4-2001

# Understanding the assessment center process: Where are we now?

Filip LIEVENS
*Singapore Management University*, filiplievens@smu.edu.sg

Richard J KLIMOSKI

# Understanding the Assessment Center Process: Where Are We Now?

Filip Lievens

University of Ghent, Belgium

Richard J. Klimoski

George Mason University

ACKNOWLEDGEMENTS

Filip Lievens is postdoctoral research fellow of the Fund of Scientific Research, Flanders. The authors would like to acknowledge Lisa Donahue for suggestions on an earlier version of this chapter.

CORRESPONDENCE

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, University of Ghent, Henri Dunantlaan 2, 9000 Ghent, Belgium. Electronic mail may be sent via Internet to [filip.lievens@rug.ac.be].

*The reference for this paper is: Lievens, F., & Klimoski, R.J. (2001). Understanding the assessment center process: Where are we now? In C.L. Cooper & I.T. Robertson (Eds.) International Review of Industrial and Organizational Psychology vol. 16. (pp. 245-286). Chicester: John Wiley & Sons, Ltd.*

# Understanding the Assessment Center Process: Where Are We Now?

## Introduction

Assessment centers have become widespread in Western Europe, Northern America, and Australia (Newell & Shackleton, 1994). The Task Force on Assessment Center Guidelines (1989) defined assessment centers as "a standardized evaluation of behavior based on multiple inputs. Multiple trained observers and techniques are used. Judgments about behaviors are made, in major part, from specifically developed assessment simulations. These judgments are pooled in a meeting among the assessors or by a statistical integration process" (p. 460).

Originally, the assessment center method was considered to be an alternative measurement instrument to estimate predictor-criterion relationships. The vast majority of research also dealt with criterion-related validity and demonstrated that assessment centers were predictive for a variety of criteria of managerial effectiveness. Yet, through the years the original conceptualization of assessment centers has changed dramatically (Howard, 1997). Three changes seem most noteworthy. First, whereas the output of assessment centers is still important, much more attention has been paid to assessment center 'processes'. This is most strongly reflected in the research on the construct validity of assessment centers. A second change is that the application of assessment centers has moved beyond selection/placement/promotion purposes. Recent surveys (e.g., Spychalski, Quinones, Gaugler, & Pohley, 1997) show that assessment centers are increasingly used for developmental purposes. As noted by Kudisch, Ladd, and Dobbins (1997) the goals of these developmental assessment centers vary from identification of participants' training needs, to formulation of personalized developmental recommendations and action plans, to skill development on the basis of immediate feedback and on-site practice. A third change is that

nowadays multiple stakeholders are involved in assessment centers. These stakeholders include assessees, assessors, assessment center users, and the organization.

This chapter aims to provide a contribution relative to two of these changes. More specifically, we aim to provide a better understanding of the individual and collective processes and factors that affect the quality of assessor decisions. Hereby we primarily focus on the factors and forces, which affect the capacity of assessment centers to provide construct valid estimates of individual attributes. This would seem to be most central to developmental assessment centers because such applications, by definition, need to produce 'true' and valid assessments of an assessee's strengths and weaknesses on the various dimensions. Moreover, developmental assessment centers assume that participants accept and act upon the feedback built around these assessments in the belief of their intrinsic validity (Thornton, Larsh, Layer, & Kaman, 1999). Thus, the quality of assessor decisions is at the core of acceptance of feedback and the motivation to thereby pursue developmental training activities. That said, it is also our view that the quality of assessor decisions in terms of construct measurement is also important for other applications (e.g., selection) as it gets to the heart of the method. In reviewing the recent literature, we will start with a relatively simple scheme adopted from the performance appraisal literature. Whereas we will treat it as a useful devise for organizing the studies of interest, we will go on to argue that a more complex view will be needed as a roadmap for future research- research that will lead to a deeper understanding of the assessment center method.

The basis for our insight into the processes and factors affecting the quality of assessor decisions in assessment centers stems from our review of the literature published between 1990 and1999. We conducted this search for relevant studies using a number of computerized databases (i.e., PsycLit, the Social Science Citation Index, Current Contents, and Dissertations Abstracts International). Additionally, we scrutinized reference lists from studies to find other published and unpublished studies. We did not only look for studies conducted in the US, but also searched for studies conducted in other countries.

We will use Landy and Farr's (1980, p. 73) component model of performance rating as a framework for organizing the studies. This framework is comprised of five classes of variables: (a) the roles (e.g., raters and ratees), (b) the rating context (e.g., rating purpose), (c) the rating vehicle (e.g., rating instrument), (d) the rating process, and (e) the results (e.g., rating information and actions based upon it). The structural relationship between these variables are as follows. Roles, context, and vehicle are expected to influence the rating process, which, in turn, should affect the results. Although this model was originally proposed in the broader field of performance rating, we feel it has heuristic value, making the various components easily transferable to (developmental) assessment centers. For instance, in this application 'roles' refer to assessors, assessees, and role-players and 'results' refer to the ratings of assessees' strengths and weaknesses, the developmental feedback formulated, and the action plans (including any training and developmental assignments) suggested. The remainder structures the studies considered in terms of these five components. This will take the form of an elaboration of the Landy and Farr (1980) framework as portrayed in Figure 1.

Note that throughout this chapter we make reference to the notion of the 'quality' of assessment center judgements. Quality is operationalized or indexed in various ways in the studies under review. In most of these studies 'quality' is a shorthand way of referring to the degree of convergent and discriminant validity present in dimension ratings. To this end, authors examined patterns found in the multitrait-multimethod matrix, exploratory or confirmatory factor analysis, and correlations with external criteria. In other studies quality of assessor ratings was operationalized as dimensional accuracy, lack of bias in ratings, and even as positive reactions to assessment center (trait) ratings. Finally, whereas the traditional application of assessment centers has often relied on the strength of the correlation of overall ratings with job performance, we will view this evidence as reassuring but not definitive when assessment center results are used as the source of developmental feedback.

---

Insert Figure 1 about here

---

# Roles

## Assessor Characteristics

Between 1990 and 1999 a first group of studies examined assessor characteristics as these may affect the assessment center rating process and, eventually, the quality of assessor ratings and decisions. Assessor characteristics refer to personal attributes (i.e., demographic and personality characteristics), assessor type, assessor source, and assessor training.

With respect to demographic characteristics of assessors, Lowry (1993) conducted a comprehensive study investigating the effects of age, race, education, rank, tenure, prior assessment center experience, managerial experience, and experience with the target job. Only assessor age and assessor rank exerted significant effects on the ratings. However, these two assessor characteristics accounted for less than 2% of the variance in ratings.

Other studies concentrated solely on the effects of assessor gender. Most of these studies found that ratings of male and female assessors did not differ significantly from each other (Binning, Adorno, & Williams, 1995; Weijerman & Born, 1995). Shore, Tashchian and Adams (1997), however, reported that in a role-play on four dimensions female assessors gave significantly higher ratings to both men and women assessees than did male assessors. In two other role-play exercises no effect of assessor gender was found. It is also possible that the gender of the assessee and the gender of the assessor interact to produce differences in assessment results. An earlier study (Walsh, Weinberg, & Fairfield, 1987) reported such a significant assessee-assessor gender interaction. In this case all-male assessor groups rated female candidates for a professional sales position significantly higher than male candidates. Yet, two recent studies failed to replicate this interaction effect (Shore et al., 1997; Weijerman & Born, 1995).

Bartels and Doverspike (1997a) focused on the personality characteristics of assessors (measured by the 16 PF) and how they impacted on leniency in assessment

center ratings. They found assessors high on intelligence, sensitivity, and poise to be more lenient.

Another variable studied was the type of assessor. Sagie and Magnezy (1997) compared ratings of psychologist assessors and managerial assessors in terms of convergent and discriminant validity. A confirmatory factor analysis of the ratings of psychologists revealed that the factors represented all five predetermined dimensions. Ratings of managers, however, yielded only two dimension factors. Lievens (1999) found that managerial assessors distinguished somewhat less among dimensions than industrial and organizational psychology students. Yet, managers provided significantly more accurate ratings than these students. In this study accuracy was determined by the extent to which ratings were consistent with the values and norms espoused by the organization.

Several studies compared assessor ratings, self-ratings, and peer ratings to each other. Shore, Shore, and Thornton (1992) concluded that construct-related evidence in assessor ratings was stronger for peer ratings than for self-ratings. Shore, Tetrick, and Shore (1998) examined whether assessor, peer, and self-ratings were based on the same types of information when making overall assessments of managerial potential. The found support for the hypothesis that self-assessments of managerial potential were based to a greater extent on information not generated in the assessment center itself. Yet, a counterintuitive finding was that assessor ratings and peer ratings (instead of self-ratings) were most dissimilar. Results of other studies (Clapham, 1998; Shechtman, 1998) reported more dissimilarity between assessor and self-ratings. In another study Nowack (1997) found that participant self-ratings were significantly associated with overall assessor ratings but not with overall job performance ratings.

A last assessor factor studied was the training given to assessors. Maher (1995) focused on the effects of different lengths of assessor training. Two days of assessor training increased accuracy more than one day. Yet, adding a third day made no significant improvement. In other words, beyond a threshold level, additional assessor training was not useful. Lievens (1999) compared different training types (i.e., data-driven assessor training,

schema-driven assessor training, and control training). The data-driven assessor training taught assessors to strictly distinguish various rating phases (e.g., observation, classification, and evaluation) from each other and to proceed to another phase, only when the previous one was finished. Alternatively, schema-driven assessor training taught raters to use a specific performance theory as a mental scheme to 'scan' the behavioral stream for relevant incidents and to place these incidents –as they were observed- in one of the performance categories. Results showed that the data-driven and schema-driven assessor training approaches outperformed the control training in terms of inter-rater reliability, dimension differentiation, and differential accuracy. The schema-driven assessor training resulted in the largest values on all three dependent variables. In a similar study Schleicher, Day, Mayes, and Riggio (1999) compared frame-of reference training, which conceptually builds on schema-driven principles, to no assessor training. Frame-of reference training resulted in ratings with significantly higher inter-rater reliability, discriminant validity, and criterion-related validity.

Assessee Characteristics

In this section we discuss studies that investigated whether characteristics of assessees impact on assessor ratings. Personal characteristics such as race, gender or age were most frequently studied. Additionally, this research stream examined effects of assessee performance variability and assessee coaching. The remainder discusses the results of these studies.

Hoffman and Thornton (1997) summarized earlier studies on assessee race effects and concluded that these studies were almost evenly split between studies showing no significant rating differences and studies showing Whites receiving higher ratings on average than other ethnic groups, usually less than one standard deviation. Recent studies confirmed this picture. Schmitt (1993) analyzed data from the selection of school administrators and found rating differences between Black and White candidates (over one half of a standard deviation). However, Bobrow and Leonards (1997) found no such differences. Whereas

these studies focused on Black-White differences, Ramos (1992) reported that assessors scored Hispanics up to half a standard deviation lower than Whites in the AT&T assessment centers on some criteria but validity against a promotion criterion was as high as for Whites. In a South African assessment center Kriek, Hurst, and Charoux (1994) did not find significant differential validity in predicting performance among Whites, Blacks and colored male supervisors.

Goldstein, Yusko, Braverman, Smith, and Chung (1998) provided a possible explanation for these mixed results regarding assessee race effects. The degree to which subgroup (Black-White) mean differences occurred in assessor ratings was found to be a function of the type of exercise rated. Moreover, Goldstein et al. (1998) reported that the subgroup differences varied by the cognitive component of the exercise. In other words, race effects were more apparent in ratings, if an assessment center consisted of more exercises with a cognitive component (e.g., in-basket). Similarly, Sackett (1998) concluded that ratings of oral exercises included in an assessment center for lawyers displayed smaller subgroup differences than ratings of written exercises. Contrary to these conclusions, Rotenberry, Barrett, and Doverspike (1999) demonstrated that the underlying structure of in-basket ratings of 3399 safety personnel was invariant between races. The lesson learned from these studies is that it may be preferable to inspect the ratings made in the specific assessment center exercises (instead of the overall assessment center ratings) for race effects. Along these lines, Baron and Janman (1995) signaled the dearth of research about possible race effects in ratings of fact-findings, presentations, group exercises, or role-plays.

The gender of assessees should not affect the ratings of assessors. In other words, assessor ratings should reflect that men and women perform equally well in assessment centers and that assessment centers are equally valid predictors of future performance for men and women. Research by Weijerman and Born (1995) confirmed this assumption, as ratings of managerial potential of 77 Dutch civil servants were not biased by the gender of the candidates. Bobrow and Leonards (1997) and Rotenberry et al. (1999) reported similar results.

