1-2009

Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises

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

Robert P. TETT

Deidra J. SCHLEICHER

ASSESSMENT CENTERS AT THE CROSSROADS: TOWARD A RECONCEPTUALIZATION OF ASSESSMENT CENTER EXERCISES

Filip Lievens, Robert P. Tett and Deidra J. Schleicher

ABSTRACT

Exercises are key components of assessment centers (ACs). However, little is known about the nature and determinants of AC exercise performance. The traditional exercise paradigm primarily emphasizes the need to simulate task, social, and organizational demands in AC exercises. This chapter draws on trait activation theory in proposing a new AC exercise paradigm. First, we develop a theoretical framework that addresses the complexity of situational characteristics of AC exercises as determinants of AC performance. Second, we argue for planting multiple stimuli within exercises as a structured means of eliciting candidate behavior. Third, we show how the new paradigm also has key insights for the rating part of ACs, namely, in selecting dimensions, designing behavioral checklists, screening assessors, and training assessors. Finally, the impact of this new AC exercise paradigm is anticipated on important AC outcomes such as reliability, internal/external construct-related validity, criterion-related validity, assessee perceptions, and feedback effectiveness.

Research in Personnel and Human Resources Management, Volume 28, 99–152

ISSN: 0742-7301/doi:10.1108/S0742-7301(2009)0000028006

First developed approximately 60 years ago, assessment centers (ACs) continue to be a popular and potent tool in organizations for evaluating candidates for selection and promotion, as well as for identifying strengths and weaknesses for developmental purposes (Spychalski, Quinones, Gaugler, & Pohley, 1997; Thornton & Rupp, 2005). As high-fidelity simulations, ACs focus on actual candidate behavior that is observed and evaluated by multiple trained assessors on multiple job-related dimensions in multiple job-related exercises. After decades of research on the construct-related validity and workings of ACs, it is now generally acknowledged that ACs are at a crossroads. In a recent series of papers on ACs (Lance, 2008), many researchers and practitioners concluded that AC exercises and behaviors are the currency of ACs. They also suggested that the traditional paradigm underlying ACs (i.e., that they work because they validly assess candidates on important dimensions across exercises) should be replaced. Specifically, Lance (2008) posited that "a 'back to basics' focus combined with contemporary psychometric rigor may lead to the development of ACs that do what ACs do, simulate important roles and stimulate behaviors related to these roles better than ACs do today" (p. 144).

However, conceptually we do not know a lot about the factors that stimulate candidate behavior in AC exercises. Moreover, on a practical level, little is also known about how AC exercises might be designed to better elicit job-relevant behavior. Therefore, the general objective of this chapter is to advance the field of ACs by providing a new theory-driven paradigm for designing and evaluating AC exercises, one based on a recent interactionist theory of job performance, namely, trait activation theory (Tett & Burnett, 2003). To be clear, the goal of this chapter is not to argue for one side or the other in the "dimensions versus exercise" debate within ACs. Regardless of which perspective researchers and practitioners take, the fact remains that the literature provides little guidance regarding determinants of candidate behavior in exercises and, correspondingly, how AC exercises should best be constructed to elicit job-relevant behavior. Thus, the paradigm presented in this chapter offers theoretically grounded yet practically useful suggestions for both approaches to ACs. Specifically, the reconceptualization of AC exercises provided in this chapter contributes to AC research and practice by suggesting (a) a theoretical framework that addresses the complexity of situational characteristics of AC exercises as determinants of AC performance; (b) recommendations about how to design exercises to better elicit candidate behaviors; and (c) recommendations about how to better evaluate candidate behaviors in ACs.

We start with a brief review of recent AC research. Next, we develop a theoretical framework of factors delineating how candidate behaviors are

elicited and evaluated in AC exercises. Finally, directions for future research are discussed.

PRIOR RESEARCH ON ASSESSMENT CENTER EXERCISES

Generally, AC exercises may be divided into three groups: individual exercises (e.g., in-basket, planning exercise, case analysis), one-to-one exercises (e.g., role play, interview, fact-finding, presentation), and group exercises (e.g., leaderless group discussion). At least three lines of research highlight the importance of exercises in the AC framework: (a) research that has tried to decompose the variability in candidate ratings into dimension variance and exercise variance, (b) research that has tested whether exercise variability represents method bias or true performance variability, and (c) research on the factors determining exercise performance.

Exercises as Key Components in Assessment Centers

More than 25 years ago, Sackett and Dreher (1982) published a seminal article in which they investigated AC ratings in three organizations. In each of these organizations, they found low correlations among ratings of a single dimension across exercises (i.e., weak convergent validity) and high correlations among ratings of various dimensions within one exercise (i.e., weak discriminant validity). Furthermore, factor analyses indicated more evidence for exercise factors than for dimension factors. Although these findings seemed ''troublesome'' at the time, the authors emphasized that they do not mean that ACs lack construct-related validity. Sackett and Tuzinski (2001) again cautioned against this misinterpretation of the basic findings, noting ''Assessment centers do not lack 'construct validity,' but rather lack clear consensus as to the constructs they assess'' (pp. 117–188).

The findings of Sackett and Dreher (1982) have proven to be very robust as they have been found in both selection and developmental ACs. In addition, they have been found in ACs conducted all over the world. Apart from the United States (Bycio, Alvares, & Hahn, 1987; Harris, Becker, & Smith, 1993; Kudisch, Ladd, & Dobbins, 1997; Reilly, Henry, & Smither, 1990; Sackett & Dreher, 1982; Schneider & Schmitt, 1992; Silverman, Dalessio, Woods, & Johnson, 1986), these results have been established in the United Kingdom (Anderson, Lievens, van Dam, & Born, 2006; Crawley, Pinder, & Herriot, 1990; Robertson, Gratton, & Sharpley, 1987), Germany

(Kleinmann & Köller, 1997; Kleinmann, Kuptsch, & Köller, 1996), Belgium (Lievens & Van Keer, 2001), France (Borteyrou, 2005; Rolland, 1999), Australia (Atkins & Wood, 2002), New Zealand (Jackson, Stillman, & Atkins, 2005), China (Wu & Zhang, 2001), and Singapore (Chan, 1996).

Interestingly, findings of situation-specific variance being larger than construct variance are not unique to ACs. Similar results have been obtained for other method-driven predictors such as structured interviews (Conway & Peneno, 1999; Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004) and situational judgment tests (Trippe & Foti, 2003). For example, convergence in constructs measured by different types of structured interviews (behavior description and situational interviews) has been low (e.g., Van Iddekinge et al., 2004). Moreover, the findings seem to extend to all fields wherein different constructs are measured in multiple performance-based exercises. For example, predominance of situation-specific variance over dimension variance has been found in studies about patient-management problems for physicians (e.g., Julian & Schumacher, 1988), military examinations (e.g., Shavelson, Mayberry, Li, & Webb, 1990), hands-on science tasks (e.g., Shavelson et al., 1991), bar examinations (e.g., Klein, 1992), and direct writing assessments (e.g., Dunbar, Koretz, & Hoover, 1991). Clearly, the ''method variance predominance'' is far from unique to ACs, suggesting that the source of the problem runs deep into the nature of human behavior in structured tasks.

In recent years, three large-scale studies have been conducted to quantitatively summarize the construct-related validity findings of AC ratings. First, Lievens and Conway (2001) reanalyzed 34 multitrait-multimethod (MTMM) matrices of AC ratings. Their main conclusion was that a model consisting of exercises (specified as correlated uniquenesses) and dimensions represented the best fit to the data. In this model, exercises and dimensions explained the same amount of variance (34%). In addition, dimensions were found to correlate substantially (.71).

A second quantitative review came to different conclusions (Lance, Foster, Gentry, & Thoresen, 2004a; Lance, Lambert, Gewin, Lievens, & Conway, 2004b). According to Lance et al., Lievens and Conway's (2001) results of exercises and dimensions explaining about the same amount of variance were due to a statistical artifact (i.e., the use of the correlated uniqueness model that systematically overestimated dimension variance). In their reanalysis, a model with correlated exercises and one general dimension prevailed. In addition, exercise variance (52%) was clearly more important than dimension variance (14%).

Recently, Bowler and Woehr (2006) conducted a third quantitative review because a limitation inherent in the two prior quantitative reviews was that each MTMM matrix was individually reanalyzed. Hence, estimates of average dimension and exercise variance were based on confirmatory factor analysis (CFA) results from models with different sample sizes, dimensions, and exercises. Bowler and Woehr used meta-analytic methods to combine 35 MTMM matrices into one single matrix. The best fit was obtained for a model with correlated dimensions and exercises. Exercises explained most of the variance (33%). The in-basket and presentation exercises, in particular, accounted for large parts of variance. Dimensions also explained a substantial amount of variance (22%). In addition, some dimensions (i.e., communication, influencing others, organizing and planning, and problem solving) explained significantly more variance than others (i.e., consideration/awareness of others, drive).

One possible explanation for the large exercise variance might be that it represents not only variability across exercises but also variability across assessors. This confounding is due to the common practice of assessors rotating through the various exercises. Indeed, to save costs, a given assessor does not evaluate each candidate in each exercise. However, two research studies have discounted this explanation (Kolk, Born, & van der Flier, 2002; Robie, Osburn, Morris, Etchegaray, & Adams, 2000). In both studies, exercise variance was separated from assessor variance by asking one assessor to rate only one dimension per exercise. Although this rating method led in both studies to more evidence for dimension factors, exercise factors were still predominant. This is conceivable in light of the fact that inter-rater reliability among assessors has typically been satisfactory (Thornton & Rupp, 2005). Hence, controlling for assessor variance seems to have only marginal effects.

In short, decades of research on the interplay of the three components of ACs (dimensions, exercises, and assessors) reveals that exercises are predominant in terms of explaining variance in assessee behavior. That is not to say that the other two components (dimensions and assessors) should be ignored or are unimportant. Rather, it suggests that more research is needed on the nature and determinants of exercise performance.

Exercise Variability as True Cross-Situational Performance Fluctuations

A second line of research has tried to determine whether the large variability in candidate ratings across exercises is actually "troublesome" – that is, does

exercise variance represent "unwanted" method bias or "true" cross-situational variability? AC architects did not originally conceptualize exercises as merely parallel measures (Howard, 2008; Neidig & Neidig, 1984). Instead, an AC is purported to consist of several exercises carefully selected to cover specific job-related competencies. Consequently, different exercises might place different psychological demands on assessees. For instance, one might expect an assessee to behave differently – even inconsistently – in a one-on-one role play as compared to a group discussion.

Evidence from both field and laboratory research supports the explanation of candidates behaving inconsistently across structurally different exercises (Lance, 2008; Lievens, 2008). In various field studies, Lance and colleagues (Lance et al., 2000; Lance et al., 2004a; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007) correlated exercise factors with external variables such as job performance and cognitive ability. They hypothesized that, if exercise factors constituted unwanted method bias, exercise factors and performance criteria should be unrelated. Conversely, if exercise effects reflect true cross-situational specificity in performance, positive relations between exercise factors and performance criteria should emerge. Results confirmed the latter option, illustrating that exercise factors do not represent unwanted method bias, but rather true performance differences. In another study, Hoeft and Schuler (2001) tried to estimate the amount of variability in AC performance across situations (i.e., exercises). Their study revealed that AC performance was more situation-specific (57%) than situation-consistent (43%).

Laboratory studies led to similar conclusions about the nature of performance in ACs. Lievens (2001a, 2002) examined the effects of both type of assessee performance and type of assessor. In particular, Lievens (2002) asked three types of assessors (I/O psychologists, managers, and students) to rate assessees whose performance varied along two continua: cross-exercise consistency (i.e., relatively inconsistent vs. relatively consistent) and dimension differentiation (i.e., relatively undifferentiated vs. relatively differentiated). Assessor ratings were analyzed for convergent validity, discriminant validity, and inter-rater reliability. Results showed large differences in evidence for convergent and discriminant validity across type of assessee performance. In fact, convergent validity was established only for consistent performance across exercises, whereas discriminant validity was established only for differentiated performance across dimensions. Granted, evidence for convergent and discriminant validity also varied across type of assessor; however, these differences were smaller.

In particular, evidence for discriminant and convergent validity was more clearly established with I/O psychologists and managers than with students. Overall, this study highlights that assessors appear to be relatively accurate and that the nature of candidate performance is a key factor to establish AC construct-related validity. Only when candidates perform consistently across exercises and heterogeneously across dimensions could evidence of construct-related validity be established. However, these would be relatively rare conditions and unreasonable expectations, given that AC exercises are designed to present diverse demands calling for diverse responses.

