5% for teamwork and 6% for promotion). Second, an examination of the beta weights demonstrates that workpace, one of the high ease of evaluation dimensions is a significant negative predictor of teamwork, t=-2.6, p < .05 (see Table 17). Workpace has the largest beta weight of any variable in the equation (β = -.16) and is therefore contributing to the greater explanatory power of the high ease dimensions. Unfortunately, the hypothesis was that all dimensions assessed in the assessment center would be positively related to all outcome variables. Also, in the prediction of promotion, teamwork has a significant negative correlation with promotion, t=-5.79, p < .05 (see Table 18). Again, this is the largest beta weight in the equation (β = -.17), adding to the greater explanatory power of the high ease of evaluation. Because of these unique findings, there is no evidence that hypothesis 4 is supported.

The negative predictive validity found for the teamwork dimension predicting promotion, and the workpace dimension predicting ratings of teamwork by supervisors were both unexpected and troubling. Although the predictive validities were small in both of these cases, the findings cause concern for a number of reasons.

First, the current assessment center was based on a thorough job analysis. Job analysts reviewed all relevant materials to the job (e.g., job descriptions, performance review forms, review of the company's mission and value statements), reviewed job analyses for similar jobs, and interviewed job incumbents and supervisors. This information was used to determine the factors linked to

success in the job of production associate or equipment services associate for the car manufacturing plant. The 7 dimensions measured in the assessment center were determined to be important competencies in the two jobs under review. One hypothesis is that the job analysis did not accurately assess the competencies needed for the job.

A second hypothesis is that certain factors like workpace and teamwork may have a curvilinear relationship with performance. In explanation of the negative relationship that was found between teamwork and promotion, it may be the case that individuals need to have good teamwork skills to be successful at the job. However, those individuals that are highest on the teamwork dimension may not be the individuals that will make the best managers. Not as surprising is the relationship that was found between workpace and the teamwork performance dimension. It is an advantage for the company if all workers have a steady workpace; however, the individuals that work the fastest may not be the best team workers. The fastest people on the production line may be more individually oriented and less team oriented.

The importance of accurate identification of competencies based on job analyses is evident if one considers what would happen if inaccurate dimensions were used for a developmental or training assessment center. Participants would be given specific feedback based on their dimensional performance. If these dimensions were negatively related to job performance dimensions or promotion, individuals would be mislead and misguided in understanding their areas for improvement.

In the case of the current assessment center, cause for concern still exists although the negative consequences are not as clear. For example, if the company hires individuals based on their assessment center performance and this performance is either not related or negatively related to success in the organization, both individuals and the organization will be harmed. In this organization, teamwork is stressed as important to successful job performance; however, there is some evidence that individuals scoring highest on this dimension are actually less likely to get promoted. Thus, it is crucial that organizations accurately understand the factors that are related to success in the organization prior to utilizing an assessment center for selection, development, or training.

Hypothesis 5

Hypothesis 5 predicted that the subset of high ease of evaluation dimensions would demonstrate predictive validity equivalent to the validity found when using all dimensions. Contrary to this prediction, high ease of evaluation dimensions only added to the prediction of promotion and did not add to the prediction of any other variables. Moreover, as explained above, most of that predictive validity was due to the negative prediction of teamwork. Because of the overwhelming lack of prediction of the entire assessment center, it is impossible to draw any substantive conclusions about the predictive validity of the subsets of dimensions.

The failure to find any significant relationship between the assessment center dimensions and most of the job performance variables was unexpected. Although there are few published studies with low or negative predictive validities for assessment centers, they do occur. Gaugler

et al. (1987) found that validity coefficients ranged from -.25 to .78, demonstrating that there is a great degree of variance in the predictive validity of assessment centers. Gaugler et al. (1987) attempted to find most of the research studies that have been done on the predictive validity of assessment centers, including those studies that had not been published. This practice helps to avoid what is commonly referred to as the 'file drawer problem' which means that often only studies with significant results get published resulting in meta-analyses that overestimate the relationship between variables. Even with the inclusion of all of the studies in the meta-analysis, Gaugler et al. still found an estimated true validity of r=.37. The predictive validity found in the present study was far below average and was similar to what was found in a number of unpublished studies (Schmitt, Noe, Meritt, & Fitzgerald, 1984; Ritchie, 1980; Klimoski & Strickland, 1981; Wissaman & Rankin, 1982 as cited in Gaugler et al., 1987).

Similar to what has been found previously (Chan, 1996), the current study did predict promotion better than job performance. Chan assessed the construct and criterion related validity of an assessment center simultaneously. He found no evidence of construct validity of the assessment center utilizing a multitrait-multimethod analysis, exploratory factor analysis and comparisons with a nomological network of constructs independent of the center. While he found assessment center ratings were predictive of subsequent promotion (r=.59), he did not find them predictive of concurrent supervisory ratings of performance (r=.06, n.s.). It has been suggested that assessment center ratings may predict promotion better than job performance

ratings because of subtle criterion contamination (Klimoski & Brickner, 1987).

Subtle criterion contamination is evidenced by shared stereotypes of what type of person is likely to be 'successful' in a particular organization. These stereotypes will influence both ratings in the assessment center and ratings of potential or promotion. The assessors are in effect attempting to 'capture the policy' of future decision makers in the company, and these policies may or may not be based on job performance. Therefore, the shared stereotypes are less likely to influence performance ratings.

Chan concluded that subtle criterion contamination was the most likely explanation for his findings of low construct validity, low criterion related validity for job performance ratings, and high criterion related validity for promotion. Given the moderate construct validity found in the current assessment center and the fact that the raters in the current center were not members of the selecting organization and are unlikely to share the same stereotypes as those responsible for determining promotion, it is unlikely that this is the primary explanation for the findings. In addition, the greater criterion related validity found in this study for predicting promotion over job performance was still extremely low ($\underline{R}$=.06).

Another possible explanation for the current results are provided by the recent findings of Goldstein et al. (1998). They investigated the degree to which subgroup (Black-White) mean differences on various assessment center exercises may be a function of the type of exercise employed. Their results suggested that subgroup differences

did vary by type of assessment center exercise and that the subgroup difference appeared to be a function of the cognitive component of the exercise. In addition, they found preliminary support that the validity of some assessment center exercises in predicting supervisor ratings of job performance is based, in part, on their cognitive component. Because of the level of the current assessment center and the types of dimensions assessed, there was less of a cognitive component in the assessed dimensions than is typically seen. The lack of a cognitive component may be partially contributing to the low criterion related validity found in the center.