Nevertheless, in other studies ratings were prone to subtle gender bias, favoring women candidates. For instance, in Schmitt (1993) ratings indicated small performance differences in favor of female candidates. Neubauer (1990) found women received slightly higher ratings in a German high school career assessment center. In another study (Shore, 1992) 375 men and 61 women were assessed on their intellectual ability, performance-related and interpersonally related skills, and overall management potential. Although there were no significant differences between men and women in overall management potential ratings or in long-term job advancement, women obtained consistently higher ratings on performance-related skills.

Related to the above, Halpert, Wilson, and Hickman (1993) investigated whether people provided significantly different ratings to the videotaped assessment center performance of either a pregnant woman or a non-pregnant woman. Ratings of 2239 undergraduates revealed that the pregnant woman was consistently rated lower and that male undergraduates assigned significantly lower ratings than females.

With respect to the effects of assessee age on ratings, the results are again equivocal. Bobrow and Leonards (1997) analyzed ratings from an operational assessment center and found very small differences between candidates younger than 40 and candidates 40 and older. However, after controlling for education, years of service, and gender Clapham and Fulford (1997) reported negative correlations between candidate age and assessment center ratings. In particular, candidates younger than 40 received significantly higher ratings than candidates older than 40.

Morrow, McElroy, Stamper, and Wilson (1990) developed eight simulated assessment center candidates which varied on physical attractiveness (high vs. low), age (less than 40 years of age vs. more than 40 years of age), and gender (male vs. female). This experimental study revealed a main effect of physical attractiveness in the promotion ratings of 40 personnel professionals, but it explained only 2% of variance. Neither assessee age nor assessee gender significantly affected the promotion ratings.

Fletcher and Kerslake (1993) and Fletcher, Lovatt, and Baldry (1997) found that about 45% of participants reported stress and anxiety during the assessment center. A related question is whether this increased stress and anxiety of some candidates also results in lower assessor ratings. Fletcher et al. (1997) tackled this problem using established measures of state, trait, and test anxiety. They did not report on a relationship between increased anxiety and lower assessment center ratings.

Gaugler and Rudolph (1992) investigated contrast effects in assessment centers. They examined both the effects of between assessee variability and within assessee variability. Regarding between assessee variability a poor candidate in a generally 'good' group was rated significantly lower than a poor candidate in a generally 'poor' group. Regarding within assessee variability a low assessee's performance was rated lower when the assessee's prior performance had been dissimilar (i.e., high) than when the assessee's prior performance had been similar (i.e., low). Finally, ratings of assessees displaying performance variation were more accurate than those obtained without performance variation. Assessee performance variability was also the focus of the study of Kuptsch, Kleinmann, and Köller (1998). Contrary to their expectations, they found that people, who perceived their own behavior as more changeable or 'chameleon-like', were rated more consistently than participants who described themselves as more consistent.

A final line of research examined the effects of assessee coaching on assessor ratings. Earlier studies concluded that coaching (e.g., a formal training course or prior experiences) might lead to higher ratings in in-baskets (Brannick, Michaels, & Baker, 1989; Brostoff & Meyer, 1984; Gill, 1982), role-plays (Moses & Ritchie, 1976), leaderless group discussions (Kurecka, Austin, Johnson, & Mendoza, 1982; Petty, 1974), and business plan presentations (Dulewicz & Fletcher, 1982). There is a paucity of recent research in this area and, hence, only a snapshot of the possible coaching tactics (e.g., casual tips, (in)correct grapevine information, behaviorally specific feedback, self-study of workbooks, or comprehensive behavior modeling programs) have been addressed so far. In one exception Mayes, Belloli, Riggio, and Aguirre (1997) used a pretest-posttest design for examining the

effects of two different management courses on assessment center ratings. Whereas the first course used lectures and discussions to teach various organizational behavior domains, the other course taught the same areas with a strong emphasis on experiential activities and skills. The conclusion was that both courses resulted in significantly better dimensional ratings in a role-play, higher overall ratings in an oral presentation, and one higher dimensional in-basket rating. The skills course emerged as significantly more effective than the traditional course in terms of higher role-play and in-basket ratings.

In addition to this lack of research on assessee coaching, we were not able to trace studies on whether assessor ratings are affected by assessee deception or impression management.

Role-player

As noted by Zedeck (1986) role-players are important factors in the assessment center. Trained role players are often used to increase standardization and to evoke dimension-related behavior from assessees. Unfortunately, little is known about their 'role' in the assessment center process. One exception is the unpublished dissertation of Tan (1996), who compared the effects of different types of role-players (i.e., active vs. passive). When role-players performed an active role (i.e., sought to elicit dimension-related behavior), assessor staff ratings showed somewhat higher convergent and discriminant validity. For 'passive' role-players these validities were very low.

## Vehicle

This section deals with the vehicles, which are used in assessment centers to obtain ratings. Logically, studies with respect to the dimensions, the various observation and rating instruments, and the integration procedures are discussed. We also include studies about the assessment center exercises because these exercises serve as vehicles to elicit job relevant information upon which ratings are based. Although there exist guidelines with

regard to the design of these vital assessment center components (Task Force on Assessment Center Guidelines, 1989), several survey studies (Boyle, Fullerton, & Wood, 1995; Lievens & Goemaere, 1999; Lowry, 1996; Spychalski et al., 1997; Van Dam, Altink, & Kok, 1992) showed that their implementation across organizations differed considerably. In this section we review whether such procedural variations influence the rating process and the quality of assessor ratings.

Dimensions

Howard (1997) noted that "[assessment center] dimensions have always been muddled collections of traits (e.g., energy), learned skills (planning), readily demonstrable behaviors (oral communication), basic abilities (mental ability), attitudes (social objectivity), motives (need for achievement), or knowledge (industry knowledge), and other attributes or behaviors" (p. 22). Studies have been conducted on the effects of varying the number, the distinctiveness, the nature, and the observability of these dimensions in terms of the quality of measurement in assessment centers.

A first group of studies varied the number and the level of abstraction of the dimensions rated. The general assumption is that asking assessors to rate a large number of dimensions (e.g., more than 4 or 5) per exercise overburdens the cognitive capabilities of the assessors. Maher (1990) confirmed this and showed that assessors' accuracy diminished when a larger number of dimensions was rated (see Gaugler & Thornton, 1989, for a similar previous study). Campbell (1991) compared the effectiveness of three general performance dimensions (i.e., intellectual/communication skills, interpersonal skills, and administrative skills) and 14 specific dimensions on various aspects of rating quality. The results partially supported the hypothesis that categorization accuracy, rating accuracy, and inter-rater reliability would be significantly greater for the general dimensions than for the specific dimensions. The general dimensions showed also substantially greater evidence of convergent validity than the specific dimensions. No effect on discriminant validity was found. Campbell (1991) concluded that the use of general dimensions showed promise as a method

of reducing the number of dimensions. In similar vein, Kolk, Born, Bleichrodt, and Van der Flier (1998) made a plea to group assessment center dimensions in three broad dimensions. They also found empirical evidence that ‘feeling’, ‘thinking’, and ‘power’ were useful labels of these meta-dimensions.

Kleinmann, Exler, Kuptsch, and Köller (1995) varied the distinctiveness of dimensions. Assessors were expected to have more difficulties distinguishing between dimensions, which were ‘naturally’ related to one another. With this respect, correlations among dimensions might be split up in true (valid) and invalid correlations (see Cooper, 1981; Murphy, Jako, & Anhalt, 1993, for the distinction between ‘true’ and ‘illusory’ halo). Kleinmann et al. (1995) found higher discriminant validity, when assessors rated assessees on conceptually distinct dimensions. With interchangeable dimensions, assessors provided interdependent ratings, which did not differ meaningfully from each other.

Another group of studies experimented with other types of dimensions/constructs. Russell and Domm (1995), for example, explored the effectiveness of an assessment center in which assessors rated candidates on seven role requirements of the target position. For example, they defined the dimension initiative as “the degree to which behaviors influence events to achieve goals by originating action rather than merely responding to events as required on the job of store manager” (p. 30). Nonetheless, there was little evidence that these task-dimensions were actually measured. Joyce, Thayer, and Pond (1994) compared the traditional dimensions to a set of constructs based on the functional structure of managerial work (e.g., internal contacts, performance management, etc.). Within-exercise ratings on these task-oriented dimensions exhibited also weak evidence of convergent and discriminant validity.

Next, many studies attempted to improve the definition and operationalization of dimensions. These studies were prompted by the fact that the behavioral domain of dimensions is often undefined or ill-defined (Kauffman, Jex, Love, & Libkuman 1993). In fact, different meanings are frequently associated with the same dimension and definitions of dimensions are not always clearly related to the behaviors elicited by the exercises.

Additionally, the interpretation of dimension constructs often changes from one exercise to another (Kauffman et al., 1993, Reilly, Henry, & Smither, 1990). For example, leadership in a group discussion (i.e., meeting leadership) would likely differ from leadership in a role-play with a subordinate (i.e., individual leadership). We will deal with the body of research on dimension definition and operationalization in the context of behavioral checklists.

Another group of studies looked at the impact of the observability of the dimensions (Reilly et al., 1990). These studies were based on the principle of aggregation (Epstein, 1979), which states that the sum of a set of measurements is more stable than any single measurement from the set. Analogous to testing, exercise ratings of a dimension can be viewed as 'single items'. When an exercise elicits few items (read behaviors) relevant to a dimension, the representativeness of the assessee behavior for the construct domain is insufficient to obtain a consistent measure of the dimension (Kleinmann & Köller, 1997).

Empirical studies reveal mixed support for this principle in the context of assessment centers. On the one hand prior research showed that there exist wide variations in the opportunity to display dimension-related behaviors across exercises (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Reilly et al., 1990). For instance, in the Reilly et al. (1990) study the number of behaviors varied from 4 behaviors for one dimension to 32 behaviors for another dimension. Further, Reilly and colleagues discovered that the opportunity for assessors to observe dimension-related behavior (indicated by the number of items in a behavioral checklist) was related to the ratings on these dimensions. This relatively strong curvilinear relationship suggested that the correlation between observed behavior and ratings was a function of the number of behavioral checklist items up to certain point (i.e., 12 items), beyond which the relationship remained stable. Finally, Shore et al. (1992) concluded that construct-related evidence in assessor ratings was stronger for more observable dimensions than for dimensions requiring more inferential processes on the part of assessors.

On the other hand prior research also raised doubt on the effects of observability on the quality of assessor ratings. Kleinmann et al. (1995) experimentally manipulated the observability of dimensions (a priori rated by expert assessors) and found no differences

between highly observable dimensions and poorly observable dimensions in terms of construct validity. In similar vein, Campbell (1991) did not report higher rating accuracy when relevant behaviors were displayed with high frequency than when they were displayed with low frequency.

A final set of studies took a closer look at the dimensions rated in assessment centers. In a sophisticated study Guldin and Schuler (1997) chose dimensions which systematically varied concerning their conceptual proximity to the trait concept. They discovered that between 34% and 55% of the true score variance was related to cross-situational relative interindividual differences. Dimensions such as activity and communication skills were most likely to be classified as trait-like. In similar vein, Tett (1998, 1999) called for careful consideration of the nature of the traits used in assessment centers and the process by which these traits find expression in behavior. He proposed the principle of trait activation, which holds that the behavioral expression of a trait requires arousal by trait-relevant situational cues (i.e., assessment center exercises). On the basis of this interactionist approach, Tett (1998, 1999) hypothesized that cross-exercise consistency in assessor ratings can be expected only when exercises shared trait-expressive opportunities. Results based on responses to two versions of an in-basket exercise (Ns = 61, 63) supported this trait activation hypothesis.

## Simulation Exercises

Generally, assessment center exercises may be divided in three groups: individual exercises (e.g., in-basket, planning exercise, case analysis), one-to-one exercises (e.g., role-play, fact-finding, presentation), and group exercises (e.g., leaderless group discussion). These exercises are developed to represent the most important elements of the target job (see Ahmed, Payne, & Whiddett, 1997, for a procedure to develop assessment center exercises). Because job demands and tasks are quite diverse, assessees often perform in different types of exercises, which may result in a weak consistency of ratings across exercises (i.e., low convergent validity). Researchers have explored several characteristics of

assessment center exercises as possible determinants of this weak across-exercise consistency in assessor ratings. These characteristics include exercise form, exercise content, and exercise instructions.

