Human Resource Management Review 19 (2009) 169–187

Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Validity of assessment centers for personnel selection

George C. Thornton III *, Alyssa M. Gibbons

Department of Psychology, Colorado State University, Fort Collins, CO 80523, United States

ARTICLE INFO

Keywords:
Assessment centers
Personnel selection
Validity

ABSTRACT

Research and practice in the application of assessment centers (AC) for personnel selection are reviewed and critiqued. Several examples of the use of ACs for external screening, internal promotion, and certification are described. Several types of evidence of validity of ACs for selection are reviewed, including representativeness of the content of dimensions and exercises in relation to job requirements, relationships among ratings within an AC, relationships of AC ratings and criteria of work effectiveness, and consequences of assessments including candidates' reactions to assessments and sub-group differences in ratings. Several controversies in research findings and practices of ACs are noted. Further research to address these controversies and new research to study emerging issues are suggested. Conclusions about the validity, fairness, and legal defensibility of ACs for personnel selection are offered.

The assessment center method (ACM) has been used for many purposes in human resource management including selection, diagnosis, and development since its introduction over 50 years ago, (Thornton & Rupp, 2006). In this chapter, we review research and practice of the method for selection purposes. We use "selection" in a broad sense to mean the use of overall assessment ratings to aid in selection of:

- external candidates into organizations,
- internal candidates into supervisory and managerial ranks,
- individuals into a pool of high potentials who will get special training,
- exemplary staff members to receive certification of competence in job skills, or
- employees for retention when there is a reduction in force and reorganization.

In all of these applications, the overall assessment rating is used as a measure of competence to be successful in some new assignment.

We begin with a definition of an assessment center (AC), and then give examples of several applications in selection. Next, we evaluate the literature to provide summaries of what is known about the validity of the method and controversies over theory, research, and practice. We conclude with a set of suggestions for research, including a new broad proposal for integrating a number of key issues percolating in the field, along with some specific research needs. In this paper, we intend to provide a summary of what will be useful to both scholars seeking key issues to investigate and practitioners facing challenges in applications.

1. Definition of assessment center

As the term has been used in the fields of industrial psychology and human resource management since its introduction over 50 years ago, "assessment center" refers to a method involving a unique combination of essential elements codified in the *Guidelines and Ethical Considerations of Assessment Center Operations* (henceforth *Guidelines*; International Task Force, 2008).

- Multiple, trained assessors observe overt behavior displayed by assessees in complex organizational simulations and make ratings of performance on dimensions deemed important for effective performance in target positions

* Corresponding author. Tel.: +1 970 491 5233.
E-mail address: George.Thornton@colostate.edu (G.C. Thornton).

1053-4822/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.hrmr.2009.02.002

- Dimensions of performance are identified by various methods of job analysis (e.g., the analysis of tasks and responsibilities in a particular job or job group) and/or competency modeling (e.g., identification of competencies needed to achieve the organization's strategic goals)
- Any dimension that can be defined in terms of observable behaviors has potential for assessment. Examples of typical dimensions include managerial skills, interpersonal effectiveness in teams, leadership, sales abilities, etc.
- Behavioral observations can also be summarized in terms of tasks or responsibilities in the target job.
- Assessors can be operational managers above the target position, human resource managers, psychologists, and external consultants.
- Assessors follow a systematic process of recording behavioral observations and making independent evaluations
- Other information from objective performance indices, background interviews, scores on cognitive ability and personality tests, performance evaluations from supervisors, and descriptions from multi-source (360°) rating systems may also be considered
- Methods of integrating information can include the classic process of sharing and discussing behavioral observations and evaluations to achieve consensus, statistical processes of combining ratings, or some combination of these.

When used for selection, the objective is to provide an overall evaluation of a candidate's ability to be successful in the future in a new assignment. Thus, ratings are typically combined into a single *overall assessment rating* (OAR), which is used for decision-making.

2. Examples of assessment centers used for selection

The assessment center method is extremely versatile; it has been used to aid selection of persons for a wide range of managerial and non-managerial positions.

2.1. External selection

The police department of Fort Collins, Colorado used an assessment center as one step in screening candidates for patrol officer (Gavin & Hamilton, 1975). The exercises simulated interpersonal situations officers may encounter on the job, such as domestic disputes and disorderly persons on the street. Assessment centers are certainly expensive and there is certainly a question of whether this practice is economically justified at the entry level. Coulton and Feild (1995) argued that the potential benefits exceed the costs of an AC for police selection. In fact, the Israeli police force has used a variety of performance exercises simulating challenging physical and leadership situations to screen candidates, and found the 2-day behavioral assessments added unique predictive accuracy over cognitive ability tests (Dayan, Kasten, & Fox, 2002). In the United Kingdom, assessment centers have been used by many organizations to select management trainees. With the rising expectations to have manufacturing employees contribute to teamwork and continuous quality improvement, organizations have turned to the ACM to assess interpersonal and decision making skills. One of the areas of great expansion of applications in the 1980s and 1990s came in large organizations such as Diamond Star Motors (Henry, 1988), Cessna (Hiatt, 2000), Coors (Thornton, 1993), and BASF (Howard & McNelly, 2000) and even small organizations such as Wilkerson Manufacturing (Thornton, personal communication, June 15, 2008). Assessment centers have also been used to select pilots (Damitz, Manzey, Kleinmann, & Severin, 2003), automotive manufacturing trainers (Franks, Ferguson, Rolls, & Henderson, 1999), and salesmen (Bray & Campbell, 1968).

2.2. Certification of competence

In many organizations, especially teaching and the IT industry, formal certification of individual employees carries considerable weight. Thus, Connecticut has used assessment centers to certify teachers (Jacobson, 2000) and Sun Microsystems developed behavioral assessments to certify the competence of consultants providing services in the design of selection and training programs for clients and of customer service representatives providing telephone support (Howard & Metzger, 2002; Rupp & Thornton, 2003). Sackett (1998) reported the use of related methods with lawyers.

2.3. Promotion of internal candidates

The most frequent application of the ACM has been the selection of non-supervisory personnel into first level management. Many early practical applications of the method in large organizations such as AT&T, Standard Oil, Sears, and IBM were designed to screen manufacturing or sales persons into first level management (Thornton & Byham, 1982). The method is particularly suited for this application, because it gives the organization a chance to observe candidates in situations that are quite different from the current job and because it gives candidates the opportunity to demonstrate competencies they may not be able to display in current assignments.

The ACM has also been used to aid promotion decisions into middle and upper management levels. In fact, the purpose of the original AT&T Management Progress Study involving men (Bray & Grant, 1966) and Management Continuity Study involving women (Howard & Bray, 1988) was to identify persons with potential for success in a wide array of middle management positions, some of which may not be known at the time of assessment. These studies provided evidence that OARs predicted management progress and performance over several years. At the top executive levels, Howard (1997) reported that the ACM has been used to assess both internal and external candidates. With the increasing pressure to evaluate top leaders, the ACM may be used even more frequently (Howard, 2001; Thornton, Hollenbeck, & Johnson, in press).

2.4. Identification of high potentials

The assessment center method has become a core feature of many succession planning programs involving systematic procedures to identify staff members, early in their careers, with high potential for long range leadership success. These "hi pos" are then moved through a set of training, mentoring, and experiential assignments (Byham, Smith, & Paese, 2000). Subsequent promotion to executive ranks is based on proven performance, but early identification is a critical first step.

3. Validity of ACM for selection

We advocate the modern view of validity embodied in the primary professional authorities of psychometrics related to personnel testing, i.e., *Standards for Educational and Psychological Tests* (American Educational Research Association, American Psychological Association, & American Council on Measurement in Education, 1999) and *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2003). Validity consists of broad bodies of accumulated sets of evidence supporting the inference to be made from test scores. The "unitarian" view of validity (Landy, 1986) holds that there are not separate types of validity but rather one validity supported by all accumulated evidence relevant to an inference in a specific application. When the ACM is used to aid selection decisions, the inference is that the overall assessment rating predicts performance in the future in some specified job or set of jobs. Thus, we will review a variety of probative evidence comprising sets of evidence set forth in the professional authorities.

3.1. Content representativeness

Having appropriate and representative content is particularly critical for ACs, as much of the popularity and legal defensibility of the method is grounded in the close resemblance of AC activities to the target job (Thornton & Rupp, 2006). The *Guidelines* (International Task Force, 2008) clearly state that assessment center dimensions and exercises should be based on a thorough job analysis. Although Norton (1981) argued that an exact match between AC content and the target job was not necessary, due to the resemblance of the AC as a whole to managerial work as a whole, few AC developers have been comfortable with this loose definition of "content." Surveys of the field suggest that most operational ACs rely on job analyses, using a variety of techniques including questionnaires, observation of job incumbents, and competency modeling (Krause & Thornton, in press; Spychalski, Quinones, Gaugler, & Pohley, 1997). Thornton and Krause (in press) found that whereas ACs for selection purposes are more likely to conduct new job analyses and create new competency models specifically for the AC, developmental ACs are more likely to use an organization's existing competency model.