This hypothesis is not supported by meta-analytic results that suggest that assessment centers are equally predictive across multiple levels of an organization (Gaugler et al., 1987). It is unlikely that the cognitive component of the dimensions assessed at these differing levels are comparable. In addition, Chan (1996) found that traditional cognitive measures were predictive of job performance ratings but assessor ratings were not. On the other hand, assessor ratings were predictive of promotion but cognitive ability measures were not.

In order to more fully understand the validity of assessment centers, it is imperative that future studies simultaneously examine criterion and construct validity of assessment center ratings. Gaugler et al. (1987) found that the degree to which validation studies were internally and externally valid, measured by ratings of the quality of the study, was related to the predictive validity of the assessment centers. Without construct validity information it is impossible to determine what is causing the high

criterion related validities that have been consistently found in the literature.

Exploratory Analyses

Exploratory analyses were conducted to examine the extent of exercise effects in the assessment center. A review of all of the data provides evidence that exercise effects were lower in this study than have been previously seen (Sackett & Dreher, 1982; Turnage & Muchinsky, 1982; Silverman et al., 1986; Schneider & Schmitt, 1992). In particular, a review of the heterotrait monomethod correlations (an indicator of exercise effects) indicates that the greatest exercise effects occurred in Group Discussion Exercise #1 and Problem Solving Exercise #2 (htmm= .37 and .63 respectively.) Based on these correlations, there appeared to be no exercise effects for Production Exercise #1 (-.02). A second method of exploring the exercise effects is to examine 'type of exercise' effects found in the confirmatory factor analyses. As pointed out previously, there were three types of exercises that made up the six exercises (i.e., production, group discussion, and problem solving).

A review of the completely standardized parameter estimates from model 7 indicates that the greatest 'type of exercise' effect occurred for the group discussion exercises.

The finding that group discussion exercises showed the greatest method effects is not surprising given previous findings of overwhelming method effects in assessment centers. The group discussion exercises in the current assessment center were most similar to the types of exercises used in assessment centers in the literature. Most of the assessment center research studies have used

‘typical’ exercises such as group discussions, role-plays, and in-baskets. Few studies have used the exercises similar to the production and problem solving exercises used in the present assessment center. It is likely that the degree of exercise effects found in assessment centers is being influenced by the type of exercises used. While this has not yet been examined in the literature, it is a promising area for future research.

The second set of exploratory analyses were conducted to examine evidence of a ‘halo’ or global trait factor accounting for the trait variance in the assessment center. Contrary to what has been found previously (Archambeau, 1979; Outcalt, 1988; Konz, 1988), there did not appear to be one global trait factor that accounted for most of the trait variance in the present assessment center. Konz (1988) found a high correlation among final dimension ratings, suggesting the ratings did not reflect entirely distinct attributes. The author concluded that assessors’ ratings of dimension performance may be influenced by some general impression of the candidates. In the present study, the finding of relatively low average latent dimension intercorrelations equal to .23 for model 3 (7 oblique dimensions and 6 orthogonal exercises) provides further support that there is not one global dimension factor pervading the data. These results, as well as the evidence found for substantial discriminant validity for both the high and low ease of evaluation dimensions further supports the hypothesis that distinct dimensions were being measured in the assessment center.

Study Limitations

The present study has several limitations worth noting. In research there is always a trade-off between maximizing internal and external validity at the expense of the other. The current study used archival data from a selection assessment center used in 1993 as part of a larger selection system to hire thousands of individuals. The largest problem inherent in using archival data is that the researcher has no control over the design or administration of the assessment center. There is no way to control potentially confounding variables.

In the domain of assessment centers, few laboratory studies have been conducted. Because of the applied nature of the centers, true gains in the assessment center construct validity literature will only occur with systematic research programs by designers and administrators of practicing centers. The involvement of these individuals in the research process will benefit all involved.

Second, while preliminary support was found for the construct validity of the assessment center, the assessed dimensions and exercises yielded no predictive validity for 8 of the 10 performance dimensions. For the 2 dependent variables that were predicted, the criterion related validity was extremely small and there were even some dimensions that yielded significant negative prediction. These findings are contrary to what has been found in previous literature and most likely hindered the potential to adequately test the ease of evaluation hypotheses.

Third, the dimensions used in the present assessment center are different than can be found in any of the published studies on assessment centers. It was expected

that this difference would be a benefit to investigating the research questions; however, the level of dimensions could have made it more difficult to adequately examine the research hypotheses. Most assessment centers use higher level dimensions (e.g., analysis, judgment) that may have greater variance in the behaviors that are generated under them. Thus, for these types of dimensions, it may be easier to rate the ease of evaluation of dimensions of behaviors more reliably. Also, although all of the dependent variables were normally distributed, it is possible that in higher level jobs, there may be more variance in job performance. Some individuals may "shine" more allowing for greater prediction of both performance and promotion. This may be a topic for future studies and future meta-analyses. As stated earlier, there is not one "typical" assessment center, although, the majority of assessment centers in use are used for middle management jobs and fewer at the lowest and highest levels of the organization (Howard, 1993). With the growing use of assessment centers, it is likely that they will increasingly be used at all levels of the organization, warranting future research on the topic. A more thorough examination of the hypotheses from the current study should be done by investigating a wide variety of dimensions across many levels of jobs and different types of organizations.

Fourth, there was no access to the reliabilities of the assessment center ratings. It is possible that even if candidate's behavior was consistent across exercises, very dissimilar ratings could result from low interrater agreement. However, research on the reliability of assessor judgements (e.g., Howard, 1974) suggests that moderately

high interrater agreement is likely in the assessment center context. Gaugler et al. (1987) estimated that most assessment centers achieve adequate levels of interrater agreement (>85%). A second possibility is that individuals may actually behave differently in different exercises, indicating low cross situational consistency and low construct validity for individual dimensions (Bycio et al., 1987). It is possible that exercises differ structurally which may limit the extent to which cross-situational consistency can be demonstrated. Structural differences may include whether candidates were observed and judged as individuals (i.e., the production and problem solving exercises) or as a group (i.e., group discussion exercises) or whether dimensions were measured through direct observation (i.e., production and group discussion exercises) or candidates were measured on their written performance (i.e., problem solving). Other less salient structural differences might include the length of the exercises and the extent to which the assessors directly interact with the candidates. Future studies must take these structural differences into account while assessing the construct validity of assessment centers.