Schneider and Schmitt (1992) experimentally manipulated the effects of exercise content and exercise form. Variance due to the form of the exercise (e.g., role-play vs. group discussion) emerged as the most important exercise factor to bolster different ratings across exercises. More specifically, exercise form explained 16% of the exercise variance in ratings. The effect of exercise content (competitive vs. cooperative) was negligible.

Highhouse and Harris (1993) examined the nature of the exercises in the typical assessment center and their effects on ratings. First, assessee behaviors were extracted from assessor report forms. Grouping similar behaviors into clusters yielded a list of 25 so-called performance constructs (e.g., maintains composure, generates enthusiasm, asks questions, etc.) used by assessors. Then, experienced assessors were asked to use these performance constructs to describe the ideal assessment center candidate in each exercise. Highhouse and Harris (1993) concluded that assessors perceived the exercise situations to be generally unrelated in terms of the behaviors required for successful performance. They also discovered some evidence for the hypothesis that assessees would be rated more consistently in exercises that were perceived to be more similar. For example, ratings of candidates in the simulated phone-call and fact finding exercises were relatively consistent, and assessors also saw these exercises as more similar. Further, assessors perceived the group discussion and scheduling exercises to be quite different situations, and ratings of candidate performance in these exercises appeared to be less consistent. However, the relationship between perceived similarity in exercise content and actual consistency in assessee performance ratings across these exercises was not confirmed in other exercises.

Besides the usual exercise instructions Kleinmann and his colleagues (Kleinmann, 1993; Kleinmann, Kuptsch, & Köller, 1996; Kleinmann, 1997) made the dimensions rated transparent to assessees. Assessees were also informed which behaviors were relevant per dimension. Because in this case assessees oriented themselves more towards the given

dimensions and demonstrated more clearly and consistently the accompanying behaviors, the quality of assessor ratings improved. Specifically, assessors were better able to provide distinct ratings (within exercises) and consistent ratings (across exercises). Nonetheless, Kleinmann (1997) discovered that divulging dimensions resulted in lower criterion-related validity for the transparent group. Smith-Jentsch (1996) also reported negative side-effects of transparent dimensions. Skill transparency was found to reduce the convergence between dimension ratings in a situational exercise and personality inventory scores, and the correlation between dimension ratings and self-reported performance one year later.

Observation and Rating Instrument

In the original AT&T assessment centers assessors took notes while observing candidates and afterwards used this information to rate the candidates. However, through the years several alternatives have been suggested to improve the quality of ratings. Behavioral checklists constitute one of the most popular options (Boyle et al., 1995; Spychalski et al., 1997). An advantage of behavioral checklists is that assessors are not required to categorize behavior. Instead, they can concentrate their efforts on the observation of relevant behaviors. As argued by Reilly et al. (1990), the checklists may further reduce cognitive demands by serving as retrieval cues to guide the recall of behaviors observed. However, according to Joyce et al. (1994) a drawback of behavioral checklists may be that they redefine a dimension from one exercise to another. In this way the increased behavioral focus and specificity of behavioral checklists may contribute to the low correlations among dimension ratings across exercises.

The research evidence with regard to the effectiveness of behavioral checklists is mixed. Reilly et al. (1990) reported positive findings because ratings made via behavioral checklists demonstrated higher convergent and somewhat higher discriminant validity than ratings without the use of behavioral checklists. In other studies behavioral checklists only enhanced discriminant validity (Donahue et al., 1997) or had virtually no effects (Fritzsche, Brannick, & Fisher-Hazucha, 1994; Schneider & Schmitt, 1992). Hennessy, Mabey, and Warr

(1998) compared three observation procedures: traditional note taking, use of a behavioral checklist, and behavioral coding. The methods were found to yield similar outcomes in terms of accuracy of judgement, accuracy of written evidence, correlation between dimension ratings, and attitude toward the method employed, with a slight preference for behavioral coding.

Recent studies also examined more specific aspects related to behavioral checklists. For example, Binning, Adorno, and Kroeck (1997) found that the discriminant validity of behavioral checklists increased only when the items were ordered in naturally occurring clusters. The discriminant validity of a randomly ordered checklist was low. Another specific aspect pertains to the number of items per dimension in checklists. With this respect, Hauenstein (1994) argued to list only the key behaviors. Reilly et al. (1990) supported this 'key behavior' approach and determined that the optimal number of statements per dimension varied between six and twelve. Lebreton, Gniatczyk, and Migetz (1999) also supported the use of shorter checklists. They demonstrated that checklists with fewer behavioral items and dimensions (e.g., 2 dimensions comprised of 14 behaviors instead of 6 dimensions made up of 45 behaviors) are to be preferred in light of predictive and construct validity.

Besides behavioral checklists, videotaping of assessees has also been used to assist assessors in their task and to improve the quality of their assessments. In particular, Ryan and colleagues (1995) hypothesized that giving assessors the opportunity to rewind and pause videotaped assessment center exercises would improve the information processing capacities of assessors. Nonetheless, they concluded that the impact of the use of videotaping assessees on ratings was minimal. In particular, rewinding and pausing the videotape had some beneficial effects on behavioral accuracy of assessors but did not increase rating accuracy.

With respect to rating procedures Harris, Becker, and Smith (1993) and Kleinmann, Andres, Fedtke, Godbersen, and Köller (1994) examined whether a variant of the behavior reporting method, the within-dimension method, showed higher convergent and discriminant

validity. In the traditional behavior reporting method "evaluation is postponed until the completion of all exercises, at which time the assessors share their observations and rate the candidates on a series of dimensions" (Sackett and Dreher, 1982, p. 402). According to the within-dimension rating method candidates are rated on each dimension upon completion of each exercise. Contrary to earlier findings (Silverman, Dalessio, Woods, & Johnson, 1986), both studies (Harris et al., 1993; Kleinmann et al., 1994) reported no beneficial effects for the within-dimension method.

Finally, the rotation scheme of assessors through the various exercises has been found to influence the quality of ratings. A first rotation scheme issue relates to the ratio of assessors to assessees. Lievens (in press) used generalizability analysis to examine the effects of reducing or increasing the number of assessors per assessee. Reducing the number of manager assessors from 3 to 1 had a serious impact on the generalizability coefficient as it dropped from .81 to .60. A second issue deals with the fact that in operational assessment centers, each assessor does not rate each candidate in every exercise. For example, a candidate might be rated by one assessor in an in-basket, and by a second assessor in a role-play exercise. Even if the candidate's behavior was consistent across exercises, very dissimilar ratings could result from low inter-rater agreement between assessors. Research by Adams and Osburn (1998) confirmed this expectation. This study also demonstrated that it is important to identify a rotation scheme, which minimizes rater inconsistencies. Andres and Kleinmann (1993) developed such a rotation system for reducing information overload, contrast effects, halo effects, and sympathy effects. No studies have empirically demonstrated the superiority of this rotation scheme.

### Integration Procedure

At the end of the assessment center assessors typically meet to discuss observations and ratings. Survey studies show that this formal assessor discussion is almost always held. For instance, the survey of Spychalski et al. (1997) of US assessment center practices indicated that 84.1% of the organizations held a consensus discussion to integrate ratings. In

Boyle's et al. (1995) survey of UK assessment center practices this percentage reached 96%.

Despite this popularity, we traced only a couple of studies (conducted between 1990-1999) on the effects of the integration procedure. Firstly, studies examined the superiority of mechanically-derived versus consensus-derived integration procedures in terms of predictiveness. Pynes and Bernardin (1992) found no difference in terms of predictive validity between mechanically-derived and consensus-derived integration procedures. Lebreton, Binning, and Hesson-McInnis (1998), however, showed that clinical judgements were superior to statistically-combined ratings. Secondly, Anderson, Payne, Ferguson, and Smith (1994) inspected how assessors integrated the information from various sources in the consensus discussion. They concluded that assessors relied more on information elicited first-hand (i.e., observational data in assessment center exercises) than on biodata or psychometric test scores.

This paucity of studies illustrates that Zedeck's (1986) point that "group dynamics seems to be totally ignored within the assessment center literature" (p. 290) is still valid. Therefore, future studies could among others investigate how personal characteristics (age, sex, status, education, and experience of the group members), group characteristics (size), and group dynamics (the development of norms, conformity, polarization) influence the integrative discussion.

## Context

The assessment center rating process does not take place in a vacuum. The rating purpose and the organizational culture are among the factors, which could affect the rating process and the quality of assessor ratings.

With respect to rating purpose, assessors may evaluate candidates differently, depending on whether their ratings will serve a selection purpose (i.e., 'yes/no' decision) or a developmental purpose (i.e., identification of strengths and weaknesses). A related concept

is the processing objective (Lichtenstein & Srull, 1987). Assessors will process the incoming information differently if they are given an evaluative goal or an observational goal. To the best of our knowledge no studies in the assessment center field have experimentally manipulated these variables.

Another relevant contextual factor is the culture of the organization. Staufenbiel and Kleinmann (1999) tested the hypothesis that assessors do not judge assessees exclusively on the basis of the prescribed dimensions but also take into account the fit of the applicants into the culture of the organization. This study examined the so-called 'subtle criterion contamination' thesis (Klimoski & Brickner, 1987). This thesis posits that assessors' implicit constructs mimic the policy factors implicitly or explicitly defined by the organization. In their study Staufenbiel and Kleinmann gave student assessors information about the job and the dominant organizational leadership culture (competitive vs. cooperative). Afterwards, assessors watched four hypothetical candidates displaying either competitive or cooperative behaviors. Results predominantly showed that applicants demonstrating behavior in line with the organizational culture were rated more favorably.

Bartels and Doverspike (1997b) investigated whether differences in organizational level (i.e., upper and middle) and business stream (i.e., chemical, corporate, distributions, and research) moderated criterion-related validity. Assessment center performance validities did not increase when disaggregated according to either level or business stream.

## Rating Process

In this section we take a closer look at the rating process in assessment centers in terms of three divergent perspectives. In particular, the sparse research on the rating process in assessment centers is grouped along three conceptual models (Lord & Maher, 1990; Thornton, 1992): the rational model, the limited capacity model, and the expert model.

### Rational Model

A rational model of the rating process (Abelson, 1981; Bobrow & Norman, 1975; Borman, 1978; Rumelhart & Ortony, 1977) assumes people are able to attend to detailed behavior, to classify these many specific pieces of factual information into distinct categories, and to form relatively objective and accurate judgements. A rational model is also known as a data-driven, behavior-driven, or bottom-up model.

Most textbooks on assessment center practice (e.g., Ballantyne & Povah, 1995; Jansen & De Jongh, 1997; Woodruffe, 1993) adhere to this rational model. This model trains assessors to carefully proceed through the following rating phases. First, assessors observe verbal and nonverbal behavior of candidates. Most assessors observe ongoing behavior ('direct observation'), although in the US assessors also frequently observe videotaped performances of candidates ('indirect observation') (Bray & Byham, 1991). When observing assessors are expected to record clear behavioral descriptions instead of vague non-behavioral interpretations. After taking notes, assessors classify behaviors according to dimensions. This requires that assessors possess a thorough understanding of the dimensions and their definitions. Finally, assessors rate candidates on multiple job-related dimensions.

Thornton (1992) argues that these systematic and standardized practices lead to data-driven and accurate judgements. Several reasons underlie this argument. Firstly, in assessment centers the goal of accuracy (Neuberg, 1989) is stressed so that assessors are to devote time and energy to the distinct processes of observing, recording, and classifying behavior. Secondly, assessors are accountable for their ratings (Tetlock, 1983) as they have to justify their ratings to fellow assessors, to candidates, and to the organization. Thirdly, more careful and complex decision making occurs when people know that their ratings and decisions may have important implications for the future (e.g., career) of the person being judged (Freund, Kruglanski, & Shpitzajzen, 1985). To date virtually no studies have manipulated the effects of these conditions (e.g., goal of accuracy, etc.). An exception is the study of Mero and Motowidlo (1995), who demonstrated that accountability promoted rating accuracy in an assessment center related context.

Limited Capacity Model

This model posits that assessors possess limited information processing capacities and, therefore, are not always able to meet the cognitive demands of the assessment center process. (Reilly et al., 1990). One source of cognitive overload is that the behavioral information is presented to assessors at a very fast rate in the various exercises which last often over 30 minutes. Cognitive overload may also come from the many inferential leaps assessors must make in order to provide dimensional ratings. The determination of relevance, dimensionality, and relative weight of behaviors are among the inferences typically required of assessors. In particular, the assignment of individually observed behaviors to dimensions is an unstructured inference process where assessors judgmentally review their notes. Additionally, they have to formulate a numerical rating for each dimension by intuitively averaging and weighing the relevant behaviors, as the performance levels often remain undefined and implicit.