In sum, the foregoing review shows that a paradigm shift has occurred in thinking about ACs over the years. First, exercises are no longer viewed as alternate measures of the same dimensions (which are also no longer seen as simply stable traits that manifest themselves consistently across all situations). Rather, differences in exercises are substantively relevant to activating assessee behavior. Second, assessors should no longer be regarded as fundamentally flawed but as capable of making relatively accurate ratings (given suitable selection and training). Third, and more generally, the large exercise variance is no longer seen as "troublesome," as reflected in the statements of Lance et al. (2004a, 2004b) – "There may be nothing wrong with assessment center's construct validity after all" (p. 23, refer also Lance, 2008) – or Ployhart (2006) – "Research may be close to solving the construct-related validity question for assessment centers" (p. 881).

Situational Factors Determining Exercise Performance

Only a handful of studies have explored situational characteristics of AC exercises as possible determinants of performance variations across exercises. Schneider and Schmitt (1992) experimentally manipulated the effects of exercise content (competitive vs. cooperative demands) and exercise form (e.g., role play vs. group discussion). Variance due to the form of the exercise emerged as the most important exercise factor to explain the variability in ratings across exercises. More specifically, exercise form explained 16% of the exercise variance in ratings. The effect of exercise content, on the contrary, was negligible.

Highhouse and Harris (1993) tried to create a more comprehensive taxonomy of the nature of exercises in the typical AC, through identification of performance constructs underlying the AC exercises, and then to determine the effect of this on ratings. First, they extracted assessee behaviors 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) used by assessors. Then, experienced assessors were asked to use these performance constructs to describe the ideal AC candidate in each exercise. Highhouse and Harris concluded that assessors perceived the exercise situations to be generally unrelated in terms of the behaviors required for successful performance, underscoring the notion that exercises are structurally different.

Highhouse and Harris also discovered some evidence that assessees were 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, two exercises that assessors saw 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.

More recently, McFarland, Yun, Harold, Viera, and Moore (2005) linked the form of the exercise to a specific type of performance in ACs, namely, impression management. They hypothesized that candidates would exhibit more impression management in AC exercises that triggered more interpersonal behavior than in exercises that primarily activated technical knowledge. Consistent with their hypothesis, impression management tactics were more frequently used and there was more variability in impression management use for exercises that placed greater demands on candidates' interpersonal competencies (e.g., role play and oral presentation vs. tactical exercise).

This finding of different exercises activating different behaviors and traits has also been confirmed in research examining the notes of assessors. Lievens, De Fruyt, and Van Dam (2001) studied trait descriptors in assessor notes and found differences between AC exercises in terms of the personality adjectives noted, with particular personality traits linked to specific exercises. For example, in group discussions, assessors reported mainly extraversion adjectives, while conscientiousness markers were more frequently noted in scoring the in-basket exercise.

In sum, a limited set of studies have tried to determine which specific exercise factors might lead candidates to perform differently across exercises. This scarce research has only just scratched the surface as the general exercise form (i.e., the type of exercise) emerged as the major determinant instead of specific situational stimuli within exercises. For

example, research has not revealed which characteristics inherent in a specific exercise form (e.g., role play) might generate a more complete and varied set of behaviors than another form (e.g., group discussion).

Summary

In this section, we reviewed various lines of research that all point in the same direction for AC research and practice: exercises are key components of ACs. So far, however, little is known about the important situational characteristics on which AC exercises vary, prompting the need for additional research that can provide a deeper conceptual understanding of the factors determining exercise performance and ratings. This call for exercise research is long overdue. For instance, in 1992, Schneider and Schmitt posited that "discovering a set of exercise-based factors ... not only can advance understanding of how assessment centers work but may also offer practical suggestions to aid the exercise development process" (p. 32). To this end, the next sections delineate a theory of performance in ACs (refer Borsboom, Mellenbergh, & Van Heerden, 2004) that could provide an impetus and organizing framework for researchers interested in AC performance and exercise issues as well as practitioners interested in AC design.

TRAIT ACTIVATION AS A RECENT PERSON-SITUATION FRAMEWORK

In ACs, candidates participate in various exercises, which are essentially different situations. Thus, to make well-grounded evaluations about a candidate's performance in an AC, it is critical to understand how behavior is expressed and how it is (or should be) evaluated in different situations (Lievens, De Koster, & Schollaert, 2008). Our approach to these issues has its foundations in the historical debate in personality and social psychology over the relative importance of traits and situations as sources of behavioral variance. Reconciliation between the extremes of "personism" and "situationism" is now generally recognized in the form of person-situation interactionism (e.g., refer Bowers, 1973; Ekehammar, 1974; Epstein & O'Brien, 1985; Johnson, 1997), which allows that people can behave consistently across different situations and that situations can cause different people to behave similarly. The key to interactionism is that

behavior is considered a multiplicative function of traits and situations. That is, the degree to which a trait affects behavior *depends on* the situation. Traits and situations, in this light, are inseparable, forming two sides of a single coin.

Building on early works by Murray (1938) and Allport (1951), trait activation theory (Tett & Burnett, 2003; Tett & Guterman, 2000) applies interactionist principles in work settings to clarify how individual traits come to be expressed as work-related behavior and how such behavior comes to be related to job performance. Fig. 1 depicts the main ideas behind trait activation theory. It starts with the common notion that a person's trait level is expressed as trait-relevant work behavior. Apart from the main effect of situations on work behavior (and vice versa), trait activation posits four key axioms. Given the central importance of this theory to our paradigm for considering AC exercises, we review each of these axioms in some detail here.

The first axiom is that traits will manifest as trait-expressive work behaviors only as responses to trait-relevant situational cues. Such cues are considered to fall into three broad and interrelated categories: task, social,

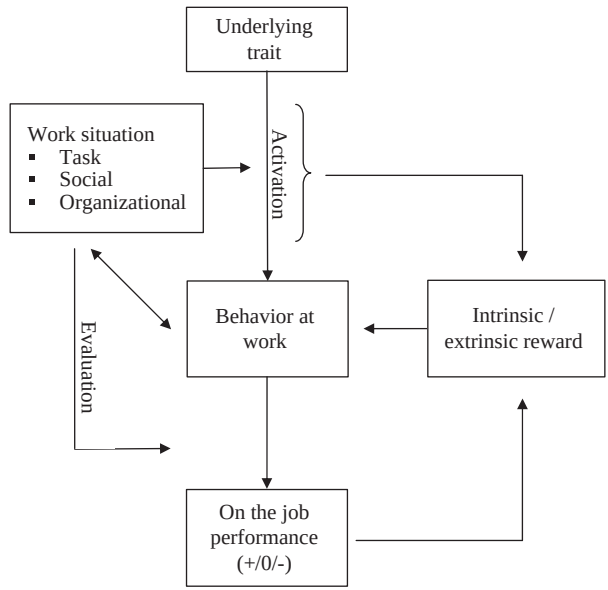

Fig. 1. Schematic Overview of Trait Activation Theory and Job Performance.

and organizational. The need for autonomy, for example, may be activated by arbitrarily structured tasks, a rule-driven boss, and in dealings with bureaucratic organizations. The common theme linking these situations is restriction in choice behavior, a cue directly relevant to the trait of autonomy. *Situation trait relevance* is a qualitative feature of situations that is essentially trait-specific. Notably, it provides a direct answer to the longstanding quest for an all-encompassing taxonomy of situational characteristics by describing situations on the basis of the traits themselves. Thus, just as the Big Five personality taxonomy offers a useful framework for organizing diverse traits (e.g., Costa & McCrae, 1992; Goldberg, 1992; Haaland & Christiansen, 2002; Lievens et al., 2001), it also offers a parallel framework for organizing and depicting diverse situations.

A second axiom underlying trait activation theory is that trait expression is dependent not only on the relevance of the situation but also on the strength of the situation. *Situation strength* is the degree to which a situation leads everyone to respond the same way (Mischel, 1973). A strong situation overrides the effects of individual differences in behavioral propensities; traits are evident as differences among individuals' behavior only to the degree the situation is weak. Using color as an analogy, situation trait relevance and situation strength are separable just as hue is from brightness. Hue, like trait relevance, is a qualitative feature (e.g., red vs. blue; autonomy-relevant vs. sociability-relevant), whereas brightness, akin to strength, captures the intensity of the hue (e.g., dull blue vs. bright blue; sociability at a funeral vs. a party). Each property conveys something uniquely critical in describing color and situations. In the latter case, differences among individuals on a particular trait will be evident to the degree that (a) the situation offers cues to express that trait and (b) the situation is not so "bright" as to demand that everyone respond the same way.

The third axiom of trait activation theory is that trait-expressive work behavior is distinct from job performance, the latter defined specifically as *valued work behavior*. As depicted in Fig. 1, trait-expressive work behavior may be rated by others positively (+), negatively (–), or neutrally (0), depending on the degree to which it is perceived to meet task, social, and organizational demands (i.e., Does it get the job done? Does it meet group expectations? Is it consistent with organizational values?). Demands at each level that serve as cues for trait expression thus also serve as reference points for evaluating behavior as performance. Separating work behavior from job performance (i.e., as valued behavior) is critical to understanding trait-performance linkages because the processes underlying trait expression

(i.e., trait activation) are fundamentally different from those underlying performance appraisal (i.e., observer judgments).

The fourth axiom of trait activation theory is that trait expression entails two parallel, motivational reward systems. Building on the need model of personality traits (e.g., Murray, 1938), *intrinsic reward* derives from trait expression per se (i.e., as need fulfillment). Concomitantly, *extrinsic reward* derives from the reactions of others to one's trait expressions. Person-situation fit, accordingly, is maximized where intrinsic and extrinsic rewards are aligned: *people want to work on tasks, work with other people, and work in organizations where they are rewarded for being themselves*. By the same token, fit will be poor in situations lacking cues relevant to the person's dispositions and, even worse, when such cues are present but invite negative reactions from others when acted upon (Tett & Burnett, 2003, called such cues "distracters"). Note also that intrinsic and extrinsic rewards clarify the concept of situation strength: a strong situation is one in which the extrinsic rewards for behaving in a particular way override the intrinsic rewards of trait expression. Thus, a trait will find expression as work behavior and, through evaluation, as job performance only when the extrinsic rewards are not so powerful as to cause everyone to respond the same way.

These four axioms of trait activation theory (bearing on situation trait relevance, situation strength, trait expression vs. job performance, and trait-based intrinsic vs. extrinsic reward systems) offer critical insights into AC principles, findings, and future developments. It is to these topics that we now turn.

TRAIT ACTIVATION IN ASSESSMENT CENTERS

Fig. 2 shows how ACs can be framed in trait activation theory. In ACs, a person's trait level is measured as a score on a dimension that is based on behavior in various AC exercises. Recent taxonomic work has sorted the various dimensions into seven major categories: communication, consideration/awareness of others, drive, influencing others, organizing and planning, problem solving, and tolerance for stress (Arthur, Day, McNelly, & Edens, 2003). All of these dimensions can be linked to deeper underlying traits. For example, stress tolerance is related to the underlying trait of emotional stability, whereas planning and organizing is related to conscientiousness, and communication is related to aspects of extraversion (Lievens, Chasteen, Day, & Christiansen, 2006).

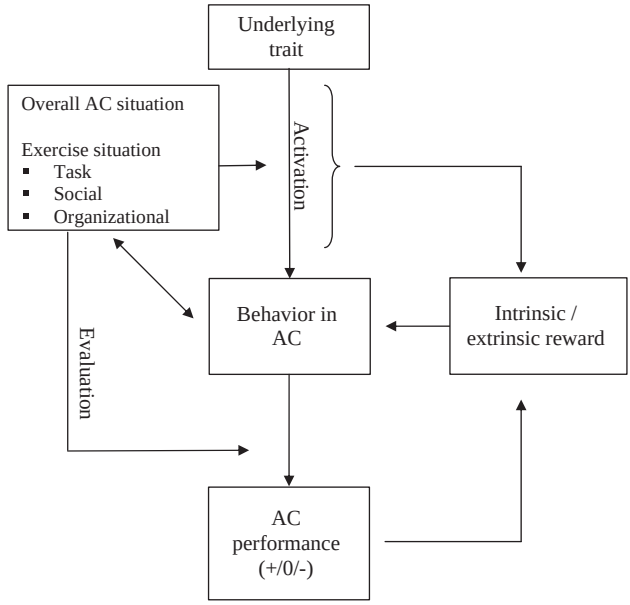

Fig. 2. Schematic Overview of Trait Activation Theory and AC Performance.