Fifth, there was no access to the reliabilities of individual exercises. Exercise reliabilities might have provided information that would help explain why some exercises were better at measuring certain dimensions than others.

Sixth, there were a limited number of behaviors rated under each dimension during the operationalization of ease of evaluation. Key behaviors ranged from 3 behaviors for workpace to 10 for meeting participation and teamwork.

These lists of key behaviors were given to raters prior to their ratings in the 1993 assessment center. Unfortunately, lists of all of the observed behaviors in the exercises were not available. Access to all of the behaviors exhibited may have allowed a more accurate assessment of the construct of ease of evaluation of the dimensions. As mentioned earlier, the addition of negative behaviors would also have been an improvement.

Finally, there were identification problems with the confirmatory factor analysis. Because some of the models yielded improper solutions, negative error variances were fixed at zero for some of the indicators resulting in suboptimal solutions. While the parameter estimates appear stable across multiple, good-fitting models, caution should be taken in interpreting the results. Despite the large sample size (n=1555), the results should be replicated in order to have greater confidence in the findings. In addition, the percentage of variance accounted for by trait and method factors was relatively low across all of the dimensions. There was a lot of error in the measured variables indicating that there is significant room for improvement in the construction of similar assessment centers.

Future Directions

While there is some preliminary support for the greater convergent and discriminant validity of high versus low ease of evaluation dimensions, there was little evidence for the predictive validity of the center. The findings from this study have generated more questions than answers. Multiple possibilities exist as to why this pattern of results was found in the present assessment center. One possibility is

that the construct of ease of evaluation was assessed correctly and the hypotheses were accurate; however, these hypotheses could not be tested fairly because of unique characteristics of the present assessment center including the lack of prediction of job performance.

A second possibility is that the construct of ease of evaluation was assessed correctly; however, the hypotheses put forth regarding the relationship of ease of evaluation of dimensions to the construct validity and predictive validity of assessment centers are inaccurate. Thus, there is no true difference in the construct validity of high versus low ease of evaluation dimensions.

A third possibility is that the ease of evaluation construct either does not exist as described here or it was not operationalized correctly in the present study. Although this seems like an unlikely alternative based on the study's findings, it is not possible to assess the verity of any of these scenarios from the results of the present study. Future empirical research must be done to determine which of these scenarios is most likely.

Another issue that should be addressed in future research is the level of the assessment center and the types of dimensions assessed. For example, would the results of the study be the same had we used a selection assessment center for a mid-level or high-level managerial job? A fairly recent survey showed that most assessment center candidates were assessed for positions at lower management levels (Bentson, Gaugler, & Pohley, 1992 as cited in Howard 1993). Dimensions assessed in assessment centers include a mixture of skills, behaviors, knowledges and motivations (Howard, 1993) and often different types of dimensions are

assessed in different levels of assessment centers. To complicate matters further, the same dimension may even be defined differently depending on the level of the assessment center. For example, in the DDI wheel of dimensions, leadership at the mid-level is defined as 'fostering teamwork, motivating others, coaching, developing and providing direction'. Leadership at the executive level involves 'attracting and developing talent, empowering others, and leadership versatility' (Howard, 1997). Future research must examine differences in the dimensions assessed at multiple levels of the organization as well as the resulting construct and predictive validity of the assessment centers.

In conclusion, this study has clearly acknowledged that there is no simple answer to the construct validity problem of assessment centers. One hypothesis — that the specific characteristics (e.g., ease of evaluation) of dimensions used in an assessment center leads to greater or lesser construct validity of the center – is still a viable one. Despite initial support for the study's main hypotheses, future research must address the characteristics of dimensions as well as other potential variables in the search for construct validity. I cannot stress enough how important it is that future research on assessment centers be done with the help and support of assessment center designers and administrators. It is uniquely this group of people that have the ability to do systematic research in the area of the construct validity of assessment centers. Systematic research will ultimately advance understanding of the causes of assessment center success at predicting job performance and it will add to the utility of the centers

(e.g., increase predictive validity, more accurate developmental information).

References

Archambeau, D.J. (1979). Relationships among skill ratings assigned in an assessment center. Journal of Assessment Center Technology, 2, 7 - 10.

Bagozzi, R.P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. Journal of Research in Personality, 27, 49-87.

Bobrow, W. (1996). Assessment centers: Is it better to score by dimension or by exercise? (online). Available: http://www.ipmaac.org/acn/oct96/asstctr.html

Bollen, K.A. (1989). Structural equations with latent variables. New York: Wiley.

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

Bray, D.W., & Campbell, R.J. (1968). Selection of salesmen by means of an assessment center. Journal of Applied Psychology, 52, 36-41.

Bray, D.W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. Psychological Monographs: General and Applied, 80, 1-27.

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

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

Cascio, W.F., and Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107 - 118.

Chan, D. (1996). Criterion and construct validation of an assessment centre. Journal of Occupational and Organizational Psychology, 69, 167- 182.

Cohen, B.M., Moses, J.L., Byham, W.C. (1974). The validity of assessment centers: A literature review. Monograph II. Pittsburgh, PA: Development Dimensions Press.

Cooper, W.W. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218 - 244.

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

Eden, D. (1984). Self-fulfilling prophecy as a management tool: Harnessing Pygmalion. Academy of Management Review, 9, 64 - 73.

Fleenor, J.W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. Journal of Business and Psychology, 10, 319-332.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., & Bentson, C. (1987). 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.

Gerbing, D.W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp.40-65). Newbury Park, CA: SAGE Publications.

Goldstein, H.W., Yusko, K.P., Braverman, E.P., Smith, D.B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. Personnel Psychology, 51, 357-374.

Gorsuch, R.L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Guion, R.M. (1987). Changing views for personnel selection research. Personnel Psychology, 40, 199 - 213.

Hampson, S.E., John, O.P., & Goldberg, L.R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. Journal of Personality and Social Psychology, 51, 37 - 54.

Hays, W.L. (1994). Statistics. New York: Harcourt Brace College Publishers.

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.