The presence of job analysis might be considered a necessary, but not sufficient, condition for establishing that the content of a given AC is appropriate. Sackett (1987) argued that, in addition to the dimensions and exercises, the scoring system, variance of expected possible responses, and instructions to the participants are also part of the content of the AC. At the time of his review, Sackett noted that these elements were seldom considered in the job analysis or content validation process, yet they have important implications for determining whether the AC is truly relevant and appropriate for the job in question. For example, if only a moderate level of leadership proficiency is necessary for the target job, but the scoring system is such that candidates who perform at that level receive low ratings, the AC cannot be said to measure the leadership behavior that is required on the job. More recently, AC developers have considered a wider range of variables in AC design. Thornton and Mueller-Hanson (2004) encapsulate many of these variables in their broader "situational analysis," which takes into account not only dimensions and tasks but also the level of performance required and the organizational context. However, the degree to which both researchers and practitioners have followed these more complex approaches in establishing the appropriateness of AC design remains unclear. Spychalski et al.'s (1997) survey addressed only traditional (dimension- and exercise-focused) job analytic techniques, and published accounts of ACs seldom describe the content validation process in any detail.

3.2. Correlations with external criteria

The ACM has a long history of demonstrating strong predictive relationships between AC ratings and criteria outside the AC such promotions, performance evaluations, salary progress (Adler, 1987; Thornton & Byham, 1982). Both overall assessment ratings (OARs) and final dimension (and sometimes exercise) ratings have been found to predict a variety of relevant external variables. In this section, we consider evidence for each type of rating separately, along with factors that appear to moderate the predictive relationships. As we consider the accumulated evidence for AC validity, it is important to remember that ACs are a method, not a construct, and that different ACs may measure different constructs (Arthur & Villado, 2008). Certainly, many ACs assess common or similar dimensions (Arthur, Day, McNelly, & Edens, 2003), but even dimensions with similar names may be defined and operationalized very differently in different ACs. The predictive validity of any particular AC may deviate considerably from the "average" or illustrative results discussed here, depending on the constructs measured, the target job, and the overall quality of implementation. Nevertheless, meta-analyses and other studies that generalize across ACs offer a basis for comparing ACs (with one another and with other methods), examining trends, and identifying factors that contribute to or detract from the predictive power of ACs in general.

In the AT&T Management Progress Study, Bray and Grant (1966) reported that assessors' judgments of candidates' management potential predicted their actual promotion to middle management level at rates far exceeding what would be

expected by chance. Further, overall effectiveness scores from the AC predicted salary progress, with correlations of.41 to .52 across four samples. Some later studies found similarly impressive predictive validities. For example, Bray and Campbell (1968) found a correlation of .51 between OARs and subsequent job performance ratings, and Borman (1982) found a corrected correlation of .48 between OARs and performance in training. A meta-analysis by Gaugler, Rosenthal, Thornton, and Bentson (1987) found an average corrected validity of .37 between OARs and various criterion measures. According to Schmidt and Hunter's (1998) review of selection methods, this average validity for ACs compares favorably with many other assessment methods. Subsequent studies have continued to find substantial correlations between OARs and performance criteria (Chan, 1996; Jansen & Stoop, 2001; Tziner, Ronen, & Hacohen, 1993). However, recent meta-analyses have found lower average validity coefficients between the OAR and performance ratings ($r = .26$, Hardison & Sackett, 2004; $r = .28$, Hermelin, Lievens, & Robertson, 2007). There are reasons to believe any of these average correlations under-estimate actual validity. Meta-analyses tend to use relatively high levels of reliability of criterion measures such as performance appraisal when making corrections (Lievens & Thornton, 2005). In addition, initial screening of candidates for selection or promotion before the AC may lead to restriction of range in AC scores.

The reasons for the ostensible decline in validity over the past 40 years are not clear. Differences in the traditional notion of publication bias (i.e., submission and acceptance of studies with higher magnitude validity; Rothstein, Sutton, & Borenstein, 2005) does not appear to explain the decline, because whereas 10 of 47 studies in Gaugler et al. (1987) were unpublished only 4 or 26 studies in Hermelin et al. (2007) were unpublished. Another form of publication bias, which might be called an iconoclastic urge, may be operating in recent years such that high-validity AC studies are considered passé, yet low-validity studies may be published because they seem to debunk an established method. In addition, it could be that recent ACs are not designed and implemented as rigorously, or that recent research studies are not as rigorous (e.g., small unrepresentative samples, unreliable criteria). Clearly, further research is needed to understand the reasons for the reported decline in validity estimates.

As noted above, because of the variety of constructs that may be assessed in ACs, meta-analytic results for ACs must be interpreted as average validities, not "true" validity (as is sometimes the case for constructs such as job satisfaction; cf. Judge, Thoresen, Bono, & Patton, 2001). Several studies have examined possible moderators of predictive validity for ACs. Gaugler et al. (1987) found that ACs showed stronger predictive validity when larger numbers of exercises and other assessments were used, when psychologists served as assessors along with managers, when there was a high proportion of female assessees, and when peer evaluation was included. Binning, Gniatczyk, LeBreton, and Melcher (2002) found that individual assessors varied in the predictive validity of their judgments. In particular, assessors who paid more attention to task behaviors had more valid ratings, while assessors who attended to teamwork and interpersonal behaviors had less valid ratings. Some studies suggest that mechanically or statistically combining assessor ratings, rather than arriving at an OAR through consensus judgment, produces higher predictive validity (Borman, 1982; Feltham, 1988). However, Chan (1996) found that assessors' consensus judgments of promotability were more strongly correlated with performance and promotion criteria than was a mechanically-derived OAR. Factors outside the design of the AC may also impact validities. Binning, Adorno, and LeBreton (1999) found that OARs were better predictors for more complex jobs, and Gaugler et al. (1987) found higher validities when the original validity study was methodologically sound.

While the bulk of AC criterion-related validity studies have focused on the OAR, some have examined the predictive validity of individual dimensions. Specifically, these studies consider the final, post-integration ratings of each dimension for each candidate, considering all relevant exercises. In the Management Progress Study, Bray and Grant (1966) found that many of their individual dimensions predicted salary progress, though none predicted quite so well as the OAR. Arthur et al. (2003) conducted a meta-analysis of the predictive validity of dimension ratings, categorizing common AC dimensions into six broad groups, e.g., consideration/awareness, communication, problem solving. They found average corrected criterion-related validities between .25 and .39 for all six dimension categories. Further, a weighted composite of the individual dimension ratings predicted performance criteria with an average corrected validity coefficient of .45.

3.3. Correlations of OAR with other measures

Several studies have also examined the relationships between AC ratings and other relevant external variables. A meta-analysis by Collins et al. (2003) found that overall AC performance was consistently correlated with measures of both cognitive ability and personality. Further, there is evidence that different dimensions and exercises correlate differentially with these external measures. Cognitive ability appears to be most strongly related to "performance style" dimensions such as problem solving and planning (Crawley, Pinder, & Herriot, 1990; Shore, Thornton, & Shore, 1990) and to exercises emphasizing such skills, such as in-baskets and project planning tasks (Spector, Schneider, Vance, & Hezlett, 2000). Similarly, Lance et al. (2000) found stronger relationships between specific ability measures such as math and writing fluency and cognitively laden exercises such as the in-basket than with more interpersonal exercises such as role plays. Personality traits, on the other hand, show stronger relationships with "interpersonal style" dimensions such as impact on others (Shore et al., 1990) and interpersonally oriented exercises such as interviews and role plays (Spector et al., 2000). Lievens, De Fruyt, and Van Dam (2001) demonstrated that assessors tended to perceive different personality traits in different exercises; for example, assessors were most likely to note instances of extraversion in a group discussion exercise and conscientiousness in an in-basket task. Interestingly, these variables do not fully account for the predictive validity of AC ratings (cf. Klimoski & Brickner, 1987). Overall AC ratings have been found to have incremental predictive validity over both cognitive ability (Dayan et al., 2002; Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Hardison, 2005; Krause, Kersting, Heggestad, & Thornton, 2006) and personality (Goffin, Rothstein, & Johnston, 1996; Hardison, 2005).

3.4. Internal structure

For many selection devices, psychometric analyses are used to verify that the individual components of the assessment (typically, items on a test or questionnaire) correspond to the theoretical structure of the construct(s) being measured (SIOP, 2003). Establishing this type of validity evidence for ACs has often proven challenging, perhaps because of disagreement regarding precisely what the "items" are and what the theoretical structure is supposed to be. Early ACs (Bray & Grant, 1966; Bray, Campbell, & Grant, 1974) considered the "items" of interest to be the final, across-exercise consensus ratings on each dimension. They examined the correlations among these final dimension ratings and concluded that logically similar dimensions were generally more strongly related than dissimilar dimensions.