In the last decade this limited capacity model received considerable research attention as many studies tried to reduce the cognitive overload on the part of assessors. Examples included limiting the number of dimensions rated, using behavioral checklists, using video technology, or increasing the ratio of assessors to assessees. As discussed in previous sections, these studies were generally effective in reducing assessor cognitive overload as inferred by improvements in the quality of ratings.

Expert Model

The basic notion of this model is that professional assessors possess and use well-established cognitive structures when rating assessors. For expert assessors these organizing prior knowledge frameworks, which develop by abstracting from previous assessment center experiences and training, are helpful because they guide attention, categorization, integration, and recall processes (Cantor & Mischel, 1977; Fiske & Taylor, 1991; Srull & Wyer, 1980, 1989; Zedeck, 1986). Conversely, novice assessors (e.g.,

students) are not expected to possess such well-established cognitive structures when rating.

Several of the studies described above supported this expert model of the assessment center rating process. An example included the finding of higher discriminant validity for psychologist assessors than for managerial assessors (Sagie & Magnezy, 1997). Another example was that assessors receiving frame-of-reference training were better able to use the dimensions differentially (Lievens, 1999; Schleicher et al., 1999). In light of the notion of the expert model this was not unexpected because frame-of-reference training provided assessors with a mental framework regarding both the assignment of behaviors by dimension and the correct effectiveness level of each behavior (in line with the organization's norms and values). Accordingly, assessors were expected to place relevant incidents -as they occurred- in the appropriate mental category. Yet, use of prior knowledge frameworks might also exert additional effects. Schuler, Moser, and Funke (1994, see also Moser, Schuler, and Funke, 1999), for example, examined how assessor-assessee acquaintance influenced assessment center validities. When assessor-assessee acquaintance was less than or equal to two years, the criterion-related validity was .09. This value increased dramatically to .50 when assessor-assessee acquaintance was greater than two years.

## Results

In developmental assessment centers the results of the rating process primarily refer to the (final or within-exercise) ratings on the various dimensions. These dimensional ratings are expected to provide a detailed and valid portrayal of managerial strengths and weaknesses. Additionally, the results also refer to the developmental feedback, training activities, and action plans suggested to participants.

An examination of the quality of these results in developmental assessment centers should comprise of three criteria (Thornton et al., 1999, Carrick & Williams, 1998). A first criterion pertains to the quality of the dimensional ratings, namely these dimensional ratings

should be valid indicants of managerial abilities. This refers to the construct validity issue in assessment centers. If the dimensions are not valid indicants of the managerial abilities, the developmental feedback and action plans could be faulty or even detrimental (Fleenor, 1996; Joyce et al., 1994; Shore, Thornton, & Shore, 1990). The following example by Kudisch et al. (1997) succinctly highlights this. "Telling a candidate that he or she needs to improve his or her overall leadership skills may be inappropriate if the underlying construct being measured is dealing with a subordinate in a one-on-one situation (i.e., tapping individual leadership as opposed to group leadership)" (p. 131).

The second and third criterion refer to the developmental feedback and developmental activities suggested to participants. In fact, participants should accept the developmental feedback provided. The literature on performance feedback (Ashford, 1986) shows that this is not as straightforward as it may seem at first sight. In addition, participants should act upon the feedback. This may imply that participants follow developmental recommendations, further develop their skills, and apply these skills on the job. The remainder of this section reviews research with respect to these three criteria and the factors affecting them.

## Distinct Dimensional Assessment as Basis for Developmental Feedback

Internal Validation Strategy. To examine whether assessor ratings on the dimensions are valid indicants of the managerial abilities the majority of studies used the multitrait-multimethod matrix (Campbell & Fiske, 1959). In these studies the dimensional ratings which assessors make after completion of each exercise (i.e., within-exercise dimension ratings) were cast as a multitrait-multimethod matrix in which assessment center dimensions served as traits and assessment center exercises as methods.

The general conclusion from earlier research (e.g., Sackett & Dreher, 1982; see Jones, 1992; Kauffman et al., 1993; Klimoski & Brickner, 1987, for reviews) was that assessment center ratings did not measure the constructs they were purported to measure. Whereas assessor ratings on the same dimensions across exercises were found to correlate

lowly (i.e., low convergent validity), assessor ratings on different dimensions in a single exercise were found to correlate highly (i.e., low discriminant validity). Between 1990 and 1999 a first line of studies sought to examine the lack of convergent and discriminant validity of assessment centers in other settings. Generally, the troubling findings were replicated in British assessment centers (Crawley, Pinder, & Herriot, 1990; Henderson, Anderson, & Smith, 1995; McCredie & Shackleton, 1994), Australian assessment centers (Carless & Allwood, 1997), Dutch assessment centers (Van der Velde, Born, & Hofkes, 1994), Belgian assessment centers (Lievens & Van Keer, 1999), German assessment centers (Kleinmann & Koller, 1997), French assessment centers (Rolland, 1999), and Singaporean assessment centers (Chan, 1996). Three studies also examined the convergent and discriminant validity of assessor ratings in developmental assessment centers. The expectation was that the quality of construct measurement would improve in developmental assessment centers because they require a detailed assessment of participants' strengths and weaknesses. However, Joyce et al. (1994) and Fleenor (1996) found that the disappointing results were also generalizable to developmental assessment centers. Kudisch et al. (1997) revealed somewhat more construct-related evidence for developmental assessment centers. In this study both exercise factors and dimension factors provided the best representation of ratings in a developmental assessment center. Unfortunately, none of these studies experimentally manipulated assessment center purpose to examine the effect on the convergent and discriminant validity of the ratings.

Along these lines, a second stream of studies aimed to single out factors, which might improve the quality of construct measurement in assessment centers. Lievens (1998) reviewed 21 studies, which manipulated specific variables to determine their impact on assessment center convergent and discriminant validity. The rationale behind many of these design and procedural interventions was that they help assessors deal with their complex task. This review study showed that dimension factors (number, conceptual distinctiveness, and transparency), assessor factors (type of assessor and type of assessor training), and exercise factors (exercise form and use of role-players) were found to slightly improve

construct validity. Conversely, the studies regarding the impact of different observation, evaluation, and integration procedures yielded mixed results.

A third stream of studies used more powerful statistical techniques such as confirmatory factor analysis to examine construct validity (see Donahue et al., 1997; Harris et al., 1993; Kudisch et al., 1997; Schneider & Schmitt, 1992; Van der Velde et al., 1994). Confirmatory factor analysis explains the multitrait-multimethod matrix in terms of underlying constructs, rather than observed variables. In factor analytic terms the question is: Do the factors underlying the ratings represent dimensions or exercises? Factors defined by multiple measures of the same trait reflect construct validity of the measures, whereas factors based on different trait measures with the same instrument indicate method effects. Additionally, separate variance estimates of dimensions, exercises, and error are available. The general conclusion was that in most of the samples the 'Exercise –only' model produced a good fit of the data (Schneider & Schmitt, 1992; Van der Velde, et al., 1994), although adding one or more dimension factors to this model often resulted in an even better fit. A trend that deserved attention was the finding that the latter were often dimensions which could be observed more easily (e.g., oral communication). In some samples the model 'Exercises and Dimensions' provided the best representation of assessment center ratings (Donahue et al., 1997; Kudisch et al., 1997). However, loadings on exercise factors were generally higher than loadings on dimension factors. Recently, alternative ways of modeling multitrait-multimethod data have also been proposed. More specifically, because of estimation problems inherent in the traditional confirmatory factor analysis approach Sagie and Magnezy (1997), Kleinmann and Köller (1997), and Lievens and Van Keer (1999) modeled method (i.e., exercise) effects as correlated uniqueness (Marsh, 1989) instead of separate method factors. They showed that this procedure was less prone to ill-defined solutions and improper estimates. Kleinmann and Köller (1997) and and Lievens and Van Keer (1999) also found that the general confirmatory factor analysis approach slightly underestimated the proportion of dimension variance.

A crucial question is whether the construct validity findings represent assessor biases or true relationships. The former interpretation hinges on both the limited capacity and expert models described above. For instance, the lack of discriminant validity may be explained by the fact that assessors often fail to meet the heavy cognitive demands of the assessment center procedure, resulting among others in the inability to differentiate among the various dimensions. Otherwise, ecologically valid, schema-based processing on the part of assessors may also be responsible for the dimension overlap (Zedeck, 1986). According to the latter interpretation assessors are not to blame for the low convergent and discriminant validities found. Instead, these findings are simply due to candidates' real performance differences across situations (Neidig & Neidig, 1984). For example, certain individuals may perform better in one-to-one exercises than in group situations, diminishing the convergence of ratings across exercises. These performance differences have been labeled as true 'exercise effects'. Low discriminant validity may then result from the fact that some candidates exhibit no performance variation on the dimensions. Recently, two studies tried to disentangle these rival interpretations. Lance, Newbolt, Gatewood, and Smith (1995) reported on several studies in which they correlated latent exercise factors and external correlates. In general, hypothesized relationships between the exercise factors and the external correlates were found, supporting the explanation that the exercise factors capture true variance instead of error. Lievens (in press) showed that assessor ratings were relatively veridical. When assessors rated videotaped candidates whose performances varied across dimensions, assessors were reasonably able to differentiate among the various dimensions. When assessors rated a videotaped candidate without clear performance fluctuations across dimensions, distinctions about dimensions were more blurred. Clearly, these two studies demonstrate that the troubling construct validity findings might reflect more true variance than previously thought and therefore shed a more positive light on assessment center construct validity.

External Validation Strategy. To examine whether developmental assessment centers yield distinct trait assessments some studies have used external criteria. These studies have

linked final dimension ratings in a nomological net (Cronbach & Meehl, 1955) with personality questionnaires and cognitive ability measures.

Using this nomological network approach Shore et al. (1990) hypothesized that final ratings on the dimensions were construct valid if the correlations between dimension scores and scores on conceptually related measures were higher than correlations between dimension scores and scores on conceptually unrelated measures. Per assessment center dimension, they classified psychological measures (e.g., measures of personality and cognitive ability) as either conceptually related or unrelated to that dimension. Conforming to their hypotheses, cognitive ability measures related more strongly to the performance-like dimensions (i.e., candidates' proficiency in performing their tasks) than to the interpersonal-style dimensions (i.e., candidates' style of behavior toward other people in work situations). Furthermore, convergent validity was found for all three interpersonal-style dimensions, and for three of six performance-like dimensions. Discriminant validity was established for two of the interpersonal-style dimensions, and for one of the performance-style dimensions. Recently, these results were confirmed by one study (Thornton, Tziner, Dahan, Clevenger, & Meir, 1997) but disconfirmed by two other studies (Chan, 1996; Fleenor, 1996). In these latter two studies the final dimension ratings failed to demonstrate most of the expected relationships with conceptually similar personality dimensions. Furthermore, the average correlations between final dimension ratings and conceptually dissimilar personality dimensions were equal or even higher than with conceptually related personality dimensions.

Scholz and Schuler (1993) conducted a meta-analysis (N = 22106) of studies in which assessment center scores (e.g., overall assessment rating, dimensional scores, etc.) were correlated with an array of external measures such as cognitive ability measures or personality inventories. Their meta-analysis included 51 studies and 66 independent samples. Intelligence correlated .33 with the overall assessment rating, which increased to .43 when corrected for unreliability. Besides intelligence, the overall assessment center rating tended also to correlate .23 (corrected for unreliability) with dominance, .30 with achievement motivation, .31 with social competence, and .26 with self-confidence.

Examining the utility and validity of selection devices generally, Schmidt and Hunter (1998) summarized 85 years of research findings. They reported that assessment center ratings did have a corrected correlation with external criteria of job success of .37. However, consistent with Scholz and Schuler (1993), they also pointed out the high correlation of assessment centers with general mental ability, which they estimated to be around .50. Because of this, when combined with a measure of general mental ability as part of a predictor battery, Schmidt and Hunter would expect an assessment center to account for very little additional variance, hence calling into question its utility. Recently, Fleenor (1996) found that the personality trait 'exhibition' was significantly correlated with all 10 assessment center dimensions, the trait 'aggression' with seven and the trait 'dominance' with five dimensions. Apparently, participants who were 'good actors' and highly competitive were rated significantly higher in the assessment center. Moser, Diemand, and Schuler (1996) correlated ratings of 58 candidates on a self-monitoring questionnaire to their ratings in an assessment center, which was designed to provide recommendations for promotion to supervisory positions. No relationship ($r$ = .02) was found between high scores on the inconsistency scale of the self-monitoring questionnaire and higher assessment center ratings (for similar results, see Arthur & Tubre, 1999). However, the social skills scale showed significant correlations ($r$ = 26) with assessee ratings. Furnham, Crump, and Whelan (1997) validated the NEO Personality Inventory using assessor ratings. A clear pattern emerged with conscientiousness and extraversion having strongest and most frequent correlations with assessor ratings. Other research does not lend support to the link between assessment centers and personality. Goffin, Rothstein, and Johnston (1996) reported a marked lack of correlation between personality and assessment center scores because both personality and dimensional assessment center scores had significant incremental validity over one another. Goffin et al. (1996) concluded that "personality and assessment centers sample different domains which in turn predict relatively different aspects of job performance" (p. 753).