Hinrichs, J.R., & Haanpera, S. (1976). Reliability of measurement in situational exercises: An assessment of the assessment center method. Personnel Psychology, 29, 31-40.

Howard, A. (1993). Will assessment centers be obsolete in the 21st century? Replies to the critics. 21st International Congress on the Assessment Center Method.

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

Hoyle, R.H. (1995). Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: SAGE Publications.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72 - 98.

Isen, A.M., & Diamond, G.A. (1989). Affect and Automaticity. In Uleman, J.S., & Bargh, J.A. (Eds.), Unintended Thought, (124 - 150). New York, NY: The Guilford Press.

James, L.R., Mulaik, S.A., & Brett, J.M. (1982). Causal analysis. Assumptions, models, and data. Beverly Hills, CA: Sage.

Joiner, D.A. (1984). Assessment centers in the public sector: A practical approach. Public Personnel Management Journal, 13, 435 - 450.

Jones, E.E. & Davis, K.E. (1965). A theory of correspondent inferences: From acts to dispositions. In L. Berkowitz (Ed.), Advances in experimental and social psychology, 2, 220 - 266. New York: Academic Press.

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

Joreskog, K.G., & Sobrom, D. (1989). LISREL 7/8 user's reference guide. Chicago: Scientific Software International.

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

Kavanagh, M.J., MacKinney, A.C., & Wollins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34 - 49.

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

Kleinmann, M., & Koller, O (1993). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. Journal of Social Behavior and Personality, 12, 65-84.

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

Konz, A.M. (1988). A comparison of dimension ratings and exercise ratings in assessment centers. Unpublished doctoral dissertation. University Of Maryland.

Kozlowski, S.W. & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus.

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

Louiselle, K.G. (1986). Confirmatory factor analysis of two assessment center rating procedures. Paper presented at the Seventh Annual IO/OB Graduate Student Conference, Minneapolis, MN.

Mayes, B.T. (1997). Insight into the history and future of assessment centers: An interview with Dr. Douglas W. Bray and Dr. William Byham. Journal of Social Behavior and Personality, 12, 3-12.

McArthur, L.Z., & Baron, R (1983). Toward an ecological theory of social perception. Psychological Review, 90, 215 - 238.

McEvoy & Beatty (1989). Assessment centers and subordinate appraisals of managers: A 7- year examination of predictive validity. Personnel Psychology, 42, 37 - 52.

Myers, D.G. (1998). Social Psychology. New York: McGraw-Hill Companies.

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

Outcalt, (1988). A research program on general motor's foreman selection assessment center: Assessor/ assessee characteristics and moderator analysis. Paper presented at the 16th International Congress on the Assessment Center Method, Tampa, Fl.

Reichardt, C.S., & Coleman, S.C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. Multivariate Behavioral Research, 30, 513 - 538.

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

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

Robertson, Gratten, & Sharpley (1987). The psychometric properties and design of managerial centers: Dimensions into exercises won't go. Journal of Occupational Psychology, 60, 187 - 195.

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

Russell, C.J. (1994). A model of assessment center construct space and an agenda for future research. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Nashville, TN.

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

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

Sackett, P.R., & Hakel, M.D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. Organizational Behavior and Human Performance, 23, 120- 137.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. Journal of Occupational and Organizational Psychology, 70, 103 - 109.

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

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.

Shoda, Y., Mischel, W., & Wright, J.C. (1989). Intuitive interactionism in person perception: Effects of situation-behavior relations on dispositional judgments. Journal of Personality and Social Psychology, 56, 41 - 53.

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

Silverman et al, (1986). Influence of assessment center methods on assessor's ratings. Personnel Psychology, 39, 565 - 578.

Squires, P., Torkel, S.J., Smither, J.W. & Ingate, M.R. (1988) Validity and generalizability of a role-play test to select telemarketing representatives. Paper presented at the meeting of the APA, Atlanta, GA.

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

Thornton, G.C. III (1992). Assessment Centers in Human Resource Management. New York: Addison-Wesley Publishing Company, Inc.

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

Thornton, G.C. III and Cleveland (1990). Developing managerial talent through simulation. American Psychologist, 45, 190 - 199.

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

Tinsley, H.E., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22, 358-376.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. Psychological Review, 93, 239 - 257.

Trope, Y. (1989). Levels of inference in dispositional judgment. Social Cognition, 7, 296 - 314.

Trope, Y., & Bassok (1982). Confirmatory and diagnosing strategies in social information gathering. Journal of Personality and Social Psychology, 43, 22-34.

Trope, Y., & Burnstein, E. (1975). Processing the information contained in another's behavior. Journal of Experimental Social Psychology, 11, 439 - 458.

Trope, Y., & Cohen, O. (1989). Perceptual and inferential determinants of behavior- correspondent attributions. Journal of Experimental Social Psychology, 25, 142 - 158.

Trope, Y., Cohen, O., & Alfieri, T. (1991). Behavior identification as a mediator of dispositional inference. Journal of Personality and Social Psychology, 61, 873 - 883.

Trope, Y., & Liberman (1993). The use of trait conceptions to identify other people's behavior and to draw inferences about their personalities. Personality and Social Psychology Bulletin, 19, 553- 562.

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

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

Wollowick and McNamara (1969). Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 53, 348 - 352.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K.A. Bollen & Long, (Eds.), Testing structural equation models (pp. 256-93). Newbury Park, CA: Sage.

Zedeck, S. (1986). A process analysis of the assessment center method. In B.M. Staw & L.L. Cummings (Eds.), Research in Organizational Behavior (Vol. 8, 259 - 296). Greenwich, CT: JAI Press.

APPENDIX A

DIAGNOSTICITY RATING SCALE

DIAGNOSTICITY RATING FORM

How representative or diagnostic is the listed behavior of the related dimension and dimension definition? In other words, how much does the indicated behavior cause one to judge an assessee as being higher or lower on the particular dimension? Please indicate your diagnosticity rating based on the exercises where the dimension was measured?

1	2	3	4	5	6	7
not diagnostic all	a minimum amount of	a small degree of diagnosticity	somewhat diagnostic	a considerable amount of diagnosticity	agreat amount of diagnosticity	completely diagnostic

APPENDIX B

DIAGNOSTICITY RATINGS FOR KEY BEHAVIORS BY DIMENSION

WORK PACE

Def.--Performing work at a specific pace without unnecessary expenditures of time or waste of supplies and materials; demonstrating a consistent rate of speed for accomplishing activities in a specific order. Sample key behaviors include performing at a consistent and appropriate speed and performing work with high accuracy.