Of course, performance on any given AC dimension is also likely to reflect cognitive ability (as in problem solving) and related skills. Tett and Burnett (2003) suggest that ability plays out in situations much as personality traits do, as responses to ability-relevant situational demands (refer also Goldstein, Yusko, Braverman, Smith, & Chung, 1998). There are two main differences between ability and personality traits. First, unlike a need-based personality trait, ability lacks intrinsic motivation potential: people engage their abilities either for the promise of extrinsic rewards (e.g., getting the job) or to satisfy companion personality traits (e.g., need for achievement, need for recognition). In this light, ability provides the "can do" and personality the intrinsic "will do" of performance in ACs. Second, whereas ability is inherently positively valued (having more ability is never a bad thing), the value of a personality trait depends on trait-specific situational demands: a highly nurturant manager, for example, may be appreciated by needy subordinates, but rejected by those more independent. Full understanding of AC performance requires integration of ability and personality effects.

In light of the foregoing analysis, we view AC exercises as situations differing in terms of their trait activation (and therefore behavior-eliciting)

potential, as determined by the availability of trait-relevant exercise cues and the extrinsic rewards governing situation strength. These two factors are discussed in the following sections.

Embedding Behavior-Eliciting Cues in Assessment Center Exercises

Situational effects in ACs operate at three levels (Fig. 3). At the broader level, the AC is, in its entirety, an evaluative situation (the top level of Fig. 3). For use in selection/promotion, it is competitive, and for use in development/training, it presents learning opportunities for assessees. Traits likely to be activated in the former case (regardless of specific exercises used) include ability, ambition, emotional control, and perhaps also risk-taking; traits relevant in the latter case include curiosity and self-actualization. At a narrower level, a given AC exercise (the second level in Fig. 3) presents a distinct set of job-relevant demands/cues relevant to the exercise as a whole. An in-basket test, for example, presents largely administrative demands, and a group exercise, largely interpersonal demands. Then, looking within each exercise (the third level in Fig. 3), there are cues varying in problem content (e.g., worker-centered vs. task-centered), complexity, clarity, importance, urgency, mode of presentation (e.g., on paper, by computer), and other variables. Traits and abilities, we suggest, are activated at each level independently of the other, with or without redundancy across levels. In

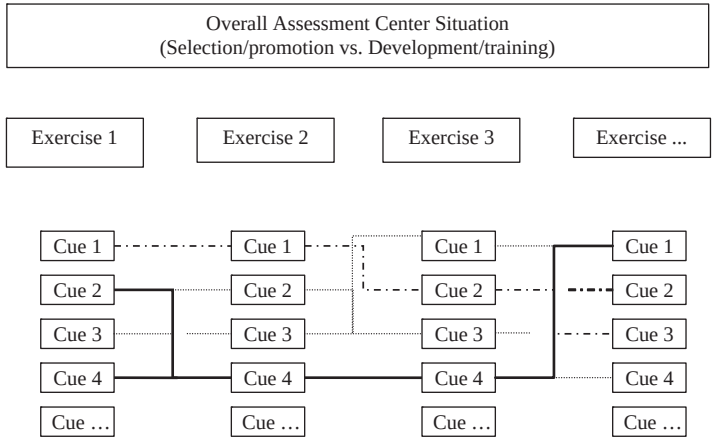

Fig. 3. Three Levels of Situational Effects in Assessment Centers.

Fig. 3, situational cues that hypothetically trigger the same traits are linked by the same type of line.

Note that, at the broader level, behavioral inconsistency across diverse exercises may belie a deeper trait consistency. For example, an ambitious and bright assessee might show detailed planning in a planning exercise and decisiveness in a decision-making exercise. Thus, what appears at the level of the exercise to be inconsistent behaviors (i.e., planning and decisiveness are negatively related) may be consistent at the broader level of the AC. Notwithstanding the trait-activating effects of the AC as a whole, specific exercises offer more varied situational cues. As responses to these more specific cues might be more richly informative with regard to the behavioral propensities of individual assessees, the following discussion focuses primarily on exercise-level and cue-level situation effects in the AC.

In current AC practice, exercises are developed with two goals in mind. First, they are developed to increase fidelity for purposes of maximizing criterion-related validity (Ahmed, Payne, & Whiddett, 1997; Thornton & Mueller-Hanson, 2004). That is, exercises are developed to represent the most important demands of the target job. Notably, using exercise fidelity as a basis for inferring on-the-job performance does not require the imputation of dimensions. Second, exercises are developed to elicit job-related behavior as indicative of specific dimensions (Howard, 2008; McFarland et al., 2005). For instance, as shown by dimension-exercise matrices of operational ACs, a cooperative leaderless group discussion is typically seen as a way of activating leadership emergence and interpersonal competencies, whereas a presentation exercise with challenging questions is expected to trigger dimensions relating to emotional stability and communication.

Regardless of whether inferences from AC performance to actual job performance are based on exercises or dimensions, the heart of such inferences lies in the nature of the exercises. The key challenge with either approach to exercise development is that an AC exercise largely remains a black box (Howard, 2008). As noted by Brummel, Rupp, and Spain (2009), stimuli in AC exercises are complex, making it difficult to ascertain which specific aspects of the exercise map onto which dimensions. As exercises provide freedom and latitude to candidates to act on the situational stimuli included, the process and outcome of the exercise might also differ across candidates. Our review of prior research on AC exercises shows that little is known about how specific exercise features might influence performance.

To address this problem, we propose that the exercise-behavior linkage be examined at a more molecular level (i.e., the third level of Fig. 3).

Essentially, this implies building various stimuli in each exercise. In this respect, Brannick (2008) cogently argued "to deliberately introduce multiple dimension-relevant items or problems within the exercise and to score such items" (p. 132). Similarly, Howard (2008) posited that

> rating dimensions in the past relied too much on serendipity rather than purposeful design. Designers should construct specific stimuli to elicit the kinds of behaviors to be measured and guide assessors to their placement and relative importance ... designers must do more than create work samples; they must develop simulations that will best elicit the desired behaviors. We need to develop a much better understanding of the kinds of assessment center challenges that will bring out the behaviors associated with current and evolving positions and diverse business challenges. (p. 101)

The search for situational stimuli that evoke particular behaviors can be framed in the distinction between incidentals and radicals in item generation theory (Irvine & Kyllonen, 2002; Lievens & Sackett, 2007). Radicals are structural item features that determine item difficulty (Irvine, Dann, & Anderson, 1990). Radicals are also known as controlling factors (Dennis, Handley, Bradon, Evans, & Newstead, 2002). Conversely, incidentals are changes in surface characteristics of items that do not determine item difficulty. Incidentals are also called non-controlling or nuisance factors (Dennis et al., 2002). For instance, in a verbal reasoning test, radicals might consist of the use of negations and class membership, whereas incidentals might entail the use of different fonts or nouns. So far, AC research is silent about which specific exercise characteristics might trigger candidate behavior, and programmatic research is therefore needed to find out which exercise characteristics are "radicals" and which are "incidentals."

On a general level, trait activation theory might help to identify which exercise factors trigger and release job-relevant candidate behavior versus those that constrain and distract such behavior (Tett & Burnett, 2003). Various approaches might be used to increase the situation trait relevance of exercises. One approach would entail adapting the content of the exercise. Let's return to our example of an oral presentation with challenging questions. Efforts might be undertaken to plant situational stimuli within the exercise. Examples of stimuli to elicit behavior relevant to a dimension such as resistance to stress (a facet of the broader trait of emotional stability) might be the inclusion of a stringent time limit, sudden obstacles, or information overload. In a more systematic way, AC developers might ensure that content cues are embedded at the task, social, and organizational levels within a given exercise. By way of example, Tables 1 and 2 suggest demands (where reacting to the situational feature will be positively valued), distracters (where reacting to the situational feature will be

Table 1. Subordinate Feedback Role Play Targeting Behaviors Related to Agreeableness (A).

Cue Type	Task	Social	Organizational
Demands	Giving constructive feedback to a poorly performing subordinate	Subordinate's succorance (i.e., need to be helped)	Organizational culture described as supportive, worker-centered
Distracters	Giving constructive feedback to a manipulative subordinate[a]	Subordinate's sob story (evoking overly lenient feedback)	Organization in the health care industry (evoking assumed warm-hearted management)
Constraints	Instructions to focus on the hard performance data	Subordinate's independence (lack of desire to be helped)	Organizational culture described as fact-driven, rational, "cold"
Releasers	Performance data suggesting subordinate's empathy	Subordinate's recognition of the benefits of feedback	Memo from CEO in support of participative goal setting

[a]Rationale: Someone high on A is more prone to believing subordinate's self-serving falsehoods.

negatively valued), constraints (which are situational features that negate the impact of a trait on behavior by restricting cues for its expression), and releasers (which are discrete events that counteract constraints) related to agreeableness in a role play and emotional stability in an oral presentation, operating at each of the three noted levels.

A second way to elicit job-related behavior is through exercise instructions. In ACs, exercise instructions provide information and expectations to candidates about what behavior to show or not to show. For example, exercise instructions might be vague (e.g., "solve the problem") or more concrete (e.g., "motivate the problem subordinate"). Similarly, exercise instructions might be unidimensional (e.g., reach consensus) or multidimensional (e.g., reach consensus and make the company more profitable). To date, we know little about how such exercise instruction variations might affect the behavior demonstrated, in terms of either main effects or in interactions with underlying traits. However, as Tables 1 and 2 suggest, trait activation theory can provide some systematic guidance for how instructions could be used (as part of the set of cues in an exercise) to explicitly target certain behaviors.

Table 2. Oral Presentation Exercise Targeting Behaviors Related to Emotional Stability (ES).

Cue Type	Task	Social	Organizational
Demands	Important problem; complex, ambiguous information; limited time	Audience described as anxious, demanding, angry, pessimistic	Industry described as fast-paced, litigious
Distracters	Plausible worst-case scenarios are camouflaged (i.e., erroneously downplayed)[a]	Audience described as compliant, highly accepting of any solution[b]	Organization suffers from group-think sense of invulnerability[c]
Constraints	Boilerplate sample presentation provided	Audience described as compassionate and supportive	Organizational culture described as relaxed and non-evaluative
Releasers	Sudden change in critical information	Audience member is described as challenging a key point in the presentation	Personal memo from CEO stressing the importance of the presentation

[a]Rationale: Someone high on ES is less likely to contemplate worst-case scenarios, even plausible ones.
[b]Rationale: Someone high on ES is less likely to worry about why the audience is so accommodating.
[c]Rationale: Someone high on ES is less likely to challenge the organization's over-optimism.

When interpersonal exercises (e.g., role plays and oral presentations) are used, role player cues are a third means for eliciting job-related behavior. In current AC practice, role players are typically given a specific list of things to do and to avoid. Role players are also trained to perform realistically albeit consistently across candidates. Although these best practices have proven their usefulness over the years, a key function of trained role players consists of evoking behavior from candidates (Thornton & Mueller-Hanson, 2004). Trait activation theory can help identify which specific behaviors might be evoked by specific role player stimuli (i.e., specific statements or actions). For instance, Schollaert and Lievens (2008) conducted a study among experienced assessors to generate role player cues, defined as predetermined statements that a role player consistently mentions during the exercise to elicit behavior. For example, to arouse behavior related to interpersonal sensitivity, the role player might state that he feels bad about a candidate's decision. Similarly, role players might trigger behavior related to planning

and organizing (deeper trait of Conscientiousness) by asking how the candidate will implement his or her solution. It is important that these role player cues should *subtly* elicit assessee behavior because the situations might otherwise become too strong (see later).

Fourth, one might consider including a large number of shorter exercises (exercise "vignettes") in the AC. For example, Brannick (2008) recommends using five 6-minute role plays instead of a single 30-minute role play (e.g., with a problem subordinate) so that one obtains samples of performance on a large number of independent tasks that are each exclusively designed to elicit behavior related to a specific trait (refer also Motowidlo, Hooper, & Jackson, 2006, for the use of 1- or 2-minute role plays). As another example, one could aim to measure communication by including "speed" role plays with a boss, peers, colleagues, customers, and subordinates.