Subsequent researchers began to view the "items" of interest as the dimension ratings made within or following each exercise (Neidig, Martin, & Yates, 1979; Turnage & Muchinsky, 1982). They expected that in a valid AC, ratings of the same dimension across exercises would be highly correlated (convergent validity) and ratings of different dimensions within an exercise would be correlated to a much lesser extent (discriminant validity). However, many of the studies that followed this approach found precisely the opposite pattern: high correlations among dimension ratings from the same exercise and lower correlations among ratings of the same dimension from different exercises (Archambeau, 1979; Neidig et al., 1979; Sackett & Dreher, 1982). Researchers in this tradition followed the logic of the multitrait-multimethod matrix (MTMM; Campbell & Fiske, 1959), arguing that dimensions were analogous to traits and exercises to methods. Trait or dimension effects were considered meaningful; method or exercise effects were considered error or noise. This MTMM paradigm came to dominate the assessment center "construct validity" literature over the next 3 decades. Despite numerous advances in analysis techniques, the pattern of strong exercise effects and somewhat weaker dimension effects persisted, leading many to conclude that assessment centers lacked discriminant validity and often convergent validity as well (Bycio, Alvares, & Hahn, 1987; Chan, 1996; Jansen & Stoop, 2001; Lance, Foster, Gentry, & Thoresen, 2004; Schneider & Schmitt, 1992).

In an early MTMM study, Turnage and Muchinsky (1982) argued that their failure to find strong evidence of discriminant validity was likely a function of error on the part of assessors. Many researchers therefore turned their attention to understanding and alleviating assessor error, with some success (Lievens, 1998). In particular, strategies that seem to be effective include providing assessors with training (Schleicher, Day, Mayes, & Riggio, 2002; Thornton & Zorich, 1980), using alternative rating strategies (Kolk, Born, & van der Flier, 2002; Robie, Osburn, Morris, Etchegaray, & Adams, 2000), and reducing assessors' cognitive load, such as limiting the number of dimensions to be assessed (Gaugler & Thornton, 1989) or providing checklists to record behavior (Reilly, Henry, & Smither, 1990). Although many of these studies produced improvements in the relative size of MTMM convergent validity and discriminant validity correlations, few if any succeeded in producing a clear pattern of dimension effects with little or no exercise effects. One exception is the work of Arthur, Woehr, and Maldegen (2000), who found strong person and dimension effects and small exercise effects using a generalizability theory analysis. On the whole, however, this body of research led many researchers to conclude that the dimensions traditionally assessed in ACs were not viable constructs and that AC researchers should focus instead on exercises as work samples (e.g., Lance, Lambert, Gewin, Lievens, & Conway, 2004; Sackett & Tuzinski, 2001). Viewing the exercise as a work sample would call for an evaluation of overall performance in the exercise, but not evaluations of separate dimensions within the exercise.

More recent research, however, has taken a different view. Whereas the early MTMM studies assumed that exercise effects reflected measurement error, contemporary research suggests that such effects may in fact represent real variation in performance from exercise to exercise (Lievens, 2002) and that that variation is related to job-relevant variables (Lance, Newbolt, et al., 2000). Further, some studies have found evidence for distinct dimensions within exercises (Rupp et al., 2006; Hoffman & Meade, 2007). Taken together, these studies suggest that both dimensions and exercises have meaningful roles to play in determining overall assessment center performance. There appears to be growing consensus in the AC field that the MTMM model, which treats exercise effects as measurement error, is not appropriate for ACs. Indeed, the members of a panel discussion at the 2008 annual meeting of the Society for Industrial and Organizational Psychology unanimously called for a moratorium on MTMM studies of within-exercise dimension ratings (Hoffman, 2008).

A recent issue of the journal *Industrial and Organizational Psychology: Perspectives on Science and Practice* highlights numerous contemporary perspectives on the use of MTMM methods and of appropriate ways to establish the construct validity of ACs (see Lance, 2008a, and subsequent commentaries). In response to the commentaries, Lance (2008b) stated:

> I suggest that the alleged construct validity problem is not a problem with the way ACs work but with the MTMM urban legend's associations of dimensions with traits and exercises with methods in the MTMM framework (p. 141).

3.5. Consequences of testing including fairness

With the evolution of the unitary or holistic view of validity has come a new emphasis on the consequences of assessments and assessment processes (Messick, 1995, 1998). Studies of the consequences include examination of bias and applicants' responses to assessment procedures. Selection methods should be free from bias, particularly relative to members of legally protected groups (SIOP, 2003), and should seek to avoid adverse impact and other negative consequences as far as possible (American Educational Research Association et al., 1999). Historically, ACs have demonstrated little evidence of systematic bias. Early comparisons found similar average ratings for men and women (Moses & Boehm, 1975) and for black and white candidates (Huck & Bray, 1976). Differences between groups appear to be greater in exercises that require cognitive ability to a greater extent (Goldstein et al., 1998).

However, because ACs generally measure non-cognitive as well as cognitive skills (see the preceding discussion of incremental validity), subgroup differences in overall ratings are likely to be less for an AC than for traditional cognitive ability measures (Schmitt & Mills, 2001). Schmitt (1993) found some evidence for interactions between the race and gender of the assessor and that of the candidate, but the effect size of these interactions was generally small. As a result of these findings, ACs have earned a reputation as a relatively fair and unbiased selection technique (Cascio & Aguinis, 2005); indeed, several court decisions have recommended ACs as an alternative to cognitive ability tests when adverse impact is present (Thornton & Johnson, 2006).

Other research, however, suggests that this picture may not be entirely clear. A recent meta-analysis (Dean, Roth, & Bobko, 2008) found an average difference of $d = 0.52$ between OARs for black and white candidates, with white candidates receiving higher ratings. A smaller difference was found between white and Hispanic candidates, and there was a slight positive effect for women as compared to men. Similarly, Anderson, Lievens, Van Dam, and Born (2006) found higher mean OARs for women than men. Comparisons of mean dimension scores showed that women scored higher on oral communication and interaction, but men scored higher on problem solving, with d values ranging from $-.31$ to $+.27$. Latent mean analyses revealed similar differences. Bobko, Roth, and Buster (2005) found larger differences between black and white candidates for exercises that emphasized technical material, with smaller differences for exercises emphasizing interpersonal skills. These results suggest that the degree to which subgroup differences appear in AC performance depends, at least to some degree, on the particular constructs being assessed (cf. Roth, Bobko, McFarland, & Buster, 2008). Some studies have also suggested a potential for age-related bias in ACs, with older participants receiving lower overall ratings (Clapham & Fulford, 1997; Dulewicz & Fletcher, 1982). More recently, Gibbons and Rupp (2004) found no association between assessor ratings and age, but the AC in their study was used for developmental purposes, whereas the previous studies were conducted in selection ACs.

These findings suggest that the fairness of ACs cannot simply be assumed. Future research is needed to identify possible moderators of subgroup differences–for example, characteristics of assessors, assessor training, number of "performance style" dimensions (i.e., those that are correlated with cognitive ability; Shore et al., 1990), and so on. Dean et al. (2008) did find a moderating effect of candidate type on the difference between black and white candidates, which was substantially smaller when candidates were incumbents of the organization than when they were applicants. Similarly, Bobko et al. (2005) found a much higher d between black and white job applicants compared to previous studies using job incumbents. They suggest that some of this effect may be due to restriction of range, particularly when the organization's existing selection procedures already create adverse impact. Many minority applicants might be eliminated earlier in the selection process, leaving only those who are already similar to majority group applicants on the constructs measured in the AC. If the AC is validated using job incumbents but used to select applicants, subgroup differences might be substantially underestimated.

Another issue relevant to the consequences of testing concerns applicants' responses to the selection procedure. Applicants who perceive a selection process as fair and job-relevant are more likely to hold more positive views of the organization and report stronger intentions to accept a position (Ryan & Ployhart, 2000; Truxillo, Bauer, Campion, & Paronto, 2002). Historically, ACs have enjoyed a strong positive reputation in this regard. Participants perceive AC exercises as relevant and realistic (Dodd, 1977; Thornton & Byham, 1982), and as more face valid than cognitive ability tests (Macan, Avedon, Paese, & Smith, 1994). However, negative reactions and consequences are possible. Incumbent employees who are unsuccessful in promotional ACs may experience lower self-esteem, competitiveness, and work ethic (Fletcher, 1991) and exhibit less self-development behavior (Noe & Steffy, 1987) following the assessment. By contrast, in a study of students, Schuler and Fruhner (1993) found that some types of self-concept ratings decreased during the actual AC process, but improved above pre-AC levels after participants received specific feedback. An important difference between this study and the preceding studies is that Fletcher (1991) and Noe and Steffy (1987) studied job incumbents in promotional ACs where AC performance was an important determinant of future job outcomes, but Schuler and Fruhner (1993) studied students in a program that appears to have been intended only for developmental purposes. The provision of specific feedback seems to be helpful in improving participants' perceptions of the AC process (Dodd, 1975), but it may be unrealistic to expect feedback to fully mitigate the effects of a negative assessment when such an assessment has tangible consequences.

3.6. Utility

Although not technically considered part of validity, utility is a critical concern for users of the AC method. ACs are expensive, requiring considerable investment of resources, personnel, and time. Organizations that implement ACs must have some assurance that the real financial benefits of the program will outweigh the costs. When utility studies have been conducted for ACs, the results have generally been positive, with estimates of the value of the AC program ranging from a few thousand dollars/hire to over a million dollars for an entire AC program (Thornton & Rupp, 2006). ACs have demonstrated utility in diverse settings, such as education (Hogan & Zenke, 1986) and public safety (Thornton & Potemra, in press), and for diverse positions, from first-level supervisors (Bobrow & Leonards, 1997) to upper management (Tziner, Meir, Dahan, & Birati, 1994).