A limitation of the majority of the aforementioned studies is that they did not relate ratings of developmental assessment centers to external criteria. Probably this explains why these studies used personality and cognitive ability as external criteria of the final dimension ratings measured. However, in assessment centers conducted for developmental purposes other constructs might serve as more relevant criteria. Examples include motivation-based constructs (Jones, 1997), extra-role performance, or general occupational interests.

Reactions and Acceptance of Developmental Feedback

As noted above, dimensional ratings serve as basis for the developmental feedback provided to participants in most applications of the assessment center method, but they are the 'raison d'être' for the developmental assessment center. The quality of these assessor descriptions provided at the end of developmental assessment centers might be examined by looking at participants' acceptance and reactions of the feedback. If participants do not understand the feedback or do not accept it, it is unlikely that they will react positively and initiate in developmental activities (Thornton et al., 1999). Positive reactions are often found but these appear to be linked to the job-relatedness and face validity of the assessment center exercises (Iles & Mabey, 1993; Kluger & Rothstein, 1993; Kravitz, Stinson, & Chavez, 1996; Macan, Avedon, Paese, & Smith, 1994; Rynes & Connerly, 1993; Sichler, 1991; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Relatively few studies have addressed how participants react to the developmental feedback, and in particular, the role that the quality of the ratings plays.

In one noteworthy exception, a comprehensive study by Harris, Paese, and Greising (1999) used organizational justice theory as a framework to investigate which variables were related to feedback reactions in a developmental assessment center. Participants' feedback reactions were measured by three criteria: procedural fairness, distributive fairness, and perceived utility of the feedback. Results showed that variables related to assessment center exercises (perceived content validity, perceived feedback validity, and affect) with the exception of fakability were generally related to all three measures of participants' feedback

reactions. Not unexpectedly, participant reactions were also predicted by feedback process variables (i.e., participation, specificity of feedback, and personableness of the assessor). Met expectations, operationalized as the degree of difference between the expected rating and the actual rating, was related to both procedural and distributive fairness, but not to perceived utility of feedback. These results mesh well with studies by Burd and Ryan (1993) and Kudisch and Ladd (1997). They showed that acceptance of developmental feedback was related among others to exercise realism, feedback favorability, and perceived assessor expertise. Other studies (Baisden & Robertson, 1993; Kudisch & Ladd, 1997) investigated whether specific personality characteristics of participants predicted feedback acceptance. However, no clear pattern emerged.

Besides the factors affecting feedback acceptance and reactions, another issue is which feedback type participants prefer. Thornton et al. (1999) distinguished between attribute feedback (i.e., organized around the dimensions) and exercise feedback (i.e., organized around the simulation exercises). Results indicated favorable reactions to both feedback types and no real differences in the extent to which participants perceived the attribute-based feedback or exercise-based feedback as accurate and useful.

Developmental Actions as a Result of the Feedback

A third and last criterion for examining the quality of assessor decisions in developmental assessment centers consists of looking whether participants actually acted upon the developmental feedback and engaged in subsequent developmental activities. Research results are mixed. Engelbrecht and Fisher (1995) discovered that 41 managers who received feedback after an assessment center experience and who engaged in subsequent developmental activities were rated higher on six performance dimensions than a comparable group of 35 managers who had not gone through the assessment process. The effects of this developmental assessment center were still measurable three months later. Unfortunately, it was unlikely that in this study managers were randomly assigned to ‘conditions’. Hence, it may be that those, who went though the center, differed in their

orientation to self-development to begin with. Other studies demonstrated the limited effectiveness of developmental assessment centers. For instance, Jones and Whitmore (1995) pointed out the lack of differences in career advancement between managers who went through a developmental assessment center and a naturally occurring control sample. Acceptance of developmental feedback was also not related to promotion and following recommended developmental activities was related to eventual promotion for only two of seven performance dimensions (i.e., career motivation and working with others). Mitchell and Maurer (1998) built on these disappointing findings and tried to explain which factors were related to participation in subsequent training and developmental activities. They showed that individuals who received lower ratings engaged in higher amounts of subsequent training. Perceived time constraints interfered with learning and developmental activities. Social support for development and managers' self-efficacy for development were related to on-the-job development constructs. Other perceived context factors and individual differences did not moderate the relationship between feedback and training/developmental activities.

## Conclusions and Discussion

### What have we learned?

From our review, it seems clear that the 'quality' of assessment center decisions (i.e., dimensional/trait ratings) can be measured and indexed. And when quality has been measured, it has been found to vary considerably - some centers have it, others do not. Moreover, the quality of the output of assessment centers appears to be linked to major assessment center 'design' parameters. The most profound insights from our review, however, are not solely associated with 'design' features of the assessment centers. Our review also has convinced us that we must have a deeper understanding of the nature of the assessor as social information processor.

Assessment center design issues. Most notably, the nature and number of the dimensions seem to affect the quality of judgements made by center staff. In general, having

to rate fewer conceptually independent dimensions, which can be clearly operationalized (and which have a real opportunity to reveal themselves in the exercises), results in higher quality. It also seems to help, if there is reasonable variability in the trait of interest and variability in the population of participants to be assessed, as relative judgements are always easier to make.

It has also been found that the nature of the exercises exerts considerable impact on judgement quality. Aspects of form, content, and the instructions given to participants make it easier to infer the existence of the traits being assessed. In similar vein, thoroughly trained role-players appear to help assessors observe relevant behaviors in exercises. It also seems likely that the order in which assessors see participants relative to exercises and the assignments given to assessors (e.g., the assignment to specialize on a particular dimension across exercises and participants) have major consequences for the ability of assessors to estimate a participant's strengths and weaknesses.

To put it simply, unless the exercises provide an opportunity to observe enough behaviors and to do so under (assessor) favorable conditions, it is very difficult to infer traits or dispositions. In this regard, most exercises appear to have been selected or designed more for their face (content) validity, than for their capacity to expose behavior that would reveal the level of specific traits possessed by the participant.

On a related matter, we might put forth as a thesis that the emphasis on exercises reflecting job content has another unintended effect. It would seem to highlight the capacity of individuals to perform well on job relevant tasks. On the 'plus' side, as pointed out by Klimoski and Brickner (1987), this may help to account for the criterion-related validity of assessment centers. Simply re-stated, assuming the content validity of the exercises, assessors are focused on estimating (predicting) likely future job performance of the candidates. However, on the 'minus' side, this may actually interfere with their major task, which is (arguably) the estimation of scores on traits or dimensions. Given the aforementioned difficulty of trait estimation from behavior elicited by exercises we feel that this negative influence is quite likely. It is also problematic because performance is usually

an imperfect indicator of key traits. In particular, the ecological validity of dimensions or traits vis-a-vis performance is rarely even considered in center design. Moreover, as we and others (e.g., Joyce et al., 1994) have noted, the particular mix of dimensions to be estimated usually varies by exercise and performance on various exercises will be driven by different combinations of traits as well.

Social judgment design issues. Based on the performance appraisal literature (Murphy & Cleveland, 1995), when poor quality ratings are encountered, it is reasonable to examine at least three factors. The first is the opportunity to observe. As described above, we do think that this is part of the story. The time with a given candidate and/or the circumstances surrounding the observations (e.g., the exercises and their opportunity to elicit dimension-related behaviors) do seem important.

A second relates to motivation, in this case, the motivation to provide quality judgements. Whereas it is possible that assessment center staff are not motivated, this may only be true in a nuanced sense. In the first place, assessors are typically a select and dedicated group. Most volunteer for the assignment. More telling, the structure of the typical assessment center would seem to emphasize accountability. For instance, as we reported, most centers make use of an integration session wherein assessors have to offer and justify their point of view regarding each participant's scores. Similarly, assessors frequently have to provide (face to face) feedback to participants. Such conditions of accountability are known to produce motivated behavior (Frink & Klimoski, 1998; Lerner & Tetlock, 1998; Mero & Motowidlo, 1995). The existence of these and other realities (e.g., most centers are 'high visibility' operations) lead us to believe that poor quality ratings are not caused by simple lack of effort. But, as will be detailed below, it just may be that effort is still part of the story if it is being allocated toward the wrong goals (e.g., estimating potential vs. dimensional accuracy)

If we have addressed the role of observation and motivation, what remains? In our opinion, the key may reside in a better understanding of a third factor, namely the capability of the assessor to make quality judgements. In the performance appraisal literature, capability is usually thought of as an amalgam of skill and capacity. In fact, we feel that both

skill and capacity are implicated in the issue of the quality of assessment center judgements. Quite clearly, most operational centers place a high premium on the training given to assessment center staff members. This is laudatory, as training is a very direct way to increase skill. However, notwithstanding some recent research examples (Lievens, 1999; Schleicher et al., 1999), training programs built around models of social information processing are still lacking in operational centers. Indeed, it is our position that assessment centers, their design (including assessor training), and their administration would profit from a better integration of current thinking in person perception, social information processing, interpersonal judgements, and decision making. But even this said, current models of social cognition, even once identified as useful, would still need to be translated into implications for assessment centers.

What do we need to know?

The framework that we used as a heuristic for this review was derived from the performance rating literature. To its credit, this included a consideration of the rating process. Moreover, we pointed out that when it comes to current thinking about the performance rating process, the field has moved towards the so-called 'expert' model perspective, implicating such phenomena as cognitive structures, decisional heuristics, case-based reasoning, and the notion of cognitive resources. In characterizing the work on assessment centers in the last decade, it should seem clear that a substantial portion of the problem of dimensional assessment accuracy may indeed be better understood in terms of what we have learned about performance-related information processing. But it is not sufficient. Consistent with some of the suggestions of Murphy and Cleveland (1995), assessment center research should be guided by a realization that we are not just trying to model 'information' processing. In fact, in trying to unravel the puzzle of assessment center rating quality, we are essentially dealing with 'social' information, gathered in social or interpersonal settings. As such, findings from the literature on social cognition and social perception must

be integrated into our thinking, into our research, and ultimately into our design solutions for assessment centers.

Although we could never do justice to the extensive social cognition domain in this chapter, we will try to highlight certain concepts and theories that have been found useful in characterizing the way people process social information. Much of the material below derives from Fiske's (1993) very useful and contemporary summary. In summarizing this material, we will, in effect, be implying a more sophisticated framework than the one at the start. Accordingly, we also hope that we will offer some guidance regarding the research needed to establish what could be thought of as 'contextualized' models of social information processing for the assessment center venue. Finally, we hope to integrate into our treatment some of the findings touched upon earlier in a manner designed to illustrate the potential of this more 'social' perspective. In this regard, the following represents our nominations of 'best bets' for future research.

Social judgement accuracy. The literature confirms that we have a propensity for and some skill in perceiving and judging others. In fact, it has been argued that we are generally pretty good at it. For example, we are quite accurate at judging dominance and warmth, with minimal opportunity to observe and to interact with someone. In fact, some might argue that the traits represented in the so-called Five Factor Model actually reflect the way that we generally perceive and describe people.

Given this, why do we not get higher quality judgements in the assessment center? One possibility is that the qualities used in the general case differ from those typically sought in assessment centers. Given the existing and natural tendencies to perceive and process people in a certain way, it may be that the assessment center dimensions represent some kind of an 'over-lay' task, that frequently comes into conflict with these tendencies. Whereas prior attempts to use other types of dimensions in assessment centers were generally unsuccessful (Joyce et al., 1994; Russell & Domm, 1995), it may be worthwhile examining the implications of selecting and using a set of dimensions with special regard to the generalized tendencies of people for trait accuracy.

The role of expectancies. Our expectancies regarding someone strongly affect our perceptions and cognitions. More to the point, we attend to and process expectancy congruent and expectancy-incongruent information differently. To date, we know little about the effects of assessment center expectancies for levels of trait information that staff think they will encounter as a function of a specific simulation (see Highhouse & Harris, 1993, for an exception) or as a function of a specific individual being observed. As noted earlier, we do know that access to prior information about a candidate counts (Moser et al., 1999; Schuler et al., 1994) but we do not know why and how these affect the judgement process.