Finally, stimuli could also be presented through videotape or virtual reality. In the former approach, resembling earlier social intelligence measures (Stricker & Rock, 1990), candidates are shown short scenes and asked to react to what they saw. Brink et al. (2008) videotaped such reactions for coding by trained assessors. They found that, in such an "AC," assessors were able to make better differentiations among various dimensions. Recent applications even enable creation of avatar-based simulation exercises wherein participants take on a virtual identity and are confronted with standardized stimuli in a virtual workplace (Rupp, Gibbons, & Snyder, 2008).

Situational Strength in Assessment Center Exercises

In the preceding section, we demonstrated how trait activation theory might be used to evoke job-related behavior by planting specific trait-relevant situational stimuli (exercise content stimuli, specific exercise instructions, role player cues, use of exercise vignettes, and videotaped stimuli) in AC exercises. However, the trait relevance of situations represents only one factor determining the trait activation potential of situations (exercises). A second factor relates to the strength of the situation. It is well known that personality best predicts performance when the behavior is unconstrained. Strong situations are distinguished by unambiguous behavioral demands that constrain people's behavior masking underlying traits because everybody responds similarly to the situation (Bem & Allen, 1974; Mischel, 1973).

Trait activation theory highlights that exercises should not represent strong situations. Exercises designed with clearly defined tasks and role players with strict rules about what to say or do leave few options for assessees to express their individual propensities. For example, exercise instructions of a role play might prescribe to candidates to fire the employee (instead of leaving this option open). Such a demand creates a strong situation for observing behavior related to decision-making or problem solving (e.g., although it may still be possible to observe and rate how assessees handle this with regard to their interpersonal competence and stress tolerance). To our knowledge, no AC studies have explicitly manipulated the strength of exercise instructions and other situational cues as a factor affecting the expression of job-relevant traits in ACs.

Nonetheless, current AC practice seems to align with this aspect of trait activation theory. Exercises are often designed with a certain amount of vagueness and ambiguity so that differences in how assessees tackle the situation are more easily elicited and observed. One exception to this is when the dimensions are revealed to candidates before the AC (Kleinmann, 1993, 1997; Kleinmann et al., 1996; Kolk, Born, & van der Flier, 2003; Smith-Jentsch, 1996). In this research on transparency of AC dimensions, candidates are informed about the dimensions and the behaviors relevant to each dimension. Clearly, this could make the AC exercise a strong situation (or at least stronger than situations wherein such information is not provided), potentially masking individual differences and thereby reducing the personality loading of AC exercises (Smith-Jentsch, 2007).

Instructions and cues about effective behavior do not come only from the exercise instructions and transparent dimensions. Candidates might also get a sense of what is effective through prior experience in AC exercises (Kelbetz & Schuler, 2003), information from other candidates, or informal/formal coaching (Brannick, Michaels, & Baker, 1989; Brostoff & Meyer, 1984; Dulewicz & Fletcher, 1982; Gill, 1982; Kurecka, Austin, Johnson, & Mendoza, 1982; Mayes, Belloli, Riggio, & Aguirre, 1997; Moses & Ritchie, 1976; Petty, 1974). So far, we do not know to what extent practice and coaching transforms the AC exercises into ''strong'' exercises. Research with a trait activation lens, we suggest, may be useful here.

Social Skill as a Moderator of Trait Activation

Planting subtle situational stimuli in AC exercises does not guarantee that these stimuli will be picked up by candidates. That is, some candidates might

act on specific stimuli, whereas others might not. It is also possible that some candidates might misinterpret the cues built into the exercises.

The idea that people differ in terms of how they assess social situations (including AC exercises) and flexibly adapt their behavior on the basis of these cues has a long history in psychology (e.g., Argyle, 1969; Murray, 1938; Thorndike, 1920). In recent years, such individual differences have known a renaissance under the general term of social effectiveness constructs. According to Ferris, Perrewé, and Douglas (2002), social effectiveness is a ''broad, higher-order, umbrella term which encapsulates a number of moderately-related, yet conceptually-distinctive, manifestations of social understanding and competence'' (p. 50). These social effectiveness constructs are known under various aliases such as social skill, social competence, social deftness, social intelligence, emotional intelligence, and self-monitoring. People high on these constructs are typically able to better ''read'' interpersonal situations than others and adapt their interpersonal behavior in line with the cues gathered.

Again, prior research examining individual difference variables as moderators of trait activation is very scarce. König, Melchers, Kleinmann, Richter, and Klehe (2007) referred to social effectiveness in selection procedures as the ability to identify the criteria used. They developed a measure for assessing this ability wherein participants were asked to assess the criteria on which they think they are being evaluated in a selection procedure. Their answers are then compared to the targeted criteria used by the assessors. Prior research showed the measure was moderately related to cognitive ability (especially verbal cognitive ability; König et al., 2007), self-reported social skill (Schollaert & Lievens, 2008), and performance in ACs (Kleinmann, 1993), interviews (Melchers et al., in press), and integrity tests (König, Melchers, Kleinmann, Richter, & Klehe, 2006).

Recently, Jansen, Lievens, and Kleinmann (2009) found even more convincing evidence for the moderating role of social skill in knowing when to show specific behavior. They discovered that when candidates effectively and quickly saw through the ambiguous interpersonal exercises of ACs and adjusted their behavior accordingly, they received high ratings. In particular, Jansen et al. showed that agreeableness (as measured by a personality inventory) was related to cooperative behavior in AC exercises only among people who perceived correctly what kind of behavior the situation demanded. Similar results were obtained for the relationship between participants' standing on conscientiousness and their rating on planning and organizing. These results shed a different light on the relationship between personality and performance in AC exercises. In line

with trait activation theory, they show that when people adequately perceive the situation, they are able to better translate their trait (e.g., agreeableness) into effective performance on a conceptually related dimension (e.g., as measured by AC ratings on cooperation). Conversely, inadequate situation perception seems to lead to significantly worse performance. Jansen et al.'s (2009) findings are an example of how AC behavior is a complex function of personality, ability, and situational factors. Further research is needed to better understand which individual difference variables moderate trait activation of behavior among candidates. Even more importantly, future studies should examine the impact of these individual difference variables on cross-situational (in)consistency in AC exercises.

Summary

In the preceding section, some key changes in AC exercise design approaches were proposed. We emphasized that AC exercises should not only be designed to represent job-related tasks but should also deliberately, albeit subtly, be structured to evoke relevant categories of behavior. For clarity, two reminders are in order. First, we are not proposing that current best practices of exercise development should be abandoned. Rather, we posit that trait activation theory should play a more prominent role in such development. Whereas current practices typically simulate key task, social, and organizational demands of the job, we see untapped potential in planting multiple stimuli within exercises as a systematic and structured means of increasing the frequency and variability of job-related behavior in AC exercises.

Second, we emphasize that our suggestions of carefully building multiple situational stimuli in exercises are not meant to "fix" the AC to obtain "better" convergent and discriminant validities (Lievens, 2008). Neither do these recommendations imply that it is personality traits per se that should be targeted for assessment in exercises. Rather, we seek to better control the expression of job-relevant *behavior*. Regardless of how that behavior is then captured and evaluated by assessors (in task-based models, dimension-based models, etc.; see later section), eliciting and observing behavior is key to effective assessment and development centers. As Howard (2008) noted "Behaviors, not exercises, are the currency of assessment centers. Exercises are merely the stimuli to elicit behaviors" (p. 101). Trait activation theory offers a much-needed framework for understanding and managing the conditions under which AC exercises elicit job-relevant behavior.

BEHAVIOR OBSERVATION AND EVALUATION IN ASSESSMENT CENTERS

Trait activation focuses on how candidate behavior might be elicited by planting various situational cues into AC exercises. As noted earlier, the trait activation logic aims to increase the range of job-related behaviors shown by assessees, which ultimately should lead to better predictions and higher quality feedback. However, as shown in Fig. 2, trait activation represents only one part of the equation. The other part of the theory is behavioral observation and evaluation on the part of *assessors*. Indeed, even though a variety of job-related behavior might be elicited, it is not guaranteed that assessors would observe and rate it. In particular, assessors might miss elicited behavior, categorize it as relevant to different traits, or even judge the appropriateness of the behavior differently. In other words, trait-expressive behavior of candidates can be washed out by judgments of assessors (Tett & Burnett, 2003).

So far, we know very little about the models and schemas that assessors use to interpret candidate behavior and to make trait judgments (Lievens & Klimoski, 2001; Zedeck, 1986). As a notable exception, Lance et al. (2004a, 2004b) compared different models of assessor cognitive processes and discovered that assessors mainly used an exercise-specific model for judging candidate behavior. Prior AC research has also examined various design considerations and rating aids to increase the quality of assessor evaluations (for detailed reviews, refer Lievens, 1998; Sackett & Tuzinski, 2001; Woehr & Arthur, 2003). Although some of these design considerations were found to be effective, they were not deliberately constructed to facilitate evaluation per the prescriptions of trait activation theory. In the following sections, we review four design considerations implied by the evaluation aspect of trait activation theory, specifically, regarding the selection of dimensions, observation methods, type of assessor, and type of assessor training.

Dimension Selection

Regarding the use of dimensions in ACs, it is important to emphasize that trait activation theory does *not* require that assessors should directly rate traits. A trait can be both positive and negative, whereas a performance dimension is inherently valued. That is, a performance dimension represents valued behavior. Traits are also situated at a deeper level, whereas performance dimensions are more at the surface level. Organizations choose

performance dimensions for various reasons, only one of which is their representation of traits. An important advantage of AC dimensions is that they are often formulated in the language of work behavior, increasing their apparent relevance to management. In fact, dimensions capture acquired work skills (e.g., negotiation and organization skills) and are closely linked to job activities and organizations' competency models (Lievens, Sanchez, & De Corte, 2004).

That said, trait activation theory does offer specific predictions about what should be expected when dimensions are included in ACs. First, it suggests that job analysis, although important, is insufficient for determining dimensions to be included in ACs. Once job analysis has identified the dimensions to be measured, trait activation theory might be used to link dimensions within an exercise to a given underlying trait (e.g., ''innovation'' and ''adaptability'' are based on behaviors that might be expressions of Openness). Thus, subject matter experts might be asked to rate the extent to which each dimension identified by job analysis is related to each of the Big Five traits (e.g., see questionnaires developed by Haaland & Christiansen, 2002; Lievens et al., 2006).

Second, on a more general level, trait activation theory highlights that care should be taken when deciding on the dimensions to be included in the AC exercises. Along these lines, 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). In particular, trait activation theory advocates using specific dimensions instead of general concepts (Tett & Schleicher, 2001). At a practical level, use of specific dimensions (short-term planning and strategic planning instead of the broad dimension planning) allows precise diagnosis for developmental purposes (Thornton, 1992) and more points of comparison in matching individuals to work environments (Tett & Guterman, 2000).

Observation Methods

In the original AT&T AC, assessors took notes while observing candidates and afterward used this information to rate the candidates. Through the years, several alternatives have been suggested to improve the quality of ratings. Behavioral checklists constitute one of the most popular options

(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.

So far, the research evidence with regard to the effectiveness of behavioral checklists is mixed. In one of the first studies, Reilly et al. (1990) reported positive findings because ratings made through behavioral checklists demonstrated higher convergent and somewhat higher discriminant validity than ratings without the use of behavioral checklists. In other studies, behavioral checklists enhanced only discriminant validity (Donahue, Truxillo, Cornwell, & Gerrity, 1997) or had virtually no effects (Schneider & Schmitt, 1992). Yet another set of studies examined more specific aspects related to behavioral checklists. Most of this research has been guided by limited cognitive capacity models. For example, Reilly et al. (1990) determined that the optimal number of statements per dimension varied between 6 and 12. 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 criterion-related and construct-related validity.

Apart from limited cognitive capacity models, trait activation theory might also be informative with regard to behavioral checklist design. As noted earlier, trait activation theory distinguishes trait-expressive work behavior from (job) performance because performance is "valued" work behavior (see Figs. 1 and 2). However, in current practice, AC dimensions are typically defined and rated more along the lines of performance than behavior because the highest ratings imply the best performance, not necessarily the highest levels of the trait. In other words, the value of the behavior is already captured by the rating scale, rather than measuring where the behavior falls along the trait dimension and later determining its optimal level for performance. Rating scales in ACs warrant reconsideration in terms of trait activation, such that the distinction between the observation of behavior and the evaluation of it is apparent in the rating checklists used. For example, sociable behavior might be equally salient in two exercises, yet positively valued in one exercise and negatively valued in the other.