Two essential keys to utility are the validity coefficient (the correlation of the predictor with subsequent job performance) and the financial impact on the organization of having good versus only average performers (Tziner et al., 1994). The latter depends on the organization and the target job. Greater utility will be realized when selecting candidates for positions with a larger impact on the organization's bottom line, but this value is generally fixed for any one selection program. Consequently, the validity coefficient is often of primary interest in choosing a selection system. Given that other selection methods (e.g., cognitive ability tests) often display superior validity coefficients with much less expense, can we ever expect an AC to demonstrate utility relative to these other methods? If we account for the necessity of considering factors such as adverse impact, the answer appears to be yes.

Selection techniques that produce relatively smaller differences between majority and minority group mean scores allow for the use of a smaller selection ratio; that is, the organization can select a smaller group of applicants while remaining within legal guidelines regarding adverse impact. This smaller selection ratio magnifies the impact of the predictor, resulting in higher utility (Hoffman & Thornton, 1997; Thornton, Murphy, Everest, & Hoffman, 2000). Even though, as discussed earlier, ACs cannot be assumed to be free of adverse impact (Dean et al., 2008), producing less adverse impact compared to another predictor can lead to greater overall utility, even if the alternative predictor has a higher validity coefficient.

4. Controversies

In the next two sections we discuss controversies and future research needs related to the use of ACs for selection. Controversies include areas which show mixed findings in past research or conflicting recommendations for practice. Future research needs include studies to not only address controversies but also to investigate topics which have not been examined empirically.

4.1. Structure: dimensions or exercises?

We have discussed the long-running debate regarding the internal structure of AC ratings and its meaning for the construct validity of ACs. The distinction between dimensions and exercises bears repeating. In this context dimensions are the attributes being measured; exercises are the methods used to measure the attributes. As summarized earlier, studies of correlations of ratings of multiple dimensions in multiple exercises have frequently found strong exercise effects and weaker dimension effects. Such findings have led some to conclude that the use of dimensions in ACs should be abandoned (e.g., Lance, 2008a; Sackett & Dreher, 1984). As an alternative, these authors recommend viewing exercises as work samples and assigning overall performance scores for each exercise, rather than attempting to rate distinct dimensions within exercises. Well-designed work samples are generally effective predictors (Roth, Bobko, & McFarland, 2005; Schmidt & Hunter, 1998), so it is reasonable to expect such an approach to yield useful predictive information. In terms of the predictive efficacy of the overall assessment rating, there may be little practical difference between dimension-based and exercise-based ACs. We will return to and extend this line of thinking with a proposal in the section on research needs.

Beyond the realm of prediction, however, the recommendation to remove dimensions from ACs raises three potential concerns. First, dimensions are not exclusive to the AC method, but are widely used in other human resource management functions as well (e.g., performance appraisal; Rupp, Thornton, & Gibbons, 2008). On a practical level, using dimensions facilitates the integration of the AC process with other organizational processes, such as training and performance improvement efforts. On a more theoretical level, one reason that dimensions (or competencies) are so often used as the language of human resource management is that job performance is widely acknowledged to be a multidimensional construct (e.g., Campbell, McCloy, Oppler, & Sager, 1993; Tett, Guterman, Bleier, & Murphy, 2000). Most studies of the dimensionality of performance examine performance as a whole, over some period of time, rather than performance on individual tasks. It is therefore difficult to say whether performance on a given task requires multiple dimensions or whether each task requires only a single dimension, with the aggregation of different tasks over time giving rise to the appearance of multidimensionality in overall performance. The evidence presented by Rupp et al. (2006) and Hoffman and Meade (2007) for distinct dimensions within exercises suggests the former, and certainly it is intuitive to suspect, for example, that effectively conducting a meeting with subordinates might require skill in both planning and communication. However, the existing research in this area is sparse, and future work is certainly needed to resolve this issue, both for ACs and for the broader field of job performance in general.

A second significant concern about the recommendation to remove dimensions from ACs is whether an exercise-based AC can fully capture the complexity of the target job. For a job that largely consists of repeatedly handling a few common situations (e.g., a human resource associate who conducts interviews and prepares candidate summaries), it is relatively easy to create a set of AC exercises that capture the essential tasks of the position. We can be fairly confident in making inferences about future job performance from this set of tasks, because the exercises represent most of the universe of tasks to which we wish to generalize. Exercise-based ACs may be quite appropriate. By contrast, for a more complex job that requires the frequent handling of novel situations and tasks (e.g., a district manager who oversees diverse work functions), the set of possible job-relevant exercises is much larger. Given the time constraints of a typical AC, we must choose a subset of these possible tasks, with the goal of generalizing to a broader task universe in which some of the tasks are unknown.

The need to generalize to a wide variety of possibly unknown situations appears to have been part of the original impetus for using dimensions in ACs. The first AC in the U.S., Station S, was used by the Office of Strategic Services in World War II to select agents for overseas intelligence assignments (Fiske, Hanfmann, MacKinnon, Miller, & Murray, 1948). The developers of Station S noted the near impossibility of writing job descriptions for these positions, due to the heterogeneity of possible assignments and the frequency with which needs changed in the field. As a result, the Station S assessment staff chose to focus on attributes of the persons being assessed rather than on their performance in specific tasks. Having a description of an individual's abilities and personality would make it easier to determine his or her fit to a new job assignment that might require entirely novel activities. In today's rapidly changing workplace, this argument still appears reasonable.

A third area of concern regarding the recommendation to use exercises as the basis for AC design is that of feedback. For many external selection programs, participants are informed only of their success or failure and feedback is not an issue. In promotional ACs, however, it is not uncommon to offer feedback in the hopes that unsuccessful candidates can remedy their deficiencies. Critics of the dimension-based approach argue that giving dimension-based feedback is misleading if the AC administrators cannot be certain that the dimension ratings are construct valid (Sackett & Dreher, 1984). Again, however, the use of exercise-based ratings raises questions of

generality: if a participant receives feedback about his or her performance in a group discussion task, will he or she be able to transfer that information to a variety of group settings in the work place, or will he or she merely have an advantage in the next AC? This is an empirical question in urgent need of an answer. To our knowledge, the only direct comparison of dimension- versus exercise-based feedback was conducted by Larsh (2001). She found no effect of feedback format on subsequent performance, and results regarding participants' responses to the feedback were mixed: exercise-based feedback had greater effects on participants' self-efficacy, with participants who were initially low in self-efficacy showing increases and those who were initially high showing decreases. However, Larsh's feedback conditions were somewhat confounded with content: the dimension-based feedback was framed in explicit trait terms (e.g., "you showed leadership"), while the exercise-based feedback was framed in behavioral terms (e.g., "you led the group"). Audience comments at a recent symposium on ACs at the meeting of the Society for Industrial and Organizational Psychology (Hoffman, 2008) suggest that ACs in practice often resolve the conundrum by giving feedback organized both ways: either by dimensions within exercises or by exercises within dimensions. There appears to be an urgent need for empirical research regarding the effectiveness of various feedback formats or even candidates' preferences for feedback.

Ultimately, perhaps, the decision to structure ACs around dimensions or exercises may be analogous to the distinction between task-oriented and person-oriented job analysis. Both may be appropriate, depending on one's purpose and on the nature of the target job. For example, exercise-based ACs may be most useful when selecting for a specific, well-defined position; when the job consists of a small set of repeated tasks; or when there is cause for concern about assessors' ability to make accurate judgments about multiple dimensions. By contrast, dimension-based ACs may be most useful when the goal of the program is to identify "high potential" candidates for a variety of future positions; when the organization is experiencing significant change, requiring adaptation and flexibility; or when the results of the AC will be integrated into a broader, competency-based HR management system. As yet, however, such distinctions are difficult to determine. Although exercise-based ACs do exist in practice (Brink et al., 2008), there have been few published studies examining their validity or comparing them directly to more traditional dimension-based ACs.

4.2. Transparency: asset or liability in selection settings?

Another question that has attracted considerable research interest in recent years concerns the degree to which the dimensions assessed in an AC should be explained, or made transparent, to the participants. Both positive and negative effects may result from making dimensions more apparent to selection candidates. Researchers and practitioners working in selection settings must consider the issue of test security. For example, in ability testing, a common concern is that some applicants might obtain advance knowledge of test items and therefore practice or memorize appropriate answers (Do & Brummel, 2006). In personality testing, many researchers have suggested that when the personality traits being assessed are transparent, applicants may "fake good," or intentionally distort their responses so as to appear more attractive to the organization (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). In either of these cases, the predictive validity of the assessment is compromised, because applicants are not presenting themselves as they actually are. In other types of selection settings, however, such as structured interviews, informing applicants about what to expect may actually improve validity. In this case, transparency is believed to reduce the influence of irrelevant factors, such as nervousness or misunderstanding instructions, that might distort candidates' normal responses, thus allowing them to display relevant knowledge or behavior to the best of their ability (Maurer, Solamon, & Lippstreu, 2008).