The role of cognitive structures. Beliefs about traits and trait structures influence how interpersonal information is assembled and used but we know little of how typical trait structures link to behaviors and performance. For example, Reeder, Prior, and Wojciszke (cited in Fiske, 1993) distinguish among frequency based traits (talkativeness), morality traits (honesty), capacity traits (ability), and those that implicate attitudes or values (work ethic) and point out the problematic inference from and to behavior for each of these types. In light of the wide range of trait types used in operational centers (Howard, 1997), this aspect deserves to be studied. As already described, Tett's (1998, 1999) trait activation model may be useful here.

Similarly, stereotypes and prototypes (exemplars) are other structures that appear to affect attention, expectancies, and cognition in assessment centers. Particularly relevant is the potential role of cognitive structures called scripts (standard narrative structures and plausible causal sequences). In this regard, it is quite likely that behavioral conformance on the part of candidates to script-like structures and especially deviations from scripts play an important part in the inference process of assessors. Here we know very little. We know even less about the manner in which assessors match their observations to the exemplars (or scripts) that they hold.

Finally, regarding structures for meaning making, there is a great deal of evidence that social perceivers often use narrative reasoning. Here, in trying to make sense of social information, we construct brief 'stories' for ourselves in order to deal with inconsistent social

information or to account for unexpected/atypical behavior on the part of someone. Given the demands of staff to communicate their impressions to one another, its quite likely that, over time, they too would develop useful prototype narratives to 'explain' anomalies or inconsistencies in the performance of candidates for whom they already developed an impression (e.g., he/she was tired, was in a high-performing group, etc.). In this regard, we really have very little information about how, when, and why staff make causal attributions.

In sum, we believe that cognitive structures are implicated in such things as attention, person perception, information processing, memory, and rating. Accordingly, they are an important mechanism to understand the conditions for rating quality.

Controlled vs. automatic processes. The literature on human cognition has highlighted that we operate at different awareness levels when it comes to information processing. Sometimes we are rather oblivious as to what stimuli we are attending to and how we are processing them. In other instances, we are most deliberate in our approach to attention, perception, and thinking. This is especially likely to be true in interpersonal relations and in the processing of social information. Generally speaking, automaticity implies cognitive efficiency. Hence, it is often the 'default' or natural approach to a complex and demanding world.

Current thinking has elaborated upon this dichotomy and offers a continuum with several noteworthy stages (Fiske, 1993). For instance, pre-conscious automaticity occurs without much awareness at all. We are not consciously attending to stimuli or to our processing of the stimulus. We do also not start or stop such processing. Post-conscious automaticity implies that we are cognizant of the stimulus but not of its effects on us. Research on the dynamics of priming shows that aspects of an ambiguous stimulus (e.g., behavior of a participant in an assessment center) can activate structures in memory. Just which structures (e.g., trait associations) are activated appears to depend on their accessibility. Accessibility, in turn, can be a function of activation frequency or recency of use (or both). There is also evidence that accessibility of structures of information is related to salience. That is, because certain actions or features of an assessee often stand out (e.g.,

different gender, more talkative, extreme performance), they are likely to trigger structures and thereby affect inferences.

Goal-dependent automaticity is triggered by motivated effort. We are aware of the stimulus, but not necessarily of all of its effects on processing. In this regard, inferring traits appears to occur rather spontaneously (and effortlessly). Further, there is some evidence that we tend to infer dispositions very fast. So is our tendency to make categorical judgements regarding other people, based on stereotypes. When we form impressions, we realize that this is occurring, but we are usually not conscious of just what cues and in what combinations are having their effects. While this is fairly automatic, it can be controlled. This is often done via assessor training. Despite a tradition for careful training of assessors with the goal of turning them into experts, the record reveals that such traditional training still results in a relatively poor capacity to make valid dimensional judgements. One idea that we have already shared in this regard is that traditional training may have inadvertently confounded the notion of skill at assessing performance with that of assessing traits. One solution would be to design developmental centers differently (e.g., having the exercises elicit trait-revealing behaviors better). However, it may be that training needs to be different as well so that assessors master the distinction between valid performance structures and valid trait structures and the appropriate use for inferring the latter form the former.

Note also that some automatic processes start as controlled processes. This is what happens as a person develops skill and proficiency through practice and experience. But here the speed up is with regard to our processing of information generally and is not target specific (e.g., as when a candidate seems to match a prototype completely). In an ideal scenario, social judgements and social categorizations are both fast and valid. This is the hallmark of the expert.

At the other end of the continuum is what Fiske (1993) refers to as fully intentional thinking. Here we are aware of our attempts at the deliberate control of attentional processes and are rather self-conscious about the way we go about processing what we see, what we think, and what we do. When thinking intentionally, we might also deliberately invoke the use

of certain meta-cognitive strategies (plans, feedback seeking, etc.). In most settings where we are trying to learn a new skill or procedure this is probably what is occurring. Conversely, in instances where we are drawn into automatic processing, but where it may not be advisable, the challenge might be to find ways enforce more deliberate cognition. Because of forces promoting automaticity (e.g., self-confidence, routineness) there are continuing risks for errors. Thus, in practice, we sometimes see the use of procedural checklists (e.g., as in a pre-flight inspection of an aircraft). Encouraging the use of checklists by assessors as described earlier in this chapter would seem to fall into this notion as well. Apart from its training value and capacity to shape valid cognitive structures, such checklists serve to raise the observation/rating process to a conscious level of awareness.

This more elaborate notion of automaticity can be used to guide assessment center research, particularly regarding the features of center design and administration that might promote or retard the formation and use of appropriate structures and the valid (and speedy) processing of social information. Specifically, we would recommend that research focuses on the motivational forces that might produce or reduce diligent processing.

Motivated cognition. Current research on social perception has highlighted that motivation is important. It is not just the level of motivation that is relevant, it is the goal that is behind it that counts. Moreover, the motivation and the underlying goals seem to have a profound effect on the strategies used in the service of social perception, impression formation, and judgements. In the end it may be that a better understanding of these strategies holds the clues for valid social information processing and trait inference in the assessment center. Fiske (1993) points out that there are two primary motives operating in social/interpersonal settings. One is the desire for accuracy and open-mindedness in the service of making valid assessments of others. The other is tied to seeking closure. This is basically an action orientation. For instance, a decision must be made or a result turned in. Generally, the motive for accuracy would imply the withholding of judgements or the willingness to revise judgements. In contrast, when action must be taken, we tend to adopt a confirmatory strategy.

Our thesis is that, despite policies to the contrary, in operational settings there appear to be forces that promote confirmation. In our analysis, the 'costs' for being wrong may frequently outweigh the 'costs' for being 'indecisive'. The literature on social cognition implicates such factors as complex and inconsistent stimuli, time pressure, an obligation to report (simple stories) to others, emotional arousal on the part of the perceiver, and the existence of well established cognitive structures (performance prototypes). We suggest that some of these be considered in future research.

Social interaction as the basis for inference. As noted earlier, assessment centers make use of a variety of measurement techniques. However, a distinctive feature is the use of simulations that involve the interaction of groups of individuals. As noted, the observations from such interactions frequently carry the weight of inference.

Social judgement theory and social cognition models recognize that making (trait) inferences from limited observations of social interaction is not easy but they also offer perspectives and models for doing so. An example is the work on attribution theory (e.g., Ross & Nesbitt, 1991). In particular, it should be possible to translate descriptive and predictive studies of how observers make dispositional (trait) attributions into prescriptions for center design. For example, it may well be that a better basis for inference is to be able to observe the same target individual in both the same group and a different group over time.

Social interaction as accountability. We have noted that accountability theory would have some use in characterizing the forces that are operating on assessors relative to both the amount and the direction of their motivation. More must be done to carefully analyze the implications of center design, policy, and practice on such forces. In particular, very little work has been done on the normative structure of assessor teams and on the dynamics of the integration session (see Klimoski, Friedman, & Weldon, 1980, for an exception). Similarly, more must be done to understand the motivational properties of different arrangements of the feedback given. Clearly, facing the prospect of a meeting with a participant personally afterwards to provide feedback should have different consequences for the cognitive and affective process controlling the quality of assessor ratings.

The portrayal of the complexities of judgement and human information processing in a diagram or figure is always difficult. Thus, it is with some trepitude that we offer Figure 2. In this figure the rectangles may be considered as antecedents of the individual and collective processes (see the two circles in Figure 2) in assessment centers. The ovals are then the dependent variables of interest. We feel that Figure 2 is one attempt to represent some of the key factors discussed in this chapter. Moreover, it is our way of 'translating' what we have learned from the social cognition literature into the assessment center context. As such, it should be viewed as a heuristic for guiding future research on the quality of assessor judgements. Using it would not only build on contemporary models of social judgement, but would also have the value of better informing practice relative to the most appropriate choice of center design features to be used in the future.

---

Insert Figure 2 about here

---

## Epilogue

The last decade of research on assessment centers has been informative but it could have been more so if investigators were more frequently working from a general plan of attack. In our opinion the next decade of research on the assessment center method would be far more informative if it were to be guided by findings and models from the social cognition literature. It would serve to promote more systematic and programmatic efforts. It would also increase the likelihood that we will get closer to solving the 'puzzle' of assessment centers (Klimoski and Brickner, 1987).

References

Abelson, R.P. (1981). 'Psychological status of the script concept', American Psychologist, 36, 715-729.

Adams, K.A., and Osburn, H.G. (1998, April). 'Reexamination of the exercise effect in assessment center ratings', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Ahmed, Y., Payne, T., and Whiddett, S. (1997). 'A process for assessment exercise design: a model of best practice', International Journal of Selection and Assessment, 5, 62-68.

Anderson, N., Payne, T., Ferguson, E., and Smith, T. (1994). 'Assessor decision making, information processing and assessor decision strategies in a British assessment centre', Personnel Review, 23, 52-62.

Andres, J., and Kleinmann, M. (1993). 'Development of a rotation system for assessors' observations in the assessment center [In German]', Zeitschrift für Arbeits- und Organisationspsychologie, 37, 19-25.

Arthur, W.E., and Tubre, T.C. (1999, May). 'The assessment center construct-related validity paradox: A case of construct misspecification?', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Ashford, S. (1986). 'Feedback-seeking in individual adaptation: A resource perspective', Academy of Management Journal, 29, 465-487.

Baisden, H.E., and Robertson, L. (1993, March). 'Predicting receptivity to feedback in a developmental assessment center', Paper presented at the International Congress on the Assessment Center Method, Atlanta, GA.

Ballantyne, I., and Povah, N. (1995). 'Assessment and development centres', Aldershot: Gower.

Baron, H., and Janman, K. (1996). 'Fairness in the assessment centre', International Review of Industrial and Organizational Psychology, 11, 61-114.

Bartels, L.K., and Doverspike, D. (1997a). 'Assessing the assessor, the relationship of assessor personality to leniency in assessment center ratings', Journal of Social Behavior and Personality, 12, 179-190.

Bartels, L. K., and Doverspike, D. (1997b). 'Effects of disaggregation on managerial assessment center validity', Journal of Business and Psychology, 12, 45-53.

Binning, J.F., Adorno, A.J., and Kroeck, K.G. (1997, April). 'Validity of behavior checklist and assessor judgmental ratings in an operational assessment center', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Binning, J.F., Adorno, A.J., and Williams, K.B. (1995, May). 'Gender and race effects on behavior checklist and judgmental assessment center evaluations', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Bobrow, D.G., and Norman, D.A. (1975). 'Some principles of memory schemata', In D.G. Bobrow and A.G. Collins (Eds.), Representation and understanding: Studies in cognitive science (pp. 131-150). New York: Academic Press.

Bobrow, W., and Leonards, J.S. (1997). 'Development and validation of an assessment center during organizational change', Journal of Social Behavior and Personality, 12, 217-236.

Borman, W.C. (1978). 'Exploring the upper limits of reliability and validity in job performance ratings', Journal of Applied Psychology, 63, 135-144.

Boyle, S., Fullerton, J., and Wood, R. (1995). 'Do assessment/development centres use optimum evaluation procedures? A survey of practice in UK organizations', International Journal of Selection and Assessment, 3, 132-140.

Brannick, M.T., Michaels, C.E., and Baker, D.P. (1989). 'Construct validity of in-basket scores', Journal of Applied Psychology, 74, 957-963.