Second, trait activation theory suggests developing scoring checklists that include behavioral clusters. That is, behaviors that are different on the surface might be clustered when they share an underlying source.

Initial evidence from Binning, Adorno, and Kroeck (1997) supports this recommendation. They found that the discriminant validity of behavioral checklists increased only when the items were ordered in naturally occurring clusters (e.g., according to stages in the exercises: start, middle, or end). The discriminant validity of a randomly ordered checklist was low.

Third, a different approach might consist of including the situational cues that were designed to elicit candidate behavior (e.g., the role player statements) in the behavioral checklists themselves. Accordingly, assessors would be reminded and prompted by the situational cues when attending to candidate behavior. It then helps them to "see the forest for the trees" in the complex stimuli triggered by AC exercises. Brannick (2008) refers to this approach as aligning the stimulus content of the exercises with the scoring rubric.

Assessor Selection

Various prior studies have examined whether the type of assessor affects the quality of AC ratings. In this stream of studies, psychologist assessors were compared to manager assessors. 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 (2001a, 2001b) found that managers had more difficulty discriminating among dimensions than did psychology student assessors. However, managerial assessors also rated candidates with higher accuracy. Other studies found that psychologists outperformed non-psychologists only when the criterion-related validity of the interpersonal ratings made was examined ($r = .24$ vs. $r = .09$; Damitz, Manzey, Kleinmann, & Severin, 2003) and that experienced assessors yielded significantly higher accuracy than inexperienced assessors (Kolk, Born, Van Der Flier, & Olman, 2002). As a whole, these studies reveal that both types of assessors have their strengths and weaknesses, in support of the common practice of including a mix of experienced line managers and psychologists in the assessor team.

Most of the aforementioned studies ascribed these findings to the educational background (e.g., degree in psychology) and expertise of assessors. From a theoretical and practical perspective, however, it is important to define and operationalize assessor expertise, and trait activation theory might provide a means to do that. In particular, Christiansen,

Wolcott-Burnam, Janovics, Burns, and Quirk (2005) demonstrated the importance of knowing how personality traits are revealed in behavior and activated by situations to become a "good judge." They referred to such declarative knowledge structures as "dispositional reasoning." On the basis of theoretical and empirical research, Christiansen et al. developed a multiple-choice measure that consisted of two parts. One part measured people's knowledge of the Big Five ("knowledge of trait-behavior linkages" and "proficiency at trait concepts"). The other part was based on trait activation theory and measured whether people knew which situations might activate specific traits and which traits might be activated by specific situations ("understanding of situation-trait relationships"). In a simulated interview context, dispositional reasoning was related to general mental ability and openness to experience and emerged as the single best predictor of accuracy. This approach might have practical value in the selection of assessors. Therefore, future research is needed to test the dispositional reasoning of various types of assessors (experienced vs. inexperienced; psychologists vs. managers) in an AC context and link them to key outcomes such as construct-related and criterion-related validity.

Assessor Training

A final group of studies has concentrated on the type of training provided to assessors. There seems to be some evidence that especially frame-of-reference training might be beneficial in terms of increasing inter-rater reliability, dimension differentiation, differential accuracy, and even criterion-related validity (Goodstone & Lopez, 2001; Lievens, 2001b; Schleicher, Day, Mayes, & Riggio, 2002). Frame-of-reference training teaches assessors to use a specific performance theory as a mental schema to "scan" the behavioral stream for relevant incidents and to place these incidents – as they are observed – in performance dimensions.

The beneficial effects of frame-of-reference training are explainable in trait activation terms. Specifically, the clear distinction in trait activation theory between observation and evaluation suggests that assessor training should provide assessors not only with information about the behaviors to be observed but also the evaluation of those behaviors. This is exactly what frame-of-reference training aims to accomplish (Lievens, 2001b; Schleicher et al., 2002). Specifically, a performance theory is imposed on assessors to ensure that they rate candidates in accordance with the norms and values of a specific organization. This performance theory consists of

competency-relevant behaviors and their effectiveness levels. Accordingly, trait activation theory provides a theory-based underpinning for the importance of providing frame-of-reference training to AC assessors.

Another implication of trait activation theory is that assessors should be familiarized with the situations that activate the behaviors. In current assessor training practice, the focus is placed on the dimensions and the accompanying behaviors. However, it is equally important that assessors know when specific behavior is being activated by various situational stimuli. We are not aware of studies that have examined such a more comprehensive assessor training approach.

Summary

Contrary to the scarce research on trait activation, various prior AC studies have examined factors that influence how assessors evaluate assessees. However, most of this body of research was not driven by an over-arching theoretical framework. We have illustrated how trait activation theory might be fruitfully used to interpret extant research and guide future studies on dimension selection, observation methods, assessor selection, and assessor training.

GENERAL DIRECTIONS FOR FUTURE RESEARCH

Frequency and Variability of Candidate Behavior in Assessment Center Exercises

It follows from trait activation theory that the inclusion of standardized situational stimuli in AC exercises will affect the amount and variability of candidate behavior to be observed in exercises, as exercises may be designed explicitly to increase their situation trait relevance. Although no studies have examined the effects of AC exercises deliberately designed to evoke candidate behavior, empirical research has underscored the importance of the observability of behavior. Prior research shows that there exist wide variations in the opportunity to display dimension-related behaviors across exercises (Donahue et al., 1997; Reilly et al., 1990). For instance, in Reilly et al. (1990), the number of behaviors varied from 4 for one dimension to 32 for another dimension. Furthermore, 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 a certain point (i.e., 12 items), beyond which the relationship remained stable.

It is important to note that generating more job-related behaviors in an exercise is a crucial outcome in developmental ACs (e.g., for leadership development), because these behaviors serve as a basis for providing participants with detailed developmental feedback about their strengths and weaknesses. Thus, future research should examine whether the frequency and variability of candidate behavior will be higher in AC exercises that are specifically designed to evoke job-related behavior.

Assessment Centers as Measures of Ability and Motivation

As noted earlier, the application of trait activation theory to ACs aims to increase the amount of job-related behavior shown by candidates by enhancing the trait relevance of the exercises. To this end, various approaches (content cues, exercise instructions, role player prompts, etc.) were proposed. At the same time, however, we warned that AC exercises should not become so strong as to demand that all assessees respond the same way. This risk is especially present in the form of exercise instructions. Use of clear-cut instructions might make the AC exercises too strong for measuring specific dimensions (see our earlier example of a termination exercise), clouding the possibility to observe variability in behavior related to specific dimensions.

As noted by Thornton and Rupp (2005), conceptualizing AC exercises as either ''strong'' or ''weak'' situations influences whether they are measures of motivation, ability, or both. When AC exercises include relatively vague instructions, and are therefore weak situations, candidates have the freedom to take action or to refrain from action. The example given by Thornton and Rupp is one of a candidate who makes hardly any verbal interventions in a leaderless group discussion. Assessors might score this particular candidate low on initiative. However, it is not possible for them to assess the verbal communication ability of that candidate as no behavioral instances related to this dimension occurred. Compare this to a leaderless group discussion that includes clear instructions (invoking a strong situation) requiring each candidate to defend his/her initial choices. Here, verbal communication ability can be assessed, although assessors do not receive information about candidates' motivation (initiative) to take such actions spontaneously. This

line of reasoning also suggests that the dimensions themselves (or at least dimensions within specific exercises) could be classified according to ability, motivation, or both.

Thus, when a weak situation is invoked (i.e., vague instructions are given to candidates in AC exercises), applicants' motivation to act (or not to act) can be assessed. Candidates will then typically use the traits they possess to guide their actions and show behaviors to the best of their ability (Smith-Jentsch, 2007). Therefore, AC exercises become confounded measures of ability and motivation. Conversely, when a strong situation is invoked (when highly directive instructions are provided and the stakes are high), ACs do not assess candidates' motivation to act or not. As the motivation to act is no longer an explanation of candidate behavior, it is possible to make "purer" judgments about candidates' abilities.

Research is needed to compare AC exercises with different instruction formats. Trait activation theory would predict that AC exercises with highly directive instructions (strong situations) will yield higher mean performance ratings and less variability in those ratings across assessees than will AC exercises with less directive instructions (weak situations), because strong AC exercises serve as maximum performance measures. This effect will be stronger on the more motivational dimensions (e.g., persuasiveness; thereby undermining their validity as motivational dimensions), as skill dimensions (e.g., problem solving) are best assessed under directive instructions.

In light of these comparisons, it is also of paramount importance to examine the *g* loading versus personality loading of AC exercises within which the exercise instructions have been manipulated. Are strong exercises more relevant to ability and maximum performance measures? Do weak exercises have a higher personality loading and a higher convergence with self-reported personality scale scores? Finally, AC exercises with different instruction formats might also have differing effects on criterion-related validity because of their presumed differential ability versus personality loading. On a more general level, a problem inherent in conducting this research is that we do not know where vague instructions end and strong instructions start. In addition, as noted earlier, instructions might be too strong for observing behavior related to some dimensions, while still allowing behavior related to other dimensions to be observed. Thus, an additional research need is to establish the correspondence between points on the strong-weak situation continuum and specific exercise instructional sets.

The stream of research on the influence of transparency on AC ratings provides some preliminary clues in answering these questions (although it should be stressed that the use of transparent vs. nontransparent dimensions

is not the same as the inclusion of different exercise instruction formats). First, candidates were able to obtain higher ratings in ACs in transparent conditions as compared to nontransparent conditions, illustrating that such transparent ACs seem to act like maximum performance measures. However, this mean effect was found only when one dimension was made transparent (Smith-Jentsch, 2007) and was not present when multiple dimensions were divulged (e.g., Kolk et al., 2003). Second, in nontransparent ACs, the convergence between ratings (on assertiveness) in an AC exercise and typical performance measures (self-reported measure of assertiveness) was much higher than in transparent ACs (Smith-Jentsch, 2007). In such transparent ACs, the convergence of AC ratings with maximum performance measures (verbal ability measure) was higher (e.g., Kolk et al., 2003). Third, use of transparent ACs resulted in lower criterion-related validity (Kleinmann, 1997; Smith-Jentsch, 1996).

These issues also have implications for viewing AC exercises as maximum performance measures, which is the typical assumption in selection and promotion contexts (Marcus, Goffin, Johnston, & Rothstein, 2007; Ployhart, Lim, & Chan, 2001; refer also Atkins & Wood, 2002; Hagan, Konopaske, Bernardin, & Tyler, 2006). Ployhart et al. (2001) argued that AC exercises tap maximum performance because the three requirements of maximum performance measures are satisfied: (1) short time frame, (2) candidates are aware they are being evaluated, and (3) candidates are motivated to put their best foot forward. The notion of candidates being able to identify the appropriate behavior in a situation (either by accurately perceiving the situation or by the use of transparent AC dimensions and behaviors) also seems consistent with testing maximal performance. When candidates are not able to perceive which dimensions and accompanying behaviors are relevant, candidates often misread the situation and simply guess which behavior is appropriate. For instance, Kleinmann (1993) showed that a substantial number of applicants did not recognize which dimensions were measured. Accordingly, they might misdirect their effort toward the wrong dimensions (Smith-Jentsch, 2007). Clearly, this practice is not ideal to show maximum performance. It is similar to a hundred meter runner who runs very fast in the wrong direction.

Construction of Alternate Forms of Assessment Center Exercises

Trait activation theory offers guidance in the development of alternate forms of AC exercises. To construct alternate forms of AC exercises,

Brummel et al. (2009) recommended changing the surface features of an AC exercise, while keeping the deep structure of the exercise intact. However, at the same time, they acknowledged that "distinguishing what constitutes surface and deep structure in a simulation exercise is difficult because the stimuli are complex." Their results showed that it was more difficult to develop alternate versions of role plays or leaderless group discussions as compared to oral presentations. These results can easily be framed in our prior discussion of the benefits of imposing a trait activation structure on AC exercises.

Trait activation cues might guide the determination of the deeper structural aspects (the so-called radicals, using our earlier analogy from item generation theory) of an AC exercise, in terms of providing a template of what aspects of the exercise map onto which dimensions and should be kept constant across exercises. For example, one might develop three role player cues to evoke behavior related to the dimension of interpersonal sensitivity in a series of role plays. Superficial differences among the cues would be incidental to their deeper similarities (i.e., as radicals) in targeting the same trait. Given that the exercises might look superficially different, this approach might serve as a deterrent to coaching or practice effects. By the same token, it might also offer cues for abilities serving detection of radicals. Thus, candidates who recognize the commonalities in the radicals, operating below the surface of less uniform incidentals, may have the advantage in presenting consistently favorable performance.