Which of these phenomena operates in the case of ACs? Does making dimensions transparent allow participants to manipulate their assessments, thereby lowering validity, or does it encourage participants to demonstrate more relevant behavior, thereby allowing for a more accurate assessment? Recent research in this area suggests a complex picture. Indeed, the degree to which individual candidates' AC ratings are influenced by dimension transparency at all is unclear. Kleinmann (1993) proposed that candidates who could correctly identify which dimensions were being assessed in each exercise would have such an advantage over other candidates, as they could then be sure to display behavior relevant to those dimensions. He found that candidates who correctly perceived that a dimension was being rated received higher scores on that dimension than those who failed to perceive it; however, candidates were generally poor at identifying dimensions. Similarly, Smith-Jentsch (2007) found that participants for whom a targeted dimension was made transparent received higher ratings on that dimension than those for whom it was not made transparent, but observed this result in only one of two studies. Other studies found no differences in mean ratings between participants who were told which dimensions were being assessed and participants who were not given this information (Denning & Grant, 1979; Kolk, Born, & van der Flier, 2003).

Studies of faking in the personality domain often note the difference between ability to fake and motivation to fake (Viswesvaran & Ones, 1999). In this light, it is interesting to note that only one of the AC studies described above used a sample of job applicants (Kolk et al., 2003, Study 2), who might be expected to be more motivated than student volunteers, yet this study found no score improvement as a result of transparency. Further, both the Smith-Jentsch and the Kolk et al. studies provided participants with examples of effective and ineffective performance, which the participants could perhaps have used to improve their own performance, but score improvement appeared in only one of the four studies described by these two sets of researchers (i.e., Smith-Jentsch, 2007, Study 1). Further research is clearly needed to determine the degree to which participants can and do adjust their behavior in response to transparency cues of various kinds.

Many of the studies described above also considered the effects of transparency on the internal structure of the AC ratings; that is, the degree to which those ratings correspond to the MTMM framework. These results address the second possible effect of transparency: that it allows candidates to display more relevant behavior, removing irrelevant variance due to candidates' attempts to guess what is expected of them. By this logic, candidates' ratings on the same dimension across exercises should be more highly correlated when dimensions are transparent than when they are not, because the participant knows to demonstrate the relevant

behavior in those exercises. Indeed, Kleinmann (1993) found that correlations between same-dimension ratings (e.g., controlling group processes) in two different exercises were higher for candidates who perceived the same dimension in both exercises than for candidates who perceived the dimension in only one exercise. Interestingly, correlations were also high for candidates who did not perceive the dimension in either exercise. This suggests that the candidates who perceived the dimension in one exercise oriented their behavior toward the dimension in that exercise, but not in the other, leading to inconsistent ratings. Kleinmann, Kuptsch, and Koller (1996) found better MTMM patterns when the dimensions were made transparent than when they were not, but this was only true for participants who reported deliberately orienting their behavior to the dimensions. Kolk et al. (2003) found no real differences in MTMM correlation patterns between transparent and non-transparent conditions in a student AC; however, in a sample of job applicants, the MTMM-based factor model fit much better in the transparent than in the non-transparent condition.

These findings have led several of the above researchers to conclude that transparency improves the construct validity of ACs. As we have already noted, however, construct validity and a good fit to the MTMM pattern are not synonymous. Smith-Jentsch (2007) took a different approach to the question of the effect of transparency on construct validity by correlating participants' AC ratings on a specific dimension of assertiveness with their self-ratings on that dimension and a different, less relevant dimension. When the targeted dimension was not made transparent to participants, the AC ratings showed a clear pattern of convergent and discriminant relationships with the self-ratings; AC ratings were correlated with self-ratings of the target dimension but not with the other dimension. When the targeted dimension was made transparent, however, AC ratings were poorly related to both self-reported dimensions. This suggests that the transparency manipulation actually decreased construct validity in the broader sense of convergent and discriminant relations with other variables. Similarly, Kuptsch, Kleinmann, and Koller (1998) used performance in one non-transparent AC to predict performance in a second AC where transparency was manipulated. The correlation between ratings from the two ACs was weaker for the transparent group. Again, making dimensions transparent reduced the degree to which AC ratings converged with other measures of the same construct. Further research is needed to replicate these findings and to determine the impact of transparency on the validity of ACs for predicting job performance.

Fairness is also a concern in determining whether to make dimensions transparent. Dodd (1977) found that a substantial majority of candidates believed that others had access to inside information about the AC, and Kleinmann (1993) showed that candidates differed in their ability to discern targeted dimensions from the exercise materials alone. Dodd suggests that making dimensions transparent can effectively level the playing field in this respect by giving all candidates access to the same information. This recommendation is in keeping with the standards for psychological testing in general (American Educational Research Association et al., 1999), which state that a test should be equally familiar or unfamiliar to all participants, and transparency is often recommended for developmental ACs (Thornton & Rupp, 2006). However, there is not yet clear evidence regarding whether or not such disclosure may affect the validity of the assessment. In fact, Smith-Jentsch (2007) argued that the ability to determine what is expected of one may actually be a job-relevant skill, and that making dimensions transparent would actually reduce this relevant source of variance in ratings. As described above, the mixed evidence regarding the effects of transparency on score inflation and construct validity makes it difficult to determine whether or under what conditions transparency is beneficial in selection settings.

4.3. ACs as maximal versus typical performance measures

A large part of the appeal of the AC method lies in the fidelity of the exercises to the target job; that is, the degree to which the AC resembles the situations that would typically be encountered by an incumbent in that position. But does this mean that AC ratings indicate how a candidate will typically perform? Early advocates of the method emphasized that ACs would allow an organization to see how a candidate would function in an upper management position before he or she was actually promoted (Byham, 1970), implying assessment of typical performance. More recently, however, several researchers have argued that ACs are more appropriately viewed as examples of maximal performance. Maximum performance represents what the candidate is capable of doing, given opportunity and sufficient motivation. Sackett, Zedeck, and Fogli (1988) commented that ACs are difficult to classify as either typical or maximal performance measures. On the one hand, an external selection or promotion AC is clearly an evaluative context and candidates are motivated to do their best. On the other hand, ACs are often long and complex, and it may be difficult for candidates to sustain maximal performance across a series of several exercises. Ployhart, Lim, and Chan (2001) took a different approach, arguing that a 2-day AC was a relatively short time frame, in which participants might be expected to remain focused, and that the AC therefore corresponded very well to the criteria for maximal performance measures set out by Sackett et al.

One piece of evidence on the side of considering ACs as maximal performance is that AC ratings are more strongly correlated with ability than with personality traits, which is both theoretically expected and often reported for maximal performance measures (Marcus, Goffin, Johnston, & Rothstein, 2007), i.e., what the candidate is capable of doing. This relationship is expected because maximal performance represents what the candidate is capable of doing, given opportunity and sufficient motivation. A candidate who is higher in cognitive ability is more likely able to discern what behaviors will be most effective and implement those behaviors, even if they are not the behaviors the candidate would naturally prefer. By contrast, personality variables (e.g., conscientiousness) are expected to reflect a candidate's habitual work and interaction styles, and therefore to be better predictors of typical than maximal performance. Reasoning along similar lines, Smith-Jentsch (2007) suggested that making dimensions transparent to participants might transform an AC from a typical performance measure to a maximal one: when participants know what they are expected to do, the AC is evaluating only whether or not they are able to do it (maximal performance). When participants are given less direction, she argued, they are more likely to behave in habitual and typical ways.

Although there are reasonable arguments on both sides of this question, empirical studies are lacking. The evidence provided by Ployhart et al. (2001) and Marcus et al. (2007) regarding the correlates of AC ratings is promising, but somewhat indirect. Dynamic investigations of participants' motivation and self-perceived performance within ACs of varying lengths would more fully resolve the question posed by Sackett et al. (1988) about whether candidates could sustain maximal performance for the duration of the AC. It also seems quite possible that whether an AC measures typical or maximal performance might vary as a result of the conditions of the AC: for example, whether the purpose of the AC is selection or development; the transparency of the dimensions (Smith-Jentsch, 2007); the use of group or individual exercises, and so on. The type of dimensions (or other constructs) assessed may also have an impact. In some ACs, the dimensions rated are described in terms very similar to personality traits (e.g., drive, interpersonal style). In others, the dimensions are described in more behavioral terms (e.g, influencing others). Personality-style dimensions may encourage assessors to draw broader, more personality-oriented inferences regarding the candidates; across the exercises, what does the candidate's pattern of behavior suggest about him or her as a person? Such dimensions seem more aligned with, and may be more related to, typical performance. By contrast, behavior-style dimensions may encourage assessors to emphasize successful or unsuccessful completion of tasks, which may correspond better with maximal performance. Task- or exercise-based ACs (Lance, 2008a) may be even further along the maximal performance continuum. Further research in this area is likely to be highly valuable for AC practitioners, allowing for better alignment between predictors and criteria and improved understanding of just what ACs measure.