Related Exercises: Production Exercise #1
Production Exercise #2

Behavioral Anchors:	Diagnosticity Rating
Performs at a consistent and appropriate speed	7
Performs work with high accuracy	4
Is able to do specific work motions at a sustained speed	5

QUALITY ORIENTATION

Def. -- Accomplishing tasks through concern for all areas involved, no matter how small; showing concern for all aspects of the job; accurately checking processes and tasks; maintaining watchfulness over a period of time. Sample key behaviors include attending to all details of the job and accurately checking processes or work outputs.

Related Exercises: Production Exercise #1
Production Exercise #2

Behavioral anchors:	Diagnosticity Rating
Shows concern for quality	5
Attends to all details of the job	6
Accurately checks processes or work outputs	6
Pays attention to detail	5
Visibly checks product outputs for appearance, errors, bad parts, and imperfections	5
Seeks to identify and eliminate root causes of quality problems	2

PROBLEM ASSESSMENT

Def.--Securing relevant information and identifying key issues and relationships from a base of information; relating and comparing data from different sources; identifying cause-effect relationships. Sample key behaviors include gathering information, organizing information, and anticipating potential problems.

Related Exercises: Problem Solving Exercise #1
Problem Solving Exercise #2

Behavioral anchors:	Diagnosticity Rating
Identifies issues and problems	5
Gathers information	2
Interprets information	4
Organizes information	3
Distinguishes relevant from irrelevant information	4
Integrates both quantitative and qualitative information to understand the cause of problems	6
Asks clear and specific questions and follow-up questions	2

PROBLEM SOLUTION

Def.--Committing to an action after developing alternative courses of action that are based on logical assumptions and factual information and that take into consideration resources, constraints, and organizational values. Sample key behaviors include developing and considering alternatives, selecting a course of action, and being decisive.

Related Exercises: Problem Solving Exercise #1
Problem Solving Exercise #2
Group Discussion Exercise #2

Behavioral Anchors:	Diagnosticity Rating
Develops alternatives	4
Recognizes when events/situations require action	3
Selects a course of action	3
Makes recommendations that are sound and have impact	6
Considers overall impact of decisions	4
Follows up on the effectiveness of a solution	2
Develops decision criteria	4
Is willing to take a stand or make a decision	3

INFLUENCE

Def.--Using appropriate interpersonal styles and methods to inspire and guide others toward goal achievement; modifying behavior to accommodate tasks, situations, and individuals involved. Key behaviors include making recommendations that are sound and have impact, and developing and considering alternatives.

Related Exercises: Group Discussion Exercise #1
Group Discussion Exercise #2

Behavioral Anchors:	Diagnosticity Rating
Identifies shared goals	4
Links actions to needs	3
Gains agreement to a course of action	6
Uses logical arguments and appeals to others' needs	6
Presents ideas/ information in a participative manner	4
Utilizes effective interpersonal styles and methods when attempting to influence others	6
Emphasizes values and principles rather than rules and regulation	2

MEETING PARTICIPATION

Def. --Using appropriate interpersonal styles and methods to motivate and guide a meeting toward its objectives; modifying behavior according to the tasks and individuals; being aware of the needs and potential contributions of others. Sample key behaviors include making procedural suggestions, summarizing information, and soliciting the ideas of others.

Related Exercises: Group Discussion Exercise #1
Group Discussion Exercise #2

Behavioral Anchors:	Diagnosticity Rating
Makes procedural suggestions	5
Summarizes information	3
Checks for understanding	5
Checks for agreement	5
Solicits others ideas; involves others in meeting	6
Resolves conflicts/ disagreements	6
Listens actively	3
Deals effectively with dominant members	5
Keeps discussion focused on the issue at hand	5
Keeps track of time during meeting situations	2

TEAMWORK

Def.-- Active participation in, and facilitation of, team effectiveness; taking actions that demonstrate consideration for the feelings and needs of others; being aware of the effect of one's behaviors on others. Sample key behaviors include acknowledging other' concerns and contributions and clearly communicates relevant ideas.

Related Exercises: Group Discussion Exercise #1
Group Discussion Exercise #2

Behavioral Anchors:	Diagnosticity Rating
Shows consideration for others' feelings and opinions	5
Acknowledges others' concerns and contributions	5
Clearly communicates relevant ideas	3
Presents information/ ideas in a participative manner	5
Actively listens to others views	4
Builds upon others' ideas	4
Provides assistance to team members	5
Encourages others	4
Demonstrates ability to compromise	5
Surfaces disagreements and deals with them constructively	5

APPENDIX C

TABLES

Table 1

Dimension Coverage Grid for the Assessment Center

	Production Exercise #1	Production Exercise #2	Group Discussion Exercise #1	Group Discussion Exercise #2	Problem Solving Exercise #1	Problem Solving Exercise #2
Work Pace	x	x				
Quality	x	x				
Influence			x	x		
Meeting Participation			x	x		
Teamwork			x	x		
Problem Assessment					x	x
Problem Solution				x	x	x

Table 2

Proportion of behaviors rated as ‘highly diagnostic’ for each dimension.

Dimension	Proportion of high diagnostic behaviors
1. Quality	5/6 or .83
2. Meeting Participation	7/10 or .70
3. Work Pace	2/3 or .66
4. Teamwork	6/10 or .60
5. Influence	3/7 or .43
6. Problem Assessment	2/7 or .28
7. Problem Solution	1/8 or .13

Table 3

Average rank orderings of ‘easiest to evaluate’ (1) to ‘hardest to evaluate’ (7)

Dimension	Average rank order
1. Work Pace	1.5
2. Quality	3.2
3. Meeting Participation	3.8
4. Teamwork	4.2
5. Problem Assessment	4.7
6. Influence	5.2
7. Problem Solving	5.5

Note: A behavior was rated as highly diagnostic if it was given a 5 or greater on the 7 point likert-type scale.