Applicant Perceptions of Assessment Center Exercises

Generally, AC exercises are favorably regarded by candidates due to their job-relatedness and face validity (Hausknecht, Day, & Thomas, 2004). How might these typically positive perceptions be affected by the increased situational relevance of AC exercises through the use of structured situational stimuli to evoke behavior?

First, candidates might appreciate the fact that these cues are consistently used across candidates. Given that multiple situational stimuli are built into AC exercises, candidates might also feel that they have more opportunity to show job-related behavior. Therefore, the consistent inclusion of various situational cues in AC exercises might lead to increases in applicants' perceptions of structural procedural justice dimensions such as consistency and opportunity to perform.

Second, on the downside, the use of situational stimuli might reduce the realism and interpersonal "warmth" of AC exercises to the degree it detracts from the natural flow of the exercise. Clearly, this will depend on the kind of situational stimuli involved. If subtle content or role player cues are included, these aspects of the exercise will probably not represent that noticeable of a change to assessees. Conversely, when a series of 2-minute role plays are being employed for eliciting more candidate behaviors, the change might be perceived as much more drastic by (especially experienced) assessees. Prior research has shown that increased interview structure leads to less favorable candidate reactions (Conway & Peneno, 1999). Therefore, the use of *some* situational stimuli to elicit behavior might lead to decreases in applicants' perceptions of interactional procedural justice dimensions such as two-way communication and interpersonal warmth.

Finally, the use of multiple situational stimuli might also affect the informational dimensions of procedural justice. For example, if more detailed exercise instructions are used as a way of triggering more behavior (while avoiding making the situation too strong), applicants might have more positive perceptions of the pre-assessment information provided to them. Similarly, feedback based on more behavioral observations might garner more positive perceptions of the post-assessment information (see section on AC feedback).

Reliability of Assessment Center Ratings

A recent meta-analysis (Connelly & Ones, 2008) on the inter-rater reliability of assessor ratings showed that reliability was adequate for various types of AC ratings. However, it was lowest for so-called within-exercise dimension ratings (ratings made on one dimension within a specific exercise). This is understandable because such ratings are often based on rather limited behavioral evidence. AC exercises designed to better elicit job-related behavior through the prescriptions of trait activation theory might further increase assessors' inter-rater reliability. This argument is based on two rationales. First, planting similar situational cues in AC exercises across candidates should increase the standardization, structure, and consistency in those exercises. Second, the opportunity to observe and take notes on dimension-related behavior should also increase the reliability of the ratings made in light of 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. Just as the reliability and content representation

of a self-report test increases with the addition of items from the same domain, assessing a given dimension in an AC exercise will improve with the addition of dimension-specific cues.

Interestingly, the use of standardized situational cues in ACs can be compared to the use of standardized questions among interviewers. Given that prior research has shown that the inter-rater reliability of structured interviews is higher than that of unstructured interviews (Conway, Jako, & Goodman, 1995), we expect the same effect when standardized situational stimuli for arousing relevant candidate behavior are employed. Future research is needed to confirm whether the specific improvements to AC exercises suggested here do in fact result in enhanced inter-rater reliability.

Internal Construct-Related Validity of Assessment Center Ratings

Trait activation provides a deeper and more sophisticated theory-based approach to looking at the convergence of ratings across AC exercises. An advantage of using trait activation theory is that it allows for the fact that convergence would not be expected among all ratings of the same dimensions across exercises. In fact, trait activation posits that convergence should be expected only between exercises that provide a sufficient opportunity to observe behavior related to the same underlying trait, and trait-expressive behavior is similarly valued across exercises. For example, consider ratings on the dimension of interpersonal influence, based on behavioral expressions of the Big Five trait of Extraversion. As both a leaderless group discussion and a role play exercise can be expected to provide cues relevant to this trait and place similar value on trait expression (i.e., they have similar demands for Extraversion), convergence between ratings on a dimension such as interpersonal influence should be expected. A planning exercise, on the contrary, provides far fewer cues for expression of Extraversion and could conceivably place a negative value on its expression (e.g., as a distraction to task completion). Accordingly, ratings on interpersonal influence from this exercise should not be expected to correlate notably with those from the first two exercises. Thus, when AC exercises differ in their trait activation potential for a given trait and place different demands on trait-relevant behavior, cross-exercise behavioral consistency will be low and convergence poor for ratings of a trait-expressive dimension.

Furthermore, the greater psychological depth of trait activation is illustrated by the fact that convergence is also expected across exercises

that look superficially different but activate the same traits at a deeper trait level. For instance, take an exercise that requires risk-taking behavior to successfully resolve the situation and a second one that involves persuading a group of people to adopt the candidates' position. Given that both these behaviors can be seen as falling within the construct domain of Extraversion, convergence on ratings from a dimension linked to this Big Five trait could be expected across these exercises (Haaland & Christiansen, 2002). Notably, only by recognizing the similarity in trait relevance at the deeper, "radical," level is such an expectation warranted.

Haaland and Christiansen (2002) provided empirical support for these implications of trait activation theory. They examined whether poor convergence of AC ratings across exercises was due to correlating ratings from exercises that differed in trait activation potential. Subject matter experts were asked to judge whether it could be possible to observe behavior relevant to the Big Five traits in a given exercise. The subject matter experts were then instructed to link the dimensions of the AC with the Big Five traits: greater convergence should be expected only on dimensions conceptually relevant to a given trait. The correlations between ratings from exercises high in trait activation potential were stronger than the correlations between ratings from exercises low in trait activation potential, providing support for the implication that the trait activation potential of the exercises plays a role in the convergent validity of ratings. Lievens et al. (2006) found similar results across a large number of ACs. The size of the effects, however, was small.

A drawback of these two prior studies is that they evaluated AC exercises in an existing operational AC, that is, without manipulation of trait activation potential in different exercises. Stronger effects might be expected under more deliberately controlled (yet realistic) conditions. Thus, so far, there have been no tests of the actual implementation of trait activation theory in AC exercise design for increasing the convergence of AC ratings across exercises. Key to such efforts will be separation of trait expression, per se, from the evaluation of that expression in light of exercise demands.

External Construct-Related Validity of Assessment Center Ratings

External validation research on ACs might also benefit from taking trait activation theory into account. In external validation, AC scores are linked in a nomological network to other instruments such as personality inventories, 360 degree feedback ratings, or cognitive ability tests. As argued

by Tett and Burnett (2003), trait activation is a framework that applies to many measurement methods, including ACs, personality inventories, 360 degree feedback inventories, or structured interviews. Therefore, trait activation theory might also provide a novel way to interpret research that correlates AC ratings with other assessment instruments. In particular, an intriguing avenue for future studies is to incorporate trait activation ideas when externally validating AC ratings with those from non-AC methods with similar activation potential.

The value of this idea can be indirectly tested by reinterpreting results of prior AC external validation research. Given that these studies did not rely on trait activation theory, it is striking that specific personality traits correlate with specific AC exercises. For instance, Spector, Schneider, Vance, and Hezlett (2000) discovered that performance in "interpersonal" exercises correlated with self-ratings on emotional stability, extraversion, and openness, whereas performance in "problem-solving" exercises correlated with cognitive ability and self-ratings on conscientiousness. In another study, Craik et al. (2002) reported that in-basket performance was related to Conscientiousness, Openness, and strategic dimensions such as decision-making. Conversely, group discussion performance was best described by interpersonal dimensions and personality constructs such as agreeableness, extraversion, and openness.

Similar a priori hypotheses have been tested about relations between AC exercises and cognitive ability. Goldstein et al. (1998) reported that the relationship between ACs and cognitive ability tests varied as a function of the cognitive "loading" of AC exercises. When exercises (e.g., in-basket exercise) tapped cognitively oriented dimensions (e.g., problem analysis), there were stronger relationships between the exercise and the cognitive ability test. Similarly, Thornton, Tziner, Dahan, Clevenger, and Meir (1997) discovered that the correlations of AC ratings with dimensions measured by comparable cognitive ability tests were higher than the correlations with dimensions measured by non-comparable cognitive ability tests. For example, AC ratings on routine problem solving correlated on average higher with tests of general intelligence, creativity, logic, and mechanical ability than with tests of spatial perception, accuracy of perception, writing ability, oral ability, and graphical ability.

Haaland and Christiansen (2002) provided more direct support for the idea of considering trait activation potential in external validation research. They asked subject matter experts to evaluate AC exercises on their trait activation potential. Higher correlations were obtained between 16PF personality scores and exercises judged high in trait activation potential for

a given personality trait than correlations with exercises low in trait activation potential.

Besides looking at personality trait inventories, another research suggestion consists of studying the relations between AC ratings and 360 degree feedback ratings. Prior studies (Atkins & Wood, 2002; Hagan et al., 2006) that validated a 360 degree feedback program against an AC found high correlations between the overall assessment rating and the aggregated 360 degree ratings. Unfortunately, no analyses at the level of the dimension ratings were conducted. Future studies might employ trait activation theory to make more fine-grained predictions and to examine under which conditions both procedures yield convergent results. For example, trait activation theory suggests that ratings of interpersonal sensitivity in AC exercises that are high in trait activation potential for agreeableness might correlate higher with peer ratings on interpersonal sensitivity in 360 degree feedback than with supervisor ratings of interpersonal sensitivity. The rationale is that peers might provide better insight into these interpersonal aspects because they have better opportunity to observe behavior related to the trait of agreeableness, themselves offering cues for its expression.

In sum, prior research has externally validated AC ratings without paying attention to trait activation theory. Trait activation theory presents a more sophisticated strategy for such validation as it consists of deliberately mapping the trait activation potential of the AC exercises and the trait activation potential of external measures, including personality inventories, 360 degree feedback ratings, situational judgment tests, situational interviews, and behavior description interviews. This theory-driven validation strategy also holds the constructs constant, while varying the assessment methods (Arthur & Villado, 2008), promoting construct-oriented validation efforts.

Criterion-Related Validity of Assessment Center Ratings

Prior criterion-related validity research has shown that ACs are good predictors of job performance and potential, at the level of both the overall assessment rating and the final dimension ratings (Arthur et al., 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hermelin, Lievens, & Robertson, 2007). In addition, a recent meta-analysis has documented the incremental validity of AC ratings over and above cognitive ability and personality (Meriac, Hoffman, Woehr, & Fleisher, 2008).

The behavior consistency model (Wernimont & Campbell, 1968) is often advanced as one of the key explanations behind the criterion-related validity of ACs. This model posits that the precision in predicting future performance improves when the point-to-point correspondence between predictor and criterion measures is increased. Trait activation theory presents a theoretical basis for understanding how such correspondence affects the validity of work samples and ACs (Fig. 4).

Fig. 4 clarifies that the criterion validity of ACs depends on a number of key conditions. First, the situational cues presented in ACs must be similar to those presented in the actual job. This is the essence of work sample methodology: to present realistic job demands in a controlled environment. As noted earlier, the AC as a whole is an evaluative situation that is competitive when used for selection and promotion and more awareness-generating when used for development. Job performance is clearly evaluative, but the actual job is undertaken over much longer time intervals than is an AC and under more dynamic conditions. It bears consideration that traits underlying long-term, typical performance on the job may be different from those serving short-term maximal performance in ACs (e.g., a low work ethic evident in poor job performance assessed over months and years may be masked in a one-day AC). Exercise-level demands offer the greatest opportunities for control, and research targeting specific exercise features and their effects on behavioral elicitation is sorely needed.

A second, related condition affecting AC criterion validity is the relative balance of intrinsic and extrinsic rewards driving behavior in ACs versus real job settings. Trait activation per se (i.e., expressing one's traits in responses to trait-relevant situational cues) is intrinsically rewarding: it feels good to express oneself (on most traits; Bakan, 1966; Cote & Moskowitz, 1998; Murray, 1938; Wiggins & Trapnell, 1996). This process operates only in weak situations, that is, where extrinsic rewards are not so strong as to override trait tendencies. ACs, especially those used for selection and promotion, are stronger than most real job situations. Accordingly, AC criterion validity will suffer to the degree that trait dispositions driving performance on the job are overpowered by situational demands driving performance in the AC. The challenge facing AC developers in this context is to design exercises offering subtle opportunities for the expression of job-relevant traits.