4.4. Integration methods

One of the long-standing controversies involves the question of whether consensus discussion, statistical integration, or some combination of the two should be used to integrate preliminary assessment observations and ratings. Over the years, a variety of integration methods have been used and evaluated. Full descriptions of several integration methods are provided in Thornton and Rupp (2006). The traditional AC method calls for individual assessors to make observations of behaviors, and possibly give ratings, related to dimensions in individual simulation exercises. These observations and ratings are then presented and discussed by assessors in the integration discussion (sometimes called the "wrap up session"). After hearing evidence, assessors typically make independent preliminary overall dimension ratings, and then discuss differences until they reach agreement on the dimension ratings. And, assessors make preliminary overall assessment ratings, and discuss these to achieve consensus. As summarized earlier in this paper, research has found mixed evidence whether the consensus discussion or statistical methods of integration result in higher correlation of the OAR and performance criteria.

Over the years, a variety of methods have been devised to statistically integrate ratings provided at various stages of the above process. At one extreme, multiple assessors observe behavior in an exercise and make independent ratings on several dimensions, but have no follow-up discussion. These ratings are then statistically averaged, first across assessors for each dimension in each exercise, then across dimensions in several exercises, and finally across dimensions to yield the overall assessment ratings. Variations on this approach include the provision for, and even requirement that, assessors discuss discrepant ratings (say, greater than one point difference on a five-point scale) along this chain. As summarized earlier, research has shown that statistical combinations of initial ratings correlate substantially with overall assessment ratings derived from the consensus method. Theory and research evidence supporting the validity of OARs derived from consensus ratings are reviewed in Thornton and Rupp (2006).

A significant limitation of research in this area is that data have come from ACs in which ratings have been made by all assessors at multiple steps in the process, so that the cognitive process for assessors is the same regardless of how the ratings are ultimately combined. This does not allow true comparisons with methods in which ratings are not made until the final steps. Assessors process information differently when they know they will be held accountable for their judgments (e.g., ratings; Mero & Motowidlo, 1995). What is needed are studies with multiple assessment conditions, in which some raters make ratings and others do not at each stage of the process.

5. Research needs

We first summarize research to address the controversies described in the previous sections, then propose research to investigate emerging issues related to the use of ACs for personnel selection. The controversy regarding whether ACs should be structured around dimensions or alternative constructs can be addressed more meaningfully with studies of ACs that are designed and implemented using such alternative constructs (e.g., exercises, roles, tasks, etc.) from the beginning, rather than the common practice of reanalyzing ratings from a dimension-based AC. Studies directly comparing dimension- and exercise-based ACs are also needed, as are studies that consider outcomes beyond the MTMM framework. Such studies could then address the conditions and purposes for which it is better to structure ACs around dimensions or exercises.

The controversy over whether to foster transparency could be addressed by further studies of its effects on different outcomes such as predictive accuracy, fairness, user reactions, and construct validity in the broad sense of relationships with external variables. The question of whether ACs measure maximal or typical performance needs to be addressed more directly, using appropriate performance criteria such as supervisor and peer ratings of various types of behavior and objective performance data, and including comparisons of ACs with different types of instructions. The length of time for which candidates can sustain focus and motivation should be examined empirically, rather than simply assumed.

The controversy over integration methods could be addressed with AC designs in which different groups of assessors follow different integration procedures (e.g., consensus discussion or statistical formulae) for the same group of assessees. Another

research need stems from our earlier observation that the reported criterion-related validity of ACs (as evidenced in individual studies and meta-analyses) has declined over the past decades. This ostensible decline may be due to lower quality ACs, less rigorous research designs, different samples, one or both of the publication biases, or the possibility that the ACM is no longer applicable to modern work, organizations, or human resource systems.

We now propose research in areas that have not been investigated, beginning with a proposition that integrates a number of important issues embedded in the use of ACs for selection. Then we identify several other more specific research topics.

5.1. General ability: an integrative proposition

We have already alluded to several aspects of the controversy regarding the use of behavioral dimensions as a framework for building ACs and for making AC ratings. We have discussed methodological (MTMM analyses), conceptual (criterion multidimensionality and job complexity), and practical (feedback) issues related to the use of dimensions, but a broader concern still remains: how can we be confident that ACs measure what they claim to measure? Arthur, Day, and Woehr (2008) note that AC researchers and practitioners often conflate evidence for the validity of the AC method with evidence that the constructs measured in a given AC are also valid. As mentioned previously, different ACs may measure widely different sets of dimensions, and even dimensions with the same name may be defined and measured very differently. Arthur et al. points out that there is seldom any empirical verification that AC dimensions are valid measures of the specific constructs they are intended to assess. For diagnostic and development purposes, of course, careful validation of the individual dimensions is essential, because decisions about further development efforts rely on the assumption that dimension scores are accurate measures of the intended constructs.

For selection purposes, however, AC researchers have historically been less concerned with the quality of dimension measurement, as the focus in selection applications is generally on the OAR. Arthur and Villado (2008) argue that this is a serious oversight, as the meaning of the OAR is dependent on the particular set of constructs assessed in the AC. An OAR from a center assessing problem solving and planning is not at all the same as the OAR from a center assessing interpersonal communication and teamwork, because they are based on different abilities and different behaviors. They go on to argue that the choice of dimensions is likely to affect the predictive validity of the AC (as some dimensions may be easier to measure well than others), the likelihood of observing subgroup differences (as groups may differ on some constructs but not others), and other important outcomes. Consequently, it would be inappropriate to treat OARs based on different constructs as equivalent or interchangeable; yet this is often done in the research literature.

Although we agree with Arthur and Villado (2008) about the importance of disentangling constructs and methods, and of carefully validating dimension measurements, we wish to offer a somewhat different perspective on the meaning of the OAR in selection contexts. A reconciliation of a number of controversies and a template for future research may reside in the extension of two classic principles of ability measurement to the AC context, namely "*g*" and "the indifference of the indicator" (Spearman, 1904, 1923/1973, 1927). Spearman proposed the presence of a general intellectual factor, designated "g," that runs through many measures of cognitive abilities. Spearman's (1904) Principle of Unity of Intellectual Function posits that "all branches of intellectual activity have in common one fundamental function (p. 85)." This ability has been labeled "intelligence," "general mental ability (GMA)," or "IQ." GMA is conceived to be a complex set of related mental abilities, the core of which is reasoning with words, numbers, and figures. GMA is typically measured with a single score resulting from the unweighted composite of several items measuring each of several components (e.g., verbal, quantitative, reasoning, spatial, perceptual, and other abilities), all subsumed under the general heading of GMA.

In the principle of "the indifference of the indicator," Spearman (1927) posited that mental ability can be measured with a variety of stimulus material: "for the purpose of indicating the amount of g possessed by a person, any test will do just as well as any other, provided that its correlation with g is equally high (p. 197)." In other words, if two tests are equally valid measures of g, both will yield the same result, even if they are composed of very different items. Research has shown that many ability tests have a large "*g*" component and are relevant indicators. Jensen (1992) provided a recent summary of Spearman's concepts. Measures of "g" following these principles have been found to relate to a wide variety of intellectual, academic, work, and life functions.

In an analogous way, when an AC is used for personnel selection, we can conceive of a "general ability" underlying performance in the target job(s). For managerial jobs we might describe a "general management ability" composed of several dimensions. Performance dimensions are likely to be correlated (cf. Viswesvaran & Ones, 2000); certainly, a high performer will likely receive high scores on most or all dimensions. These intercorrelations suggest a possible common factor: g_{mgmt}, a general ability underlying performance in the family of management positions that require similar skills. Job analysis is used to understand the components (dimensions) of this general factor, just as content analysis is used to specify the components of GMA. In a managerial AC, the overall assessment rating is composed of observations by multiple assessors (say 3) of several behaviors (say 5) displayed in multiple exercises (say 4) relevant to the dimensions (say 5). Thus, the OAR reflects g_{mgmt} and is conceived to be a 300 item test ($3\times5\times4\times5=300$).

In ACs for other positions, there are conceptual equivalents, such as general sales ability, general teaching ability, general customer service ability, etc. In each example, the construct is a general ability made up of several components; the OAR is an omnibus measure consisting of multiple observations of behavior tapping different components. The behaviors that are elicited from the participants in assessment exercises are indicators of the processes that are important in the performance domain. In mental ability testing, according to Spearman's principle, any number of possible items could be written to measure the process of interest. For example, reasoning with words could be tapped with items composed of different sets of words. Analogously, in the AC, ability to plan could be tapped with AC exercises depicting very different issues within an organization.

These concepts of what an AC measures when the method is used for selection do not obviate the need for careful job analysis, validation of dimensions, or systematic preparation of assessment materials. Not just any attributes and not just any exercise material will do. Job analysis/competency modeling is necessary to specify the domain of job tasks and performance attributes that

are essential for effective performance. If the challenges in the assessment exercises and the dimensions or tasks to be evaluated are not a sample of the job domain, the AC will not be valid.