Table 4

Correlation matrix of within-exercise dimension scores

Ex/Dim	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Production Exercise #1																	
1. Work Pace	3.2	1.23	1.00														
2. Quality	4.0	.91	.22	1.00													
Production Exercise #2																	
3. Work Pace	4.5	.77	.11	-.01	1.00												
4. Quality	3.8	.55	.06	-.03	-.02	1.00											
Group Discussion Exercise #1																	
5. Influence	3.4	.71	.07	-.04	.06	-.01	1.00										
6. Meet Part	3.8	.75	.06	-.03	.02	.00	.45	1.00									
7. Teamwork	3.5	.87	-.01	-.04	-.02	.00	.28	.37	1.00								
Group Discussion Exercise #2																	
8. Influence	3.4	.70	.03	-.05	.04	-.01	.38	.33	.28	1.00							
9. Meet Part	3.7	.73	.02	-.04	.01	-.03	.33	.41	.29	.39	1.00						
10. Teamwork	3.5	.65	-.01	-.02	-.05	.00	.21	.29	.37	.33	.38	1.00					
11. Prob Sol.	3.0	.69	.02	.01	.11	-.05	.07	.06	-.02	.10	.05	-.04	1.00				
Problem Solving Exercise #1																	
12. Prob Ass.	3.1	.68	-.01	.00	.08	-.06	.01	.01	.01	.02	.02	-.03	.07	1.00			
13. Prob Sol.	3.4	.67	-.03	.00	.02	-.01	.03	-.04	-.01	.01	-.10	-.07	.07	.22	1.00		
Problem Solving Exercise #2																	
14. Prob Ass.	2.3	.77	.07	.01	.12	-.02	.11	.06	-.04	.00	-.01	-.10	.17	.11	.12	1.00	
15. Prob Sol.	3.6	.87	.04	.03	.07	-.04	.06	.10	.03	.01	.03	-.02	.09	.02	.04	.63	1

Note: High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution

Table 5

Intercorrelations of the dependent variables

	OVERALL	PA	PS	WP	LEARN	FIT	TEAM	QUAL	KNOWL
OVERALL	1.00								
PA	.74	1.00							
PS	.76	.82	1.00						
WP	.69	.54	.56	1.00					
LEARN	.74	.69	.67	.69	1.00				
FIT	.76	.67	.71	.60	.63	1.00			
TEAM	.58	.53	.53	.42	.49	.68	1.00		
QUAL	.74	.67	.64	.65	.64	.62	.48	1.00	
KNOWL	.79	.73	.74	.69	.75	.69	.47	.71	1.00

NOTE: all correlations significant at the .001 level. Dependent variable abbreviations: PA-problem assessment, PS-problem solution, WP-work pace, LEARN-applied learning, FIT-job fit/motivation, TEAM-teamwork, QUAL-quality orientation, KNOW-technical/professional knowledge. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution

Table 6

Confirmatory Factor Analysis results for 10 nested hierarchical multitrait-multimethods for dimension scores

Model	X^2	df	Chi/df	NFI	CFI	AGFI	RMR	RFI
Null: 15 orthogonal factors	1345.08	105	12.80					
1. 7 oblique dimensions and 6 oblique exercises	12.07	46	FIT IS PERFECT					
2. 7 orthogonal dimensions and 6 orthogonal exercises	351.74	85	4.10	.74	.78	.93	.043	.68
3. 7 oblique dimensions and 6 orthogonal exercises	34.59	60	.58	.97	1.00	.99	.014	.95
4. 7 oblique traits and 3 orthogonal exercises	58.90	64	.92	.96	1.00	.98	.018	.93
5. 1 dimension, 6 orthogonal exercises	131.02	81	1.62	.90	.96	.97	.027	.87
6. 1 dimension, 6 oblique exercises	34.50	63	.55	.97	1.00	.99	.014	.96
7. 1 dimension, 3 orthogonal exercises	57.62	77	.75	.96	1.00	.99	.018	.94
8. 7 oblique dimensions	165.75	77	2.15	.88	.93	.96	.03	.83
9. 6 orthogonal exercises	582.89	94	6.20	.57	.61	.89	.056	.52
10.6 oblique exercises	131.75	79	1.67	.90	.96	.97	.027	.87

Table 7

Model comparisons using chi-square difference tests

Model	X^2	df	Comparisons	X^2	df	significant improvement in fit
$Model_{null}$	1345.08	105	M7-M6	23.12	14	*p>.01
$Model_1$	12.07	46	M6-M3	-.09	3	ns.
$Model_2$	351.74	85	M3-M1	22.52	14	ns.
$Model_3$	34.59	60	M4-M3	24.39	4	*p<.01
$Model_4$	58.90	64	M10-M6	97.25	16	*p<.01
$Model_5$	131.02	81				
$Model_6$	34.50	63				
$Model_7$	57.62	77				
$Model_8$	165.75	77				
$Model_9$	582.89	94				
$Model_{10}$	131.75	79				

Note: *p<.01 indicates that the second (less restrictive) model was a significant improvement in fit over the first (more parsimonious model

Table 8

Parameter estimates from a completely standardized solution for model 3: 7 oblique dimensions and 6 orthogonal exercises

	Trait	Method	Error
Production Exercise #1			
Work Pace	.26	.97	0[a]
Quality	.97	.23	0[a]
Production Exercise #2			
Work Pace	.43	-.02	.82
Quality	-.03	1.00	0[a]
Group Discussion Exercise #1			
Influence	.72	-.11	.47
Meeting Participation	.77	-.08	.40
Teamwork	.74	.67	0[a]
Group Discussion Exercise #2			
Influence	.53	.37	.58
Meeting Participation	.53	.44	.53
Teamwork	.74	.02	.58
Problem Solution	.22	.01	.90
Problem Solving Exercise #1			
Problem Assessment	.11	.14	.93
Problem Solution	.08	.67	0[a]
Problem Solving Exercise #2			
Problem Assessment	.54	.55	0[a]
Problem Solution	.27	.55	.51

Note: [a] Parameter fixed at zero for LISREL to converge. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 9

Parameter estimates from a completely standardized solution for model 6: 1 dimensions and 6 oblique exercises