A third factor affecting AC criterion validity, as portrayed in Fig. 4, is the similarity between the AC and the actual job in the way that work behavior is evaluated as performance. The measurement of AC performance is typically undertaken with considerable care and deliberation under controlled

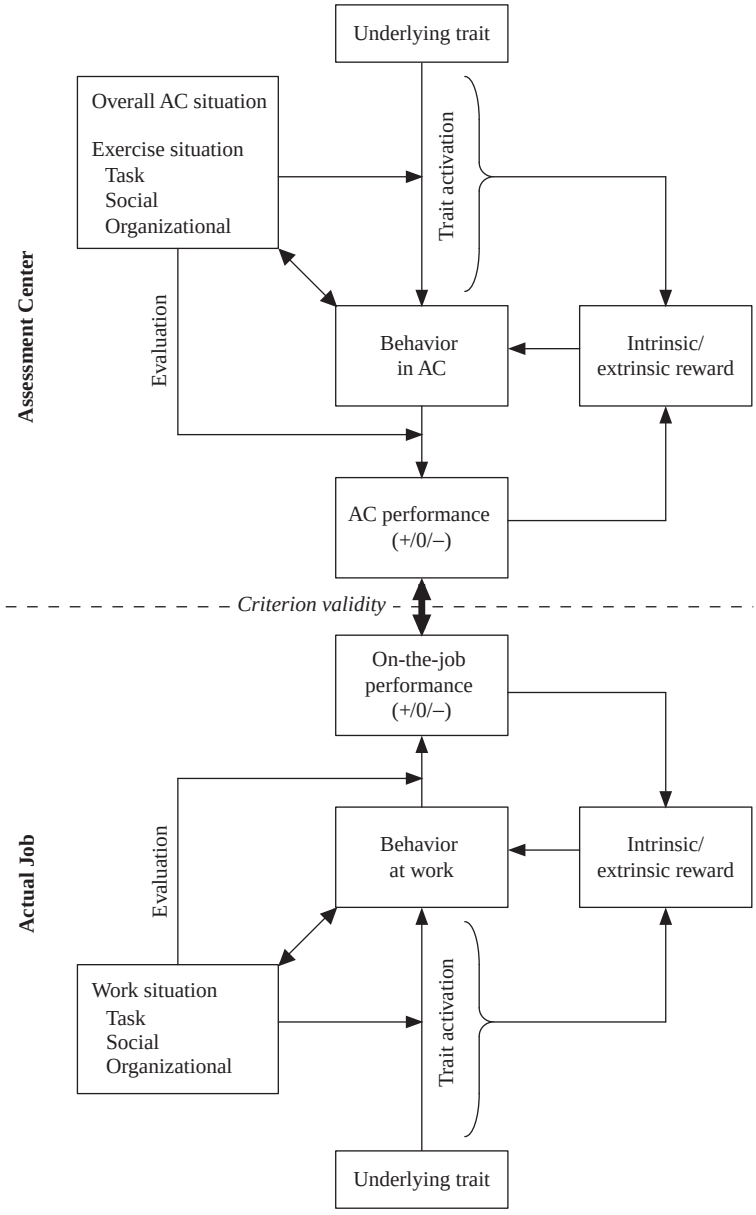

Fig. 4. Schematic Overview of Criterion-Related Validity of ACs on the Basis of Trait Activation Theory.

conditions, guided by explicit checklists and following extensive assessor training. Job performance, on the contrary, is usually assessed by untrained judges under "noisy" conditions typified by varying job demands, rater opportunity to observe, and rater goals. Evaluation in ACs is likely to focus more on targeted task demands; evaluation in real job settings, on the contrary, is likely to be biased by performance judged in relation to less structured social and organizational demands. Thus, a boss may overrate a subordinate's job performance to the degree that the employee meets the boss' own social needs. Such biases are explicitly "trained out" in ACs.

In sum, trait activation theory posits that AC criterion validity may be understood in terms of the similarity between the AC and target job on (a) relevant traits operating at multiple levels (task, social, organizational); (b) the relative balance of intrinsic versus extrinsic reward systems, bearing on situation strength and maximal versus typical performance; and (c) methods and processes of evaluating behavior as performance. Higher similarity in each case can be expected to confer higher criterion validity. Whether the impact of such factors might be assessed (as moderators) using meta-analysis is doubtful, as it would require that AC descriptions in prior research be detailed enough to allow reliable distinctions to be drawn among ACs, and such detail is generally lacking. Rather, research is more likely to be fruitful if based on a more proactive, confirmatory strategy explicitly targeting the noted factors. Better understanding of the conditions affecting AC criterion-related validity can be expected to serve improvements in selection and promotion decisions based on AC performance.

Effects of Feedback in Assessment Centers

In ACs (and especially in developmental ACs), the ratings and observations provided by assessors are merely the means to an end, providing a detailed and valid portrayal of the assessee's strengths and weaknesses. Subsequently, these ratings serve as the basis for developmental feedback, training activities, and action plans. The implications of trait activation theory for understanding reactions to AC feedback as well as performance change upon receiving feedback might be twofold.

First, trait activation theory can inform the debate about whether AC feedback reports should be built around dimensions versus exercises (Thornton, Larsh, Layer, & Kaman, 1999). When feedback is built around dimensions (e.g., "You scored low on resilience"), the advantage is that dimension-specific feedback is relevant across a wide variety of situations.

Yet, this feedback assumes that these dimensions are indeed measured across many situations (exercises). As noted earlier, research has shown this is usually not the case. If the dimensions are not valid indicators of the managerial abilities, the developmental feedback and action plans could be misguided and detrimental. Conversely, feedback might also be built around exercises (e.g., ''You scored low in the oral presentation''). As mentioned, such exercise-based or task-based feedback is in line with most of the research evidence showing that exercises capture the lion's share of variance in AC ratings. Yet, this feedback might lack depth, as it generalizes to only one specific situation (one exercise/task). As noted by Howard (2008), specific tasks frequently change for many jobs such that it seems not very insightful to know that a participant masters a specific task. Thornton et al. (1999) examined candidate reactions toward dimension-based feedback versus exercise-based feedback. Results indicated favorable reactions to both feedback types and no real differences in their perceived accuracy and usefulness.

Notably, trait activation theory promotes a combined position involving both dimension-based and exercise-based feedback, suggesting that feedback reports be built around the situations that activate job-relevant traits (e.g., ''You scored low on resilience in situations where you were put under pressure''). The link with job settings might then be explicated. So far, applicant reactions toward feedback based on trait activation principles have not been explored. Similarly, no research has examined whether feedback based on trait activation principles might lead to performance improvement. Such research seems warranted.

Second, trait activation theory has implications for the quality of feedback provided. AC exercises designed to elicit a broader array of job-related behaviors might be expected to generate feedback based on richer behavioral material. In turn, this might lead to more favorable reactions to the feedback. For instance, prior research (Burd & Ryan, 1993; Harris, Paese, & Greising, 1999; Kudisch & Ladd, 1997) has shown that both feedback process factors (e.g., assessor expertise) and exercise factors (e.g., perceived content validity of the exercise) are related to favorable feedback reactions such as the perceived utility of the feedback. Therefore, when AC exercises are deliberately designed to activate more job-related behaviors, candidates might perceive the exercise as offering more useful feedback.

A key remaining question then is whether participants will actually act upon this improved developmental feedback and engage in subsequent developmental activities. The general feedback literature shows that feedback is not always effective (Kluger & DeNisi, 1996). Similar results

are obtained in the AC field, although research is again scarce. For instance, Jones and Whitmore (1995) pointed out the lack of differences in career advancement between managers who went through a developmental AC and a non-AC 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). However, recent research paints a more positive picture. Woo, Sims, Rupp, and Gibbons (2008) found that more favorable feedback was related to higher behavioral engagement during a developmental AC and in the subsequent follow-up activities.

CONCLUSIONS

After decades of research, it is now generally acknowledged that ACs are at a crossroads. Research has converged in suggesting that exercises are predominant in terms of explaining variance in assessee behavior, that dimensions measured in ACs will manifest themselves differently in different situations, and that these different manifestations are substantively based on the nature of the exercises. Consequently, there is a distinct need moving forward to focus more on the exercises in ACs, with regard to both a theoretical understanding of what factors stimulate candidate behaviors in ACs and practical implications for how AC exercises might be better designed to more fully elicit relevant behavior. In this chapter, we proposed that trait activation theory (Tett & Burnett, 2003) is uniquely suited to providing the framework for a more detailed and systematic examination of the role of the exercise in eliciting candidate behavior, with the goal of improving the usefulness of ACs for both administrative and developmental applications.

The four key axioms in trait activation theory have central relevance to understanding how behavior in ACs is elicited and expressed in response to the situation (i.e., exercise). The first axiom is that relevant behavior will only be expressed in response to relevant situational cues. The second axiom is that behavioral expression is also dependent on the strength of the situation. Thus, these first two tenets of trait activation theory suggest that differences among assessees on a behavioral dimension will only be expressed in the AC to the extent that the exercise (i.e., situation) offers cues to express the relevant behaviors and these cues are not so strong as to

demand that all assessees behave in the same way. The third axiom distinguishes between trait-expressive work behavior and job performance, defining the latter specifically as valued work behavior. The fourth axiom is that trait expression entails both internal reward systems (i.e., inherent in the expression of the trait itself) and external reward systems (i.e., from the reactions of others). In short, AC exercises differ in terms of their trait activation (and therefore behavior eliciting) potential, and trait activation theory allows for a more fine-grained approach to the linkage of exercises and behaviors. As such, it provides both a theoretical framework for interpreting findings in the extant AC literature (regarding, e.g., criterion-related and construct-related validity evidence) and prescriptions for actually improving the design of AC exercises. These prescriptions regarding the design of AC exercises are summarized in the following section and then concluded with important directions for future research.

Summary of AC Exercise Prescriptions

At a broad level, a trait activation theory approach to ACs involves recognition of the importance of building various stimuli into the AC exercises. More attention needs to be paid to the exercise factors that trigger and release relevant candidate behavior versus those that constrain and distract such behavior (Tett & Burnett, 2003). The view taken in this chapter is that exercises can be designed explicitly to increase their situation trait relevance, and we offered some prescriptions for doing so.

First, it is possible (and advised) to adapt the *content of the exercise* to include explicit and specific relevant cues. These content cues should be embedded at the task, social, and organizational levels (e.g., see Tables 1 and 2). Second, situation trait relevance can also be explicitly manipulated through the *instructions* accompanying exercises (e.g., whether such instructions are vague vs. directive, or unidimensional vs. multidimensional). Third, situation trait relevance can be further manipulated through the *cues provided by role players* in exercises. The training of role players should include explicit discussion of how they should act to best elicit relevant behavior in each exercise. Fourth, across each of these cue areas (i.e., content of the exercise, instruction sets, role player behavior), the provision of stimuli needs to be explicit enough to activate candidates' propensities, while at the same time subtle enough to avoid presenting the candidate with too strong of a situation (in terms of behavioral demands). Fifth, we noted that some of these explicit provisions of situational cues

might be best accomplished through the inclusion of a larger number of shorter exercises in ACs, and possibly as videotaped/virtual reality stimuli presented to the candidates. Finally, this identification of the important components of the situation offers some guidance in the development of *alternate forms of AC exercises*. Specifically, exercise features critical for triggering expression of targeted dimensions need to be distinguished from the more incidental features, permitting alternate forms of exercises sharing the former, not the latter.

There are also several implications stemming from the trait *evaluation* part of trait activation theory, regarding how behaviors in ACs should be observed and evaluated. First, trait activation theory offers specific predictions about what should be expected when dimensions are included in ACs (and highlights that care must be taken when deciding on which dimensions to include). Second, trait activation theory is informative with regard to how behavioral checklists might be designed to better capture behavior in ACs, including that (a) there should be a distinction made on the checklist between the observation of behavior and its evaluation; (b) behavioral checklists might be fruitfully organized around behavioral clusters; and (c) the situational cues embedded in exercises to better elicit behavior – for example, cues provided by role players – might themselves be included on the checklists. Third, trait activation theory also offers an approach to operationalizing assessor expertise – one based in assessors' ability to understand how traits are revealed in behavior and activated by situations – that could be useful in terms of both selecting and training assessors.