Bray, D.W., and Byham, W.C. (1991). 'Assessment centers and their derivates', The Journal of Continuing Higher Education, 39, 8-11.

Brostoff, M., and Meyer, H.H. (1984). 'The effects of coaching on in-basket performance', Journal of Assessment Center Technology, 7, 17-21.

Burd, K.A., and Ryan, A.M. (1993, May). 'Reactions to developmental feedback in an assessment center', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Campbell, D.T., and Fiske, D.W. (1959). 'Convergent and discriminant validation by the multitrait-multimethod matrix', Psychological Bulletin, 56, 81-105.

Campbell, W.J. (1991). 'Comparison of the efficacy of general and specific performance dimensions in an operational assessment center', Unpublished dissertation, Old Dominion University, VA.

Cantor, N., and Mischel, W. (1977). 'Traits as prototypes: Effects on recognition memory', Journal of Personality and Social Psychology, 35, 38-48.

Carless, S.A., and Allwood, V.E. (1997). 'Managerial assessment centres: What is being rated?', Australian Psychologist, 32, 101-105.

Carrick, P., and Williams, R. (1998). 'Development centres: A review of assumptions', Human Resource Management Journal, 9, 77-92.

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

Clapham, M.M. (1998). 'A comparison of assessor and self dimension ratings in an advanced management assessment center', Journal of Occupational and Organizational Psychology, 71, 193-203.

Clapham, M.M., and Fulford, M.D. (1997). 'Age bias in assessment center ratings', Journal of Managerial Issues, 9, 373-387.

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

Crawley, B., Pinder, R., and Herriot, P. (1990). 'Assessment centre dimensions, personality and aptitudes', Journal of Occupational Psychology, 63, 211-216.

Cronbach, L.J., and Meehl, P.E. (1955). 'Construct validity in psychological tests', Psychological Bulletin, 62, 281-302.

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

Dulewicz, V., and Fletcher, C. (1982). 'The relationship between previous experience, intelligence and background characteristics of participants and their performance in an assessment centre', Journal of Occupational Psychology, 55, 197-207.

Engelbrecht, A.S., and Fischer, A.H. (1995). 'The managerial performance implications of a developmental assessment center process', Human Relations, 48, 387-404.

Epstein, S. (1979). 'The stability of behavior: I. On predicting most of the people much of the time', Journal of Personality and Social Psychology, 37, 1097-1126.

Fiske, S.T. (1993). 'Social cognition and social perception', Annual Review of Psychology, 44, 155-194.

Fiske, S.T., and Taylor, S.E. (1991). 'Social cognition', Singapore: McGraw-Hill.

Fleenor, J.W. (1996). 'Constructs and developmental assessment centers: Further troubling empirical findings', Journal of Business and Psychology, 10, 319-333.

Fletcher, C., and Kerslake, C. (1993). 'Candidate anxiety and assessment centre performance', Journal of Managerial Psychology, 8, 19-23.

Fletcher, C., Lovatt C., and Baldry,C. (1997). 'A study of state, trait, and test anxiety, and their relationship to assessment center performance', Journal of Social Behavior and Personality, 12, 205-214.

Freund, T., Kruglanski, A.W., and Shpitzajzen, A. (1985). 'The freezing and unfreezing of impressional primacy: Effects of the need for structure and the fear of invalidity', Personality and Social Psychology Bulletin, 11, 479-487.

Frink, D., and Klimoski, R.J. (1998). 'Toward a theory of accountability in organizations and human resource management', In G.R. Ferris (Ed.), Research in personnel and human resource management, vol. 16, Greenwich, CT. JAI press.

Fritzche, B.A., Brannick, M.T., and Hazucha, J.F. (1994, April). 'The effects of using behavioral checklists on the predictive and construct validity of assessment center

ratings', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

Furnham, A., Crump, J., Whelan, J. (1997). 'Validating the NEO Personality Inventory using assessor's ratings', Personality and Individual Differences, 22, 669-675.

Gaugler, B.B., and Rudolph, A.S. (1992). 'The influence of assessee performance variation on assessors' judgments', Personnel Psychology, 45, 77-98.

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

Gill, R.W.T. (1982). 'A trainability concept for management potential and an empirical study of its relationship with intelligence for two managerial skills', Journal of Occupational Psychology, 52, 185-197.

Goffin, R.D., Rothstein, M.G., and Johnston, N.G. (1996). 'Personality testing and the assessment center: incremental validity for managerial selection', Journal of Applied Psychology, 81, 746-756.

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

Guldin, A., and Schuler, H. (1997). 'Consistency and specificity of assessment center criteria: A new approach for construct validation of assessment centers [In German]', Diagnostica, 73, 230-254.

Halpert, J.A., Wilson M.L., and Hickman, J.L. (1993). 'Pregnancy as a source of bias in performance appraisals', Journal of Organizational Behavior, 14, 649-663.

Harris, M.M., Becker, A.S., and Smith, D.E. (1993). 'Does the assessment center scoring method affect the cross-situational consistency of ratings?', Journal of Applied Psychology, 78, 675-678.

Harris, M.M., Paese, M., and Greising, L. (1999, August). 'Participant reactions to feedback from a developmental assessment center: an organizational justice theory approach', Paper presented at the Academy of Management Meeting, Chicago, IL.

Hauenstein, P.C. (1994, April). 'A key behavior approach for improving the utility of developmental assessment centers', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

Henderson, F., Anderson, N., and Rick, S. (1995). 'Future competency profiling: Validating and redesigning the ICL graduate assessment centre', Personnel Review, 24, 19-32.

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

Highhouse, S., and 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.

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

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

Iles, P.A., and Mabey, C. (1993). 'Managerial career development techniques: effectiveness, acceptability and availability', British Journal of Management, 4, 103-118.

Jansen, P.G.W., and Jongh, de F. (1997). 'Assessment centers: a practical handbook', Chicester: John Wiley.

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

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

Jones, R.G., and Whitmore, M.D. (1995). 'Evaluating developmental assessment centers as interventions', Personnel Psychology, 48, 377-388.

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

Kauffman, J.R., Jex, S.M., Love, K.G., and Libkuman, T.M. (1993). 'The construct validity of assessment centre performance dimensions', International Journal of Selection and Assessment, 1, 213-223.

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. (1997). 'Transparency of the required dimensions: A moderator of assessment centers' construct and criterion validity [In German]', Zeitschrift für Arbeits und Organisationspsychologie, 41, 171-181.

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

Kleinmann, M., Exler, C., Kuptsch, C., and Köller, O. (1995). 'Independence and observability of dimensions as moderators of construct validity in the assessment center [In German]', Zeitschrift für Arbeits- und Organisationspsychologie, 39, 22-28.

Kleinmann, M., Kuptsch, C., and Köller, O. (1996). 'Transparency: A necessary requirement for the construct validity of assessment centres', Applied Psychology: An international Review, 45, 67-84.

Kleinmann, M., Andres, J., Fedtke, C., Godbersen, F., and Köller, O. (1994). 'The influence of different rating procedures on the construct validity of assessment center methods [In German]', Zeitschrift für Experimentelle und Angewandte Psychologie, 41, 184-210.

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

Klimoski, R.J., Friedman, B.A., and Weldon, E. (1980). 'Leader influence in the assessment of performance', Personnel Psychology, 33, 389-401.

Kluger, A.N., and Rothstein, H.R. (1993). 'The influence of selection test type on applicant reactions to employment testing', Journal of Business and Psychology, 8, 3-25.

Kolk, N.J., Born, M., Bleichrodt, N., and Flier, H., van der (1998, August). 'A triarchic approach to assessment center dimensions: Empirical evidence for the Feeling-Thinking- Power model for AC dimensions', Paper presented at the International Congress of Applied Psychology, San Francisco, CA.

Kravitz, D.A., Stinson, V., and Chavez, T.L. (1996). 'Evaluations of tests used for making selection and promotion decisions', International Journal of Selection and Assessment, 4, 24-34.

Kriek, H.J., Hurst, D.N., and Charoux, J.A.E. (1994). 'The assessment centre: Testing the fairness hypotheses', Journal of Industrial Psychology, 20, 21-25.

Kudisch, J.D., and Ladd, R.T. (1997, April). 'Factors related to participants' acceptance of developmental assessment center feedback', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Kudisch, J.D., Ladd, R.T., and 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.

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

Kurecka, P.M., Austin, J.M., Johnson, W., and Mendoza, J.L. (1982). 'Full and errant coaching effects on assigned role leaderless group discussion performance', Personnel Psychology, 35, 805-812.

Lance, C.E., Newbolt, W.H., Gatewood, R.D., and Smith, D.E. (1995, May). 'Assessment center exercise factors represent cross-situational specificity, not method bias', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Landy, F. J., and Farr, J.L. (1980). 'Performance rating', Psychological Bulletin, 87, 72-107.

Lebreton, J.M., Binning, J.F., and Hesson-McInnis, M.S. (1998, August). 'The effects of measurement structure on the validity of assessment center dimensions: The clinical-statistical debate revisited', Paper presented at the Annual Meeting of the Academy of Management, San Diego, CA.

Lebreton, J.M., Gniatczyk, L.A., and Migetz, D.Z. (1999, May). 'The relationship between behavior checklist ratings and judgmental ratings in an operational assessment center: An application of structural equation modeling', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Lerner, J.S., and Tetlock, T.E. (1998). 'Accounting for the effects of accountability', Psychological Bulletin, 125, 255-275.

Lichtenstein, M., and Srull, T.K. (1987). 'Processing objectives as a determinant of the relationship between recall and judgement', Journal of Experimental Social Psychology, 23, 93-118.

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

Lievens, F. (1999, May). 'The effects of type of assessor training on the construct validity and accuracy of assessment center ratings', Paper presented at the European Congress of Work and Organizational Psychology, Espoo- Helsinki, Finland.

Lievens, F. (in press). 'Assessors and use of assessment center dimensions: A fresh look at a troubling issue'. Journal of Organizational Behavior.

Lievens, F., and Goemaere, H. (1999). 'A Different Look at Assessment Centers: Views of Assessment Center Users', International Journal of Selection and Assessment, 7, 215-219.

Lievens, F., and Van Keer, E. (1999, May). 'Modeling method effects in assessment centers: An application of the correlated uniqueness approach', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Lord, R.G., and Maher, K.J. (1990). 'Alternative information-processing models and their implications for theory, research, and practice', Academy of Management Review, 15, 9-28.

Lowry, P.E. (1993). 'The assessment center: An examination of the effects of assessor characteristics on assessor scores', Public Personnel Management, 22, 487-501.

Lowry, P.E. (1996). 'A survey of the assessment center process in the public sector', Public Personnel Management, 25, 307-321.

Macan, T.H., Avedon, M.J., Paese, M., and Smith, D.E. (1994). 'The effects of applicants' reactions to cognitive ability tests and an assessment center', Personnel Psychology, 47, 715-738.

Maher, P.T. (1990, March). 'How many dimensions are enough?', Paper presented at the International Congress on the Assessment Center Method, Orange, CA.

Maher, P.T. (1995, May). 'An analysis of the impact of the length of assessor training on assessor competency', Paper presented at the International Congress on the Assessment Center Method, Kansas City.

Marsh, H.W. (1989). 'Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions', Applied Psychological Measurement, 13, 335-361.

Mayes, B.T., Belloli, C.A., Riggio, R.E., and Aguirre, M. (1997). 'Assessment centers for course evaluations: A demonstration', Journal of Social Behavior and Personality, 12, 303-320.

McCredie, H., and Shackleton, V. (1994). 'The development and interim validation of a dimensions-based senior management assessment centre', Human Resurce Management Journal, 5, 91-101.

Mero, N.P., and Motowidlo, S.J. (1995). 'Effects of rater accountability on the accuracy and the favorability of performance ratings', Journal of Applied Psychology, 80, 517-524.

Mitchell, D.R.D., and Maurer, T.J. (1998, August). 'Assessment center feedback in relation to subsequent human resource development activity', Paper presented at the Annual Meeting of the Academy of Management, San Diego, CA.

Morrow, P.C., McElroy, J.C, Stamper, B.G., and Wilson, M.A. (1990). 'The effects of physical attractiveness and other demographic characteristics on promotion decisions', Journal of Management, 16, 723-736.

Moser, K., Diemand, A., and Schuler, H. (1996). 'Inconsistency and social skills as two components of self-monitoring [In German]', Diagnostica, 42, 268-283.

Moser, K., Schuler, H., and Funke, U. (1999). 'The moderating effect of raters' opportunities to observe ratees' job performance on the validity of an assessment centre', International Journal of Selection and Assessment, 7, 133-141.