These notions may help resolve many of the controversies embedded in the discussion of the validity of ACs when they are used for personnel selection. The "*g*" theory of ACs would obviate the need to resolve whether ACs measure either abilities or personality characteristics; both might be appropriate indicators of the general ability being assessed. It would be expected that different dimensions would have some moderate correlation, and that all dimensions identified in a job analysis would be reflective to some extent of the general ability. The principle of "indifference of the indicator" would obviate the controversy over whether dimensions or tasks are the currency of ACs, because what is really important is the overall set of complex behaviors that are elicited, observed, and evaluated. These behaviors are the "items" of the test and the overall AC score is indifferent to the specific behaviors that are elicited. Furthermore, the debate over convergent and discriminant validity of post-exercise dimension ratings is rendered moot, because the outcome of interest is the total test score or OAR. If a dimension-based AC and an exercise-based AC yield similar validity coefficients for the same job, we might infer that both measure "g" for the job in question. It may not matter whether a particular set of dimensions (Thoresen, 2002), roles (Russell, 1987) or exercises (Lance, 2008a) were the vehicle to obtain that estimate of "*g*."

In summary, when the ACM is used for personnel selection, the organization wishes to select individuals who will be successful in future assignments. Sometimes the future assignments are not clearly understood; for example, because the jobs and organizations may be changing rapidly. We propose that in each selection context, the organization is striving to measure a "general ability" permeating the requirements in the target job or set of jobs. Research into the efficacy of this proposition and whether it enhances the use of the ACM for personnel selection is needed.

5.2. Assessee training

With the exception of the transparency research reviewed above, the bulk of AC accuracy studies have focused on the role of assessors and their ability to make accurate ratings. Far less attention has been given to the question of whether assessees can manipulate their own assessments. For example, virtually no studies have been conducted to evaluate the efficacy of various techniques to train candidates to perform in ACs. Such techniques include orientation programs typically offered by public jurisdictions before promotional exams, books on how to prepare for an AC (e.g., Page, 1995; Rowe, 2006), and one- to three-day training programs offered by various consultants (e.g., PoliceCareer.com; PolicePromotion.com). It is not at all clear whether these programs have any effect on candidates performance, or whether such an effect might be negative (e.g., reducing validity) or positive (e.g., reducing nervousness and other distractions; cf. Maurer et al., 2008). Future research should investigate the effects of such programs on various outcomes, including participants' confidence entering an AC, performance on different dimensions assessed (e.g., administrative skills vs. basic decision making abilities vs. interpersonal relations vs. personality-type dimensions), and efficacy for candidates with different demographic and work backgrounds.

5.3. Reusing exercises

A related area begging for research is the question of the advantages and disadvantages of reusing exercises in subsequent selection and promotional exams. Building new exercises is time consuming and organizations may wish to reuse well developed and validated exercises. However, information is likely to "leak out" over time about the content of exercises, which may give later participants an advantage. The concern here is with pre-knowledge of exercise content, which is somewhat different from the question of whether or not dimensions should be made transparent to participants (as discussed above). Transparency of dimensions allows a candidate to orient behavior toward those dimensions; exercise pre-knowledge might allow a candidate to prepare possible response strategies in advance. Research could address the effects of different levels of knowledge about exercise content on the performance of AC participants on different types of dimensions. Research may reveal that performance on administrative dimensions such as delegation may be enhanced but performance on basic aptitude dimensions such as decision making abilities or leadership skill may not be enhanced by prior knowledge.

5.4. Standardization

Lack of standardization, whether real or perceived, has not been studied, but is critical in high stakes selection and promotional examinations, especially in contentious settings such as police and fire jurisdictions where administrative and legal challenges abound. Lack of standardization can lead to challenges on the basis of equal employment opportunity laws and public charters guaranteeing fair treatment and promotions based on merit and fitness. Exercises that may be vulnerable to allegations of lack of standardization include leaderless group discussions (e.g., group composition may differ in terms of fellow participants who are talkative or reticent, more or less experienced, similar or different in race, gender, or age and exercises) and exercises involving role players with whom the candidates interact (e.g., any single role player may not portray the "problem employee" consistently over time or different role players may portray a different "problem employee" to different candidates). The composition of the group of candidates participating in a group exercises (e.g., leaderless group discussions, games) may not be standardized. It is reasonable to question whether differences in the mix of gender, race, age, etc. may affect behavior of participants and ratings of assessors, but only one study provides empirical evidence. Schmitt and Hill (1977) found that sex and race composition of assessee groups had little effects on self, peer, or assessor ratings. Perceptions of candidates, potential candidates, managers, and other stakeholders

(Caldwell, Gruys, & Thornton, 2003) regarding the standardization of ACs, and the actual effects of lack of standardization on behavior and outcomes warrant investigation.

5.5. Impression management

The preceding discussion about transparency and maximal versus typical performance raises an important question: Is the outcome of an AC influenced by impression management techniques by the candidate? Impression management is a process in which people try to influence the impression other people have of them (Schlenker, 1980). It would be informative to understand the types of impression management techniques candidates use, the extent to which assessors recognize these techniques, and whether there are discernable positive or negative relationships between display of these techniques, assessor ratings, and performance criteria, as well as whether the use of impression management moderates the relationship between assessment ratings and performance criteria. That is, can participants distort their behavior to such an extent as to give an entirely misleading impression of their likely performance on the job?

We have already discussed the question of whether participants adjust their behavior to display what they think is expected (e.g., Kleinmann, 1993; Kolk et al., 2003). Here, we consider the question of whether AC candidates employ intentional strategies to present themselves in a more favorable light. This question has not been extensively studied, but the evidence that exists suggests that further research attention is needed. McFarland, Yun, Harold, Viera, and Moore (2005) found that AC candidates did engage in impression management tactics, and that these tactics affected assessor ratings, even on non-interpersonal dimensions (where the possible job-relevance of impression management behavior is much less). If candidates can improve their ratings on all dimensions through impression management, the validity of the overall assessment (and of the non-interpersonal dimension ratings in particular) is compromised. However, the fact that some participants use such tactics does not necessarily mean that they do so successfully. Kuptsch et al. (1998) found that candidates who were high in self-monitoring received more consistent ratings from exercise to exercise, but that their ratings overall were no higher than those of candidates low in self-monitoring.

To our knowledge, there are very few empirical studies evaluating the prevalence and effectiveness of such behaviors, nor are there proposals for how such faking might be detected and addressed (e.g., via assessor training). Considering ACs as maximal performance measures removes some concern about faking, as it is not expected that typical job performance will match the performance levels seen in the AC. However, this perspective requires the assumption that a candidate's high performance truly reflects his or her ability–that he or she is capable of highly skilled behaviors, even if he or she does not engage in them all the time–and is not simply an artifact of self-presentation techniques. Research is needed to test such assumptions and establish the degree to which faking behavior affects the results of operational ACs.

5.6. Building exercises (including instructions) to elicit behavior

Another promising line of research would use trait activation theory (TAT; Tett et al., 2000) to guide the construction of content and instructions of exercises. TAT argues that different situations differ in the degree to which they "activate" various traits; that is, the degree to which the trait determines behavior in that situation. This argument is loosely analogous to the idea of strong versus weak situations (Mischel, 1973). In the former, situational cues exert a strong influence, and there is little variance in behavior. In the latter, situational cues do not exert much influence, and behavior is determined to a much greater extent by the personality of the individual. For example, a job interview presents very strong situational cues about what behavior is expected and appropriate, whereas a lunch date with a friend does not. We would expect behavior to vary much more from person to person in the latter situation than in the former. TAT goes even further by arguing that different situations suppress or enable the expression of particular traits; that is, some situations offer strong cues about the expected degree of, say, extraversion, while other situations do not. The latter set of situations are said to activate the trait of extraversion, because they allow more opportunity for the trait of extraversion to be displayed.

Haaland and Christiansen (2002) have made the argument that trait activation is an essential concept for AC design. If an exercise does not sufficiently activate a particular trait (or dimension), there will be little variance in behavior, and therefore little variance in ratings. This lack of variance, then, reduces the possible correlation between ratings of the same dimension in different exercises (convergent validity, in the MTMM framework). In support of their argument, Haaland and Christiansen found higher correlations among dimension ratings from exercises that were rated by experts as having greater potential to activate those particular dimensions. Subsequent studies have yielded similar results (Bush, 2004; Lievens, Chasteen, Day, & Christiansen, 2004) and Lievens (2007) continues to study this area.

In a sense, there is nothing really new in these arguments. AC researchers have argued for years that not every dimension can be observed in every exercise, and that exercises must be intentionally designed to elicit relevant behavior (Thornton, 1992). However, it is not clear that this recommendation is consistently followed in practice. Further, determining the appropriateness of particular exercises for assessing particular dimensions is typically a matter of expert judgment. If TAT research can provide a more theoretical framework for aligning exercises and dimensions, it will provide a significant service to the AC community. Future empirical studies of experts' ratings of trait activation may be helpful, but studies that directly examine the degree to which exercise characteristics produce variation in assessee behavior may be even more valuable.