It is important to note that we are not proposing that current best practices of exercise development (and AC design in general) be abandoned. Rather, we argue that trait activation theory should play a more prominent role in such development, with the goal of making a good tool even better. Whereas current practices typically simulate key task, social, and organizational demands of the job, we see untapped potential in planting multiple stimuli within exercises as a systematic and structured means of increasing the frequency and variability of job-related behavior in AC exercises. Neither do these recommendations necessarily imply that it is personality traits per se that should be targeted for assessment in AC exercises. Rather, we seek to better elicit expression of job-relevant *behavior*. Regardless of how that behavior is then captured and evaluated by assessors (in exercise- or task-based models, dimension-based models, etc), eliciting and observing behavior is key to effective assessment and development centers.

Summary of Future Research Needs

The foregoing analysis identifies several important areas for future research, which are briefly summarized here. First, there is the most obvious need to empirically confirm each of the preceding prescriptions regarding how manipulating situational cues (through content, instructions, etc.) in exercises would affect the behavior elicited in such exercises. Second, there is a similar need regarding the effects of the strength of such cues. We need to know both how cue strength affects the behavior elicited and the correspondence between points on the strong-weak situation continuum and specific aspects of the exercise such as content and instructional sets. Third, and more generally, AC research has been silent with regard to which specific exercise characteristics might trigger specific candidate behavior, and we feel that much more extensive programmatic research is needed to uncover the important elements of exercises (including distinguishing between those characteristics that are "incidentals" and those that are "radicals"). In the end, this could generate a theory of AC performance.

Fourth, as discussed in earlier sections of this chapter, there is the need for research to examine how the abovementioned trait activation processes might vary for different types of dimensions (e.g., motivational vs. skill-based dimensions) as well as how social perceptiveness and other individual differences might moderate trait activation. A final important agenda for research in this area involves the investigation of how the increased situational relevance of AC exercises would impact the important "outcomes" of ACs, including the (a) perceptions of assessees, (b) inter-rater reliability of assessors, (c) convergence of AC ratings across exercises (i.e., "internal" construct-related validity), (d) convergence between AC ratings and other similar constructs in a nomological network (i.e., "external" construct-related validity), (e) the criterion-related validity of ACs, and (f) the effectiveness of feedback provided after the AC. We hope this chapter will encourage research in each of these important areas, with the ultimate goal of identifying theoretically rich and practically useful guidelines for further improving ACs through systematic attention to AC exercises as behavior-eliciting situations.

REFERENCES

Ahmed, Y., Payne, T., & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment*, *5*, 62–68.

Allport, G. W. (1951). *Personality: A psychological interpretation*. London: Constable.
Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology*, *91*, 555–566.
Argyle, M. (1969). *Social interaction*. Great Britain: Butler & Tanner Ltd.
Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*, 125–154.
Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435–442.
Atkins, P. W. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, *55*, 871–904.
Bakan, D. (1966). *The duality of human existence: Isolation and communion in western man*. Boston: Beacon.
Bem, D. J., & Allen, A. (1974). On predicting some of people some of time: Search for cross-situational consistencies in behaviour. *Psychological Review*, *81*, 506–520.
Binning, J. F., Adorno, A. J., & Kroeck, K. G. (1997). 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, April, St. Louis, MO.
Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.
Borteyrou, X. (2005). *Intelligence, personnalité, mises en situation et prédiction de la réussite professionnelle: La construction d'un centre d évaluation pour des officiers de marine*. Unpublished doctoral dissertation, Université Victor Segalen Bordeaux, Bordeaux.
Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review*, *80*, 307–336.
Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*, 1114–1124.
Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 131–133.
Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology*, *74*, 957–963.
Brink, K. E., Lance, C. E., Bellenger, B. L., Morrison, M. A., Scharlau, E. A., & Crenshaw, J. L. (2008). Discriminant validity of a "next generation" assessment center. In: B. J. Hoffman (Chair), *Reexamining assessment centers: Alternate approaches*. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.
Brostoff, M., & Meyer, H. H. (1984). The effects of coaching on in-basket performance. *Journal of Assessment Center Technology*, *7*, 17–21.
Brummel, B. J., Rupp, D. E., & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology*, *62*, 137–170.

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

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

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

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, *18*, 123–149.

Connelly, B. S., & Ones, D. S. (2008). Interrater unreliability in assessment center ratings: A meta-analysis. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, April, San Francisco, CA.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*, 565–579.

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology*, *18*, 94–104.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Cote, S., & Moskowitz, D. S. (1998). On the dynamic covariation between interpersonal behavior and affect: Prediction from neuroticism, extraversion, and agreeableness. *Journal of Personality and Social Psychology*, *75*, 1032–1046.

Craik, K. H., Ware, A. P., Kamp, J., O'Reilly, C., Staw, B., & Zedeck, S. (2002). Explorations of construct validity in a combined managerial and personality assessment programme. *Journal of Occupational and Organizational Psychology*, *75*, 171–193.

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

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology: An International Review*, *52*, 193–212.

Dennis, I., Handley, S. J., Bradon, P., Evans, J., & Newstead, S. E. (2002). Towards a predictive model of the difficulty of analytical reasoning items. In: S. H. Irvine & P. C. Kyllonen (Eds), *Item generation and test development* (pp. 53–71). Mahwah, NJ: Erlbaum.

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

Dulewicz, V., & 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.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289–303.

Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin*, *81*, 1026–1048.

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.

Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin*, *98*, 513–537.

Ferris, G. R., Perrewé, P. L., & Douglas, C. (2002). Social effectiveness in organizations: Construct validity and research directions. *Journal of Leadership & Organizational Studies*, *9*, 49–63.

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

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.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structures. *Psychological Assessment*, *4*, 26–42.

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

Goodstone, M. S., & Lopez, F. E. (2001). The frame of reference approach as a solution to an assessment center dilemma. *Consulting Psychology Journal: Practice and Research*, *53*, 96–107.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, *55*, 137–163.

Hagan, C. M., Konopaske, R., Bernardin, H. J., & Tyler, C. L. (2006). Predicting assessment center performance with 360-degree, top-down, and customer-based competency assessments. *Human Resource Management*, *45*, 357–390.

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

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

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*, 639–683.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, *15*, 405–411.

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

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

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

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

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology*, *81*, 173–195.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation and test development*. Mahwah, NJ: Erlbaum.

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

Jansen, A., Lievens, F., & Kleinmann, M. (2009). The importance of situation perception in the personality-performance relationship. Paper to be presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Johnson, J. A. (1997). Units of analysis for description and explanation in psychology. In: R. Hogan, J. A. Johnson & S. R. Briggs (Eds), *Handbook of personality psychology* (pp. 73–93). San Diego, CA: Academic Press.

Jones, R. G., & Whitmore, M. D. (1995). Evaluating developmental assessment centers as interventions. *Personnel Psychology*, *48*, 377–388.

Julian, E. R., & Schumacher, C. F. (1988). CBT pilot examination: Results and characteristics of CBX. Paper presented at the conference of the National Assessment Centers Board of Medical Examiners on Computer-based Testing in Medical Education and Evaluation, March, Philadelphia, PA.

Kelbetz, G., & Schuler, H. (2003). Does practice improve assessment center performance? *Zeitschrift für Personalpsychologie*, *1*, 4–18.

Klein, S. P. (1992). The effect of content area and test type on bar exam scores. Paper presented at the National Conference of Bar Examiners.

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. *Zeitschrift für Arbeits und Organisationspsychologie*, *41*, 171–181.

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

Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review*, *45*, 67–84.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284.

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

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

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

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U.-C. (2006). The relationship between the ability to identify evaluation criteria and integrity test scores. *Psychology Science*, *48*, 369–377.

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U.-C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment*, *15*, 283–292.

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

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

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

Lance, C. E. (2008). Why assessment centers (ACs) don't work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 84–97.

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004a). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*, 22–35.

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

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

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

Lebreton, J. M., Gniatczyk, L. A., & Migetz, D. Z. (1999). 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, May, Atlanta, GA.

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. (2001a). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, *65*, 1–19.

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

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

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

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

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

Lievens, F., De Fruyt, F., & Van Dam, K. (2001). Assessors' use of personality traits in descriptions of assessment centre candidates: A five-factor model perspective. *Journal of Occupational and Organizational Psychology*, *74*, 623–636.

Lievens, F., De Koster, L., & Schollaert, E. (2008). Current theory and practice of assessment centers. In: S. Cartwright & C. Cooper (Eds), *Oxford handbook of personnel psychology* (pp. 215–233). Oxford: University Press.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? *International Review of Industrial and Organizational Psychology*, *16*, 246–286.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, *92*, 1043–1055.

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, *57*, 881–904.

Lievens, F., & Van Keer, E. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organizational Psychology*, *74*, 373–378.

Marcus, B., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. *Human Performance*, *20*, 275–285.

Mayes, B. T., Belloli, C. A., Riggio, R. E., & Aguirre, M. (1997). Assessment centers for course evaluations: A demonstration. *Journal of Social Behavior and Personality*, *12*, 303–320.

McFarland, L. A., Yun, G. J., Harold, C. M., Viera, L., & Moore, L. G. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology*, *58*, 949–980.

Melchers, K. G., Klehe, U. C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (in press). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*.

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

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*, 252–283.

Moses, J. L., & Ritchie, R. J. (1976). Supervisory relationships training: A behavioral evaluation of a behavior modeling program. *Personnel Psychology*, *29*, 337–343.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749–761.

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

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

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.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*, 868–897.

Ployhart, R. E., Lim, B. C., & Chan, K. Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology*, *54*, 809–843.

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

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centers: Dimensions into exercises wont go. *Journal of Occupational Psychology*, *60*, 187–195.

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

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

Rupp, D. E., Gibbons, A. M., & Snyder, L. A. (2008). Transforming our models of learning and development: Web-based instruction as enabler of third-generation instruction. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 454–467.

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

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

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

Schleicher, D. J., Day, C. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*, 735–746.

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

Schollaert, E., & Lievens, F. (2008). The effects of exercise instructions on the observability of assessment center behavior. In: K. G. Melchers (Chair), *Assessment centers: Organizational practices, individual differences correlates and influencing factors on construct validity*. Symposium conducted at the International Congress of Psychology, July, Berlin, Germany.

Shavelson, R. J., Baxter, G. P., Pine, J., Yure, J., Goldman, S. R., & Smith, B. (1991). Alternative technologies for large-scale science assessment: Instrument of education reform. *School Effectiveness and School Improvement*, *2*, 1–8.

Shavelson, R. J., Mayberry, P., Li, W., & Webb, N. (1990). Generalizability of job performance measurements: Marine Corps rifleman. *Military Psychology*, *2*, 129–144.

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

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

Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions transparent on relations with typical performance predictors. *Human Performance*, *20*, 187–203.

Spector, P. E., Schneider, J. R., Vance, C. A., & Hezlett, S. A. (2000). The relation of cognitive ability and personality traits to assessment center performance. *Journal of Social Psychology*, *30*, 1474–1491.

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

Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences*, *11*, 833–839.

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

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

Tett, R. P., & Schleicher, D. J. (2001). Assessment center dimensions as "traits": New concepts in AC design. In: M. Born (Chair), *Assessment Center Dimension Validation: Are we asking the wrong questions?* Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, April, San Diego, CA.

Thorndike, E. L. (1920). Intelligence and its use. *Harper's Magazine*, *140*, 227–235.

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

Thornton, G. C. III, Larsh, S., Layer, S., & Kaman, V. (1999). 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, May, Atlanta, GA.

Thornton, G. C., III., & Mueller-Hanson, R. A. (2004). *Developing organizational simulations: A guide for practitioners and students*. Mahwah, NJ: Erlbaum.

Thornton, G. C., III., & Rupp, D. E. (2005). *Assessment centers in human resources management*. Mahwah, NJ: Lawrence Erlbaum.

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

Trippe, D. M., & Foti, R. J. (2003). An evaluation of the construct validity of situational judgment tests. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, April, Orlando, FL.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, *17*, 71–93.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376.

Wiggins, J. S., & Trapnell, P. D. (1996). A dyadic-interactional perspective on the Five-Factor Model. In: J. S. Wiggins (Ed.), *The Five-Factor Model of personality: Theoretical perspectives* (pp. 88–162). New York: Guildford Press.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*, 231–258.

Woo, S. E., Sims, C. S., Rupp, D. E., & Gibbons, A. M. (2008). Development engagement within and following developmental assessment centers: Considering feedback favorability and self-assessor agreement. *Personnel Psychology*, *61*, 727–759.

Wu, Z. M., & Zhang, H. C. (2001). The construct validity and structure modeling of assessment centers. *Acta Psychologica Sinica*, *33*, 372–378.

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