Moses, J.L., and Ritchie, R.J. (1976). 'Supervisory relationships training: a behavioral evaluation of a behavior modeling program', Personnel Psychology, 29, 337-343.

Murphy, K.R., and Cleveland, J.N. (1995). 'Understanding performance appraisal', Thousands Oaks: Sage.

Murphy, K.R., Jako, R.A., and Anhalt, R.L. (1993). 'The nature and consequences of halo error: a critical analysis', Journal of Applied Psychology, 78, 218-225.

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

Neubauer, R. (1990). 'Women in the career assessment center--a victory? [In German]', Zeitschrift für Arbeits und Organisationspsychologie, 34, 29-36.

Neuberg, S.L. (1989). 'The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies', Journal of Personality and Social Psychology, 56, 374-386.

Newell, S., & Shackleton, V. (1994). Guest Editorial: International differences in selection methods. International Journal of Selection and Assessment, 2, 71-73.

Nowack, K.M., (1997). 'Congruence between self-other ratings and assessment center performance', Journal of Social Behavior and Personality, 12, 145-166.

Petty, M.M. (1974). 'A multivariate analysis of the effects of experience and training upon performance in a leaderless group discussion', Personnel Psychology, 27, 271-282.

Pynes, J., and Bernardin, H.J. (1992). 'Mechanical vs. consensus-derived assessment center ratings: A comparison of job performance validities', Public Personnel Management, 21, 17-28.

Ramos, R.A. (1992). 'Testing and assessment of Hispanics for occupational and management positions: A developmental needs analysis', In K.F. Geisinger (Ed.), Psychological Testing of Hispanics (pp. 173-194). Washington, DC: American Psychological Association.

Reilly, R.R., Henry, S., and 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.

Rolland, J.P. (1999). 'Construct validity of in-basket dimensions'. European Review of Applied Psychology, 49, 251-259.

Ross L., and Nesbitt, R.F. (1991). 'The person and the situation', Perspectives of Social Psychology, New York: McGraw Hill.

Rotenberry, P.F., Barrett, G.V., and Doverspike, D. (1999, May). 'Determination of systematic bias for an objectively scored in-basket assessment', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Rumelhart, D.E., and Ortony, A. (1977). 'The representation of knowledge in memory', In R.C. Anderson, R.J. Spiro, and W.E. Montague (Eds.), Schooling and the acquisition of knowledge (pp. 99-136). Hillsdale, NJ: Lawrence Erlbaum.

Russell, C.J., and Domm, D.R. (1995). 'Two field tests of an explanation of assessment centre validity', Journal of Occupational and Organizational Psychology, 68, 25-47.

Ryan, A.M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., and McCormick, S. (1995). 'Direct, indirect, and controlled observation and rating accuracy', Journal of Applied Psychology, 80, 664-670.

Rynes, S.L., and Connerly, M.L. (1993). 'Applicant reactions to alternative selection procedures', Journal of Business and Psychology, 7, 261-278.

Sackett, P.R. (1998). 'Performance assessment in education and professional certification: Lessons for personnel selection ?', In M.D. Hakel (Ed.), Beyond multiple choice (pp. 113-129). Mahwah, NJ: Lawrence Erlbaum.

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

Sagie, A., and Magnezy, R. (1997). 'Assessor type, number of distinguishable dimension categories, and assessment centre construct validity', Journal of Occupational and Organizational Psychology, 70, 103-108.

Schleicher, D.J., Day, D.V., Mayes, B.T., and Riggio, R.E. (1999, May). 'A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Schmitt, N. (1993). 'Group composition, gender, and race effects on assessment center ratings', In H. Schuler, J.L. Farr, and M. Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives (pp. 315-332). Hillsdale, NJ: Lawrence Erlbaum.

Schneider, J.R., and Schmitt, N. (1992). 'An exercise design approach to understanding assessment center dimension and exercise constructs', Journal of Applied Psychology, 77, 32-41.

Scholz, G., and Schuler, H. (1993). 'The nomological network of the assessment center: A meta-analysis [In German]', Zeitschrift für Arbeits- und Organisationspsychologie, 37, 73-85.

Schuler, H. Moser, K., and Funke, U. (1994, August). 'The moderating effect of rater-ratee acquaintance on the validity of an assessment center', Paper presented at the International Congress of Applied Psychology, Madrid, Spain.

Shechtman, Z. (1998). 'Agreement between lay participants and professional assessors: Support of a group assessment procedure for selection purposes', Journal of Personnel Evaluation in Education, 12, 5-17.

Shore, T.H. (1992). 'Subtle gender bias in the assessment of managerial potential', Sex Roles, 27, 499-515.

Shore, T.H., Shore, L.M., and Thornton, G.C. III (1992). 'Construct validity of self- and peer evaluations of performance dimensions in an assessment center', Journal of Applied Psychology, 77, 42-54.

Shore, T.H., Taschian, A., and Adams, J.S. (1997). 'The role of gender in a developmental assessment center', Journal of Social Behavior and Personality, 12, 191-203.

Shore, L.M., Tetrick, L.E., and Shore, T.H. (1998). 'A comparison of self-, peer, and assessor evaluations of managerial potential', Journal of Social Behavior and Personality, 13, 85-101.

Shore, T.H., Thornton, G.C. III, and Shore, L.M. (1990). 'Construct validity of two categories of assessment center ratings', Personnel Psychology, 43, 101-116.

Sichler, R. (1991). 'Experiences and activities with an assessment center procedure: An empirical contributoin to the "social validity" of test situations [In German]', Zeitschrift für Arbeits- und Organisationspsychologie, 33, 139-145.

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

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., and Stoffey, R.W. (1993). 'Applicant reactions to selection procedures', Personnel Psychology, 46, 49-78.

Smith-Jentsch, K.A. (1996, April). 'Should rating dimensions in situational exercises be made transparent for participants? Empirical tests of the impact on convergent and predictive validity', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Spychalski, A.C., Quinones, M.A., Gaugler, B.B, and Pohley, K.A. (1997). 'A survey of assessment center practices in organizations in the United States', Personnel Psychology, 50, 71-90.

Srull, T.K., and Wyer, R.S. (1980). 'Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgment', Journal of Personality and Social Psychology, 38, 841-856.

Srull, T.K., and Wyer, R.S. (1989). 'Person memory and judgment', Psychological Review, 96, 58-83.

Staufenbiel, T., and Kleinmann M. (1999, May). 'Does P-O fit influence the judgments in assessment centers', Paper presented at the European Congress of Work and Organizational Psychology, Espoo- Helsinki, Finland.

Steiner, D.D., and Gilliland, S.W. (1996). 'Fairness reactions to personnel selection techniques in France and the United States', Journal of Applied Psychology, 81, 134-147.

Tan, M. (1996). 'The effects of role-player standardization on the construct validity of dimensions in assessment exercises [in Dutch]', Unpublished doctoral dissertation, University of Amsterdam.

Task Force on Assessment Center Guidelines (1989). 'Guidelines and ethical considerations for assessment center operations', Public Personnel Management, 18, 457-470.

Tetlock, P.E. (1983). 'Accountability and complexity of thought', Journal of Personality and Social Psychology, 45, 74-83.

Tett, R.P. (1998, April). 'Traits, situations, and managerial behavior: Test of a trait activation hypothesis', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Tett, R.P. (1999, May). 'Assessment center validity: New perspectives on an old problem', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Thornton, G.C. III (1992). 'Assessment centers in Human Resource Management', Reading, MA: Addison-Wesley.

Thornton, G.C. III, Larsh, S. Layer, S., and Kaman, V. (1999, May). 'Reactions to attribute-based feedback and exercise-based feedback in developmental assessment centers', Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Van Dam, K., Altink, W.M.M., and Kok, B. (1992). 'Assessment center practice: A summary of problems encountered [In Dutch]', De Psycholoog, 7, 509-514.

Van der Velde, E.G., Born, M.P., and Hofkes, K. (1994). 'Construct validity of an assessment center using confirmatory factor analysis [In Dutch]', Gedrag en Organisatie, 7, 18-26.

Walsh, J.P., Weinberg, R.M., and Fairfield, M.L. (1987). 'The effects of gender on assessment centre evaluations', Journal of Occupational Psychology, 60, 305-309.

Weijerman, E.A.P., and Born, M.P. (1995). 'The relationship between gender and assessment center scores [In Dutch]', Gedrag en Organisatie, 8, 284-292.

Woodruffe, C. (1993). 'Assessment centres: Identifying and developing competences', London: Gower.

Zedeck, S. (1986). 'A process analysis of the assessment center method', Research in Organizational Behavior, 8, 259-296.

Figure 1. Component Model of Assessment Centers

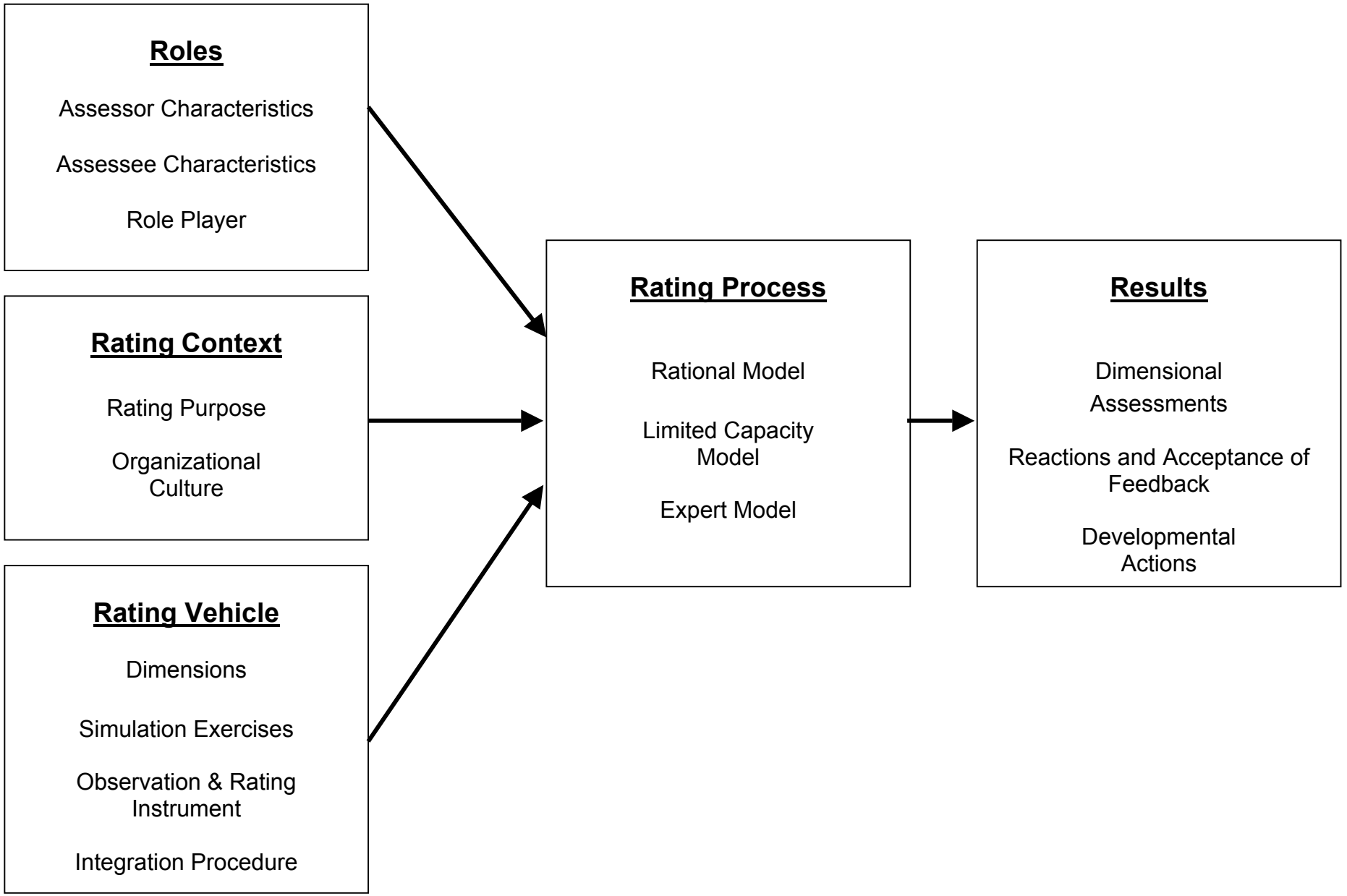

Figure 2. Assessment Centers and the Social Judgment Process