	Trait	Method	Error
Production Exercise #1			
Work Pace	.18	.98	0[a]
Quality	.02	.22	.95
Production Exercise #2			
Work Pace	.34	.03	.88
Quality	-.09	1.00	0[a]
Group Discussion Exercise #1			
Influence	.20	.59	.61
Meeting Participation	.09	.69	.52
Teamwork	-.14	.56	.67
Group Discussion Exercise #2			
Influence	.10	.60	.63
Meeting Participation	.02	.65	.58
Teamwork	-.22	.58	.61
Problem Solution	.34	.06	.88
Problem Solving Exercise #1			
Problem Assessment	.17	.27	.90
Problem Solution	.16	.70	.48
Problem Solving Exercise #2			
Problem Assessment	.42	.56	.51
Problem Solution	.18	.98	0[a]

Note: [a] Parameter fixed at zero for LISREL to converge. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 10

Parameter estimates from a completely standardized solution for model 7: 1 dimension and 3 orthogonal exercises

	Trait	Method	Error
Production Exercise #1			
Work Pace	.14	.99	0[a]
Quality	.00	.22	.95
Production Exercise #2			
Work Pace	.27	.07	.92
Quality	-.08	.07	.99
Group Discussion Exercise #1			
Influence	.27	.56	.61
Meeting Participation	.19	.65	.55
Teamwork	-.06	.54	.70
Group Discussion Exercise #2			
Influence	.09	.57	.67
Meeting Participation	.03	.61	.62
Teamwork	-.22	.58	.61
Problem Solution	.34	.03	.88
Problem Solving Exercise #1			
Problem Assessment	.19	.01	.96
Problem Solution	.14	.04	.98
Problem Solving Exercise #2			
Problem Assessment	.45	.89	0[a]
Problem Solution	.24	.58	.60

Note: [a] Parameter fixed at zero for LISREL to converge. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 11

Proportion of Variance Explained by Trait Factors for Models 3: 7 oblique dimensions and 6 orthogonal exercises, 6: 1 dimension and 6 oblique exercises, and 7: 1 dimension and 3 orthogonal exercises

	Model		
	3	6	7
Production Exercise #1			
Work Pace	07%	03%	02%
Quality	95%	0%	0%
Production Exercise #2			
Work Pace	19%	12%	07%
Quality	0%	01%	01%
Group Discussion Exercise #1			
Influence	52%	04%	09%
Meeting Participation	59%	01%	07%
Teamwork	55%	02%	0%
Group Discussion Exercise #2			
Influence	28%	01%	01%
Meeting Participation	28%	0%	0%
Teamwork	55%	05%	05%
Problem Solution	05%	12%	12%
Problem Solving Exercise #1			
Problem Assessment	01%	03%	04%
Problem Solution	01%	03%	02%
Problem Solving Exercise #2			
Problem Assessment	29%	18%	20%
Problem Solution	07%	03%	06%
AVERAGE	**29%**	**5%**	**5%**

Note: High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 12

Convergent and discriminant validity coefficients for high ease of evaluation dimensions

Dimension	r (monotrait-heteromethod correlations)	mean heterotrait-monomethod correlation by dimension
Quality	-0.03	0.12
Teamwork	0.37	0.29
Workpace	0.11	0.12
Meeting Participation	0.41	0.29
Grand mean	0.22	

Heterotrait-heteromethod correlation	
Grand mean	0.05

Exercise	r (heterotrait-monomethod correlations)
Production Exercise #1	-0.02
Production Exercise #2	0.22
Group Discussion Exercise#1	0.37
Group Discussion Exercise#2	0.20
Grand mean	0.19

Table 13

Criterion and discriminant validity coefficients for low ease of evaluation dimensions

Dimension	r (monotrait- heteromethod correlations	mean heterotrait-monomethod correlation by dimension
Problem Solution	0.07	0.35
Problem Assessment	0.10	0.43
Influence	0.39	0.29
Grand mean	0.19	

Heterotrait-heteromethod correlation	
Grand mean	0.08

Exercise	r (heterotrait-monomethod correlation)
Group Discussion Exercise #1	0.37
Group Discussion Exercise #2	0.20
Problem Solving Exercise #1	0.22
Problem Solving Exercise #2	0.63
Grand mean	0.35

Table 14

LISREL estimates of the latent factor intercorrelations for model 3

Factor	1	2	3	4	5	6	7
Dimension							
1. Work Pace	1.00						
2. Quality	- .03	1.00					
3. Influence	.25	-.07	1.00				
4. Meeting participation	.15	-.05	.82	1.00			
5. Teamwork	-.08	-.05	.67	.74	1.00		
6. Problem Assessment	.40	.01	.14	.07	-.11	1.00	
7. Problem Solution	.49	.07	.28	.25	.02	.83	1.00

Note: High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution

Table 15

Correlations of the within-exercise dimension scores and the dependent variables.

	OVERALL	PA	PS	WP	LEARN	FIT	TEAM	QUAL	KNOW
Production Exercise #1									
workpace	.01	.03	.05	-.06	.04	-.02	**-.15**	-.03	.07
quality	.05	.01	.03	-.04	.05	-.01	-.05	.01	.06
Production Exercise #2									
workpace	0	.01	.06	-.01	.02	.04	-.03	-.01	.07
quality	.03	.03	.06	-.04	0	-.06	.04	0	-.04
Group discussion exercise #1									
teamwork	-.01	.02	-.02	.05	0	.02	.11	-.01	-.04
meeting	.09	.07	.12	.03	.04	.11	.11	.01	.08
influence	**.14**	.13	.13	-.01	.10	.07	.07	.04	.08
Group Discussion exercise #2									
teamwork	.10	.12	.11	.11	.02	.12	**.14**	.05	.08
meeting	.06	.07	.07	.05	.02	.09	.09	.01	.11
influence	.03	.10	.12	-.03	-.03	.09	.05	-.02	.02
problem sol.	.06	.07	.09	.02	.01	.02	.04	.03	.09
Problem Solving Exercise #1									
Problem id	-.04	0	0	-.09	-.03	0	-.02	.01	.02
Problem sol.	-.04	-.06	-.08	-.09	-.08	-.04	-.02	-.07	-.05
Problem Solving Exercise #2									
Problem id	-.04	.02	.01	-.06	-.04	-.02	0	-.03	-.07
Problem sol.	-.08	-.10	-.11	-.12	-.06	-.12	-.02	-.05	-.11

Note: Bolded correlations are significant at $p < .01$. Dependent variable abbreviations: PA-problem assessment, PS-problem solution, WP-work pace, LEARN-applied learning, FIT-job fit/motivation, TEAM-teamwork, QUAL-quality orientation, KNOW-technical/professional knowledge. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 16

Regression equations predicting the dependent variables

Predicting overall performance:

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.112	.013	.855	.013	.855	.953	.491
LOW EASE	.146	.021	.832	.009	.803	.954	.493

Predicting quality average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.040	.002	.109	.002	.109	.925	.979
LOW EASE	.064	.004	.158	.002	.224	.929	.880

Predicting knowledge average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.157	.025	1.73	.025	1.73	.84	.143
LOW EASE	.169	.029	1.14	.004	.37	.839	.772

Predicting problem assessment average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.104	.011	.755	.011	.556	.855	.556
LOW EASE	.168	.028	1.13	.017	.187	.852	.187

Predicting problem solution average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.138	.019	1.34	.019	1.34	.931	.257
LOW EASE	.193	.037	1.51	.018	1.73	.927	.161

Table 16 (cont.)

Regression equations predicting the dependent variables

Predicting workpace average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.118	.014	.970	.014	.970	.901	.424
LOW EASE	.171	.029	1.17	.015	1.43	.899	.235

Predicting learning average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.066	.004	.299	.004	.299	.824	.879
LOW EASE	.110	.012	.474	.008	.710	.825	.547

Predicting jobfit average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.132	.018	1.23	.018	1.23	.980	.300
LOW EASE	.158	.025	.994	.007	.69	.982	.559

Predicting teamwork average

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.223	.050	3.58	.050	3.58	.900	*.007
LOW EASE	.224	.051	2.05	.001	.05	.905	.985

Predicting promotion

Variable	Mult R	R2	F(Eqn)	ΔR2	Fch	SE	Sig.Fch
HIGH EASE	.192	.037	14.81	.037	14.82	.277	*.000
LOW EASE	.248	.062	14.51	.025	13.61	.274	*.000

Table 17

Coefficients for predicting teamwork

Model	Dimension	Beta	t	sig.
1	WP	-.16	-2.59	*.010
	QUAL	.02	.30	.765
	MP	.08	1.22	.222
	TMWRK	.11	1.64	.102
2	WP	-.16	-2.60	*.010
	QUAL	.02	.31	.755
	MP	.08	1.05	.297
	TMWRK	.11	1.63	.104
	PRID	.02	.34	.733
	PRSOL	.00	.01	.993
	INFL	.00	.08	.937

Note: High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 18

Coefficients for predicting promotion

Model	Dimension	Beta	t	sig.
1				
	WP	.10	4.04	*.000
	QUAL	.04	1.49	.136
	TMWRK	-.16	-5.74	*.000
	MP	.04	1.45	.147
2				
	WP	.08	3.24	*.001
	QUAL	.05	1.86	.063
	TMWRK	-.17	-5.79	*.000
	MP	-.01	-2.52	.801
	PRID	.09	3.02	*.003
	PRSOL	.07	2.53	*.011
	INFL	.09	2.93	*.003

Note: High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 19

Intercorrelations of the exercise scores

	Prodex1	prodex2	group1	group2	problem1	problem2
Prodex1	1.00					
Prodex2	.04	1.00				
Group1	.09	-.01	1.00			
Group2	.02	.02	**.47**	1.00		
Problem1	-.01	.04	-.01	.01	1.00	
Problem2	.14	.02	.04	.01	.08	1.00

Note: Bolded correlations indicates $p < .01$. Abbreviations for the exercises: Prodex1-production exercise #1, Prodex2-production exercise #2, Group1-group discussion exercise #1, Group2-group discussion exercise #2, Problem1-problem solving exercise #1, Problem2-problem solving exercise #2. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 20

Correlations of exercise scores with dimension scores

	Mean	SD	WP	QUAL	MP	TMWRK	INFL	PRID	PRSOL
Prodex1	3.62	.84	**.73**	**.62**	.01	-.03	.01	.04	.03
Prodex2	4.10	.47	**.51**	**.27**	.01	-.03	.05	**.08**	.06
Group1	3.56	.54	.06	-.05	**.76**	**.65**	**.72**	.05	**.08**
Group2	3.40	.44	.04	-.06	**.69**	**.61**	**.67**	.04	**.23**
Problem1	3.25	.53	.01	-.03	-.05	-.03	.03	**.59**	**.45**
Problem2	2.95	.74	**.10**	0	.06	-.04	.06	**.68**	**.69**

Note: Bolded correlations indicate $p < .01$. Abbreviations for the exercises: Prodex1-production exercise #1, Prodex2-production exercise #2, Group1-group discussion exercise #1, Group2-group discussion exercise #2, Problem1-problem solving exercise #1, Problem2-problem solving exercise #2. High ease of evaluation dimensions were work pace, quality, meeting participation, and teamwork. Low ease of evaluation dimensions were influence, problem assessment, and problem solution.

Table 21

Summary of Multitrait-Multimethod Correlations from Previous Construct Validity Studies

Study	Purpose	N	mthm	htmm	hthm
Archambeau(1979)	selection	29	.10-.51	.33	na
Sakett & Dreher(1982)	selection	86	.07	.64	.06
	selection	311	.11	.4	.07
	selection	162	.51	.45	.45
Turnage & Muchinsky(1982)	selection	1028	.18-.70	.51-.90	na
	selection	1028	.20-.69	.52-.90	na
Silverman et al.(1986)	selection	45	.54	.65	.44
	selection	45	.37	.68	.31
Robertson et al.(1987)	selection	41	.28	.64	na
	recruitment	48	.26	.66	na
	recruitment	84	.23	.60	na
	selection	49	.11	.49	na
Bycio et al.(1987)	selection/ development	1170	.36	.75	na
Russell(1987)		75	.53	.25	na
Schneider & Schmitt(1992)	development/ research	89	.25	.72	.22
Harris et al.(1993)	selection	237	.32	.42	na
		556	.33	.41	na
		63	.33	.46	na
Fleenor(1996)	development	102	.22	.42	.16
Kudisch & Ladd(1997)	diagnosis	138	.29	.41	.16
CURRENT STUDY	selection	1555			
High Ease Dimensions			.21	.19	.05
	Range (-.03 to .41)				
Low Ease Dimensions			.19	.35	.08
	Range (.07 to .39)				

Note: Table adapted from Schneider & Schmitt, 1992 and Kudisch et al., 1997

APPENDIX D

FIGURES see Chapter3.pdf