Different types of exercise instructions may have different effects on candidate performance. Speculation has suggested that whereas a simple instruction prior to a case study such as "Analyze the situation and make a recommendation" yields a measure of the participant's propensity to engage in systematic thinking, a more complex and even prescriptive instruction such as "Specify the problem,

list possible causes of the problem, state the most probable cause, list 3 alternative solutions with costs and benefits of each, then state your judgment of the best balanced decision" yields a measure of the person's ability to make good decisions. Studies considering such questions are clearly useful both for exercise design and for addressing the typical vs. maximal performance controversy.

5.7. Effects of retesting

Another topic in need of research is the effect of retesting with the AC method. In several situations, candidates may participate in more than one AC over time. Candidates for promotion in police and fire departments may not be promoted during the typical two-year life of the eligibility list, and thus subsequently retest for the same position in a parallel AC. Or, a successful officer candidate who gets promoted to sergeant may participate in an AC for promotion to lieutenant or captain.

During their careers managers in large organizations may participate in multiple ACs used for promotion, developmental planning, leadership training, and succession planning. The effects of being re-assessed in these different patterns have not been studied. Two research designs would yield valuable information. A with-in person design would examine the effects of a group retaking parallel ACs. A between-person design would examine how re-takers (i.e., one group of persons taking the AC a second time) compared with first-timers (i.e., persons taking the AC the first time). Similarly, to our knowledge, there has been no research on the effects of serving as assessor at one point in time on that person's performance as a candidate in a subsequent AC. Presumably, service as an assessor helps a person later to be a better assessee; this is one reason that some police and fire managers are willing to serve as assessors for lower level promotional exams in brother/sister jurisdictions.

5.8. Additional dependent variables

In general, a wider array of dependent variables needs to be studied in the process of accumulating evidence related to the validity of the ACM for selection. Certainly past research has focused on relevant evidence, as reviewed in prior sections of this paper. In addition, other fertile fields of evidence have not been explored. Studies of the trade-off of predictive accuracy, differential validity, and adverse impact as a function of design features in the AC would be valuable for both research insights and practical guidance. Studies of reactions of various stakeholders (Caldwell et al., 2003) beyond assessees and assessors would be relatively easy and highly informative to the human resource community and executives.

5.9. Response processes

Understanding of the validity of the assessment center method would certainly benefit from studies of the response processes (American Educational Research Association et al., 1999) of participants and assessors. What are participants thinking about when they respond to the array of items in an in-basket? Are they taking items one-by-one in methodical order or scanning all items to see inter-relationships? Are they thinking about general objectives for influencing subordinates or protecting higher management interests? It is common practice for assessors to ask questions like these after assessees complete written responses to an in-basket, but to our knowledge no research has systematically studied these introspections with, say, standardized qualitative research methods. In a similar fashion, what are assessors thinking about as they observe a group interaction? Are they mechanically recording discrete behaviors, seeking out positive (or negative) instances of performance, monitoring their own gender or racial prejudices?

5.10. Assessee inconsistency

In our review of the literature regarding the psychometric structure of ACs, we described recent research suggesting that tendency for candidates to receive different ratings in different exercises may reflect genuine inconsistency in performance, not simply assessor error (Lance, Newbolt, et al., 2000; Lievens, 2002; Hoffman & Meade, 2007; Rupp et al., 2006). The reasons for this inconsistency, however, are unclear. Lievens (2002) suggests that it may be a function of instructions or other features of the AC; Kuptsch et al. (1998) suggest that it may be the result of individual differences in self-monitoring. Gibbons and Rupp (in press) proposed that candidates might vary in consistency, and that this consistency might be job-related. In an earlier study, a quantitative index of consistency predicted candidates' job performance ratings even when controlling for the OAR (Gibbons & Rupp, 2007). If consistency or the lack thereof is related to organizational outcomes, it is important to understand what it is and whether and how it should be considered in the selection process.

5.11. Technology

AC designers have long sought strategies to make their assessments more efficient and less expensive. The growing prevalence of high-powered computers, inexpensive telecommunications, and sophisticated software is driving a great deal of innovation in operational ACs. Computerized in-baskets, simulating an employee's email and access to web resources, are becoming commonplace. Some ACs are conducted entirely over the telephone, using role players in remote locations to simulate the kinds of tasks that might be encountered in a call center or similar setting (Gowing, 2006). Digital video technology is used to connect assessors and candidates in different parts of the world, or to allow asynchronous assessment (Rupp, 2006). Artificial intelligence and video game technology have allowed some AC practitioners to offer entirely computer-based simulations, in which the candidate interacts not with peers or role

players but with digital characters (or avatars) in a virtual world. Although these technologies are expensive to develop, they can reduce or eliminate many of the operational costs of a traditional AC (travel, lodging, professional role players, etc.).

Unfortunately, although many practitioners have eagerly adopted these technologies, there has been very little empirical evaluation of their effectiveness and almost no comparison with traditional ACs. Although there is some evidence to suggest that assessing via video does not diminish the accuracy of assessment (Ryan et al., 1995), technology may have a negative impact on other outcomes such as applicant reactions (e.g., Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). Of course, the effects of technology are not necessarily negative. On the contrary, the proponents of many of these technologies argue that they can improve on traditional techniques (e.g., by being more realistic, by capturing responses in greater detail, etc.), although empirical verification of these claims is lacking. Ultimately, the ACM is a technique grounded in human judgment. The effects of automating portions or all of the process must be better understood in order to determine whether such technologically enhanced assessments can truly be considered assessment centers in the traditional sense of the term.

6. Summary

Assessment centers have been used extensively to aid in the selection of personnel for various new assignments. Such applications include external screening, internal promotion, early identification of potential, and certification of competence. A wide variety of evidence has accumulated demonstrating the validity of ACs for selection. Job analysis and competency modeling are typically used to study the performance domain of target jobs. Results of these analyses identify the dimensions to be assessed and the content of assessment exercises. Multiple assessors observe overt behavior in exercises simulating important job situations. Ratings of dimensions and overall performance have been found to relate to a variety of criteria including measures of comparable constructs and job performance. Sub-group differences in AC scores tend to relatively small in comparison with differences in cognitive ability test scores.

Controversies are present as a result of conflicting research findings and recommendations for practice. Additional research is needed to understand what types of dimensions (e.g., cognitive abilities, interpersonal skills, and affective characteristics) can be assessed, the impact of faking and impression management on AC evaluations, the desirability of using human attributes, tasks, roles, or some other variables as the structural framework for the design and implementation of ACs, the extent to which dimensions should be made known to candidates, and the optimal way to integrate observations and ratings from multiple assessors and exercises. Research into new areas is also needed, including the effectiveness of programs to train candidates prior to participation in ACs, whether exercises can be reused without compromising validity and fairness, the impact of instances of lack of standardization of administration, and methods that can be employed to ensure that behavior relevant to dimensions is elicited in the simulation exercises.

Two additional general research areas are suggested. First, classic measurement principles of "general mental ability" and "the indifference of the indicator" are proposed to apply to the AC method when used for selection. Analogous to test measures of "*g*." selection ACs can be considered measures of a general ability to perform a job or set of jobs, e.g., "general managerial ability" designated g_{mgmt}. A variety of different indicators of that ability can be garnered from performance in simulation exercises without reference to some intermediate set of variable such as dimensions or tasks or exercises. The second general area calls for a re-conceptualization of intra-individual differences in behavior within an AC. Whereas inconsistency across exercises has traditionally been considered error of measurement or situation specificity, a new avenue of research can investigate the meaning of individual differences in consistency and inconsistency in performance. Such individual differences may be stable characteristics that are related to other personality characteristics and are predictive of future job performance.

The study of ACs in the context of personnel selection provides opportunities to advance theory, research, and practice in human resource management. Basic human processes of faking and impression management among candidates as well as observation and decision making among assessors can be studied in the rich yet semi-controlled environment of simulation exercises. Innovations in the design and implementation of simulations and in methods of integration of complex information hold promise for improving the effectiveness of ACs for matching human capabilities and organizational requirements.

In conclusion, evidence suggests that the AC method offers a viable alternative and supplement to other personnel selection methods. Research and practice suggest that ACs are valid, fair, legally defensible, and acceptable to candidates and other stakeholders in a wide variety of jobs.

References

Adler, S. (1987). Toward the more efficient use of assessment center technology in personnel selection. *Journal of Business and Psychology*, *2*, 74—93.

American Educational Research Association, American Psychological Association, & American Council on Measurement in Education (1999). *Standards for Educational and Psychological Tests.* American Psychological Association: Washington, D.C.

Anderson, N., Lievens, F., Van Dam, K., & Born, M. (2006). A construct investigation of gender differences in a leadership role assessment center. *Journal of Applied Psychology*, *91*, 555—566.

Archambeau, D. J. (1979). Relationships among skill ratings assigned in an assessment center. *Journal of Assessment Center Technology*, *2*, 7—19.

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternative view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology*, *1*, 105—111.

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., Jr., & 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.

Arthur, W., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, *26*, 813—835.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment*, *13*, 1—10.
Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment*, *12*, 135—148.
Binning, J. F., Adorno, A. J., & LeBreton, J. M. (1999, April). *"Sociotechnical" moderators of assessment center criterion-related validity*. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.
Binning, J. F., Gniatczyk, L. A., LeBreton, J. M., & Melcher, K. M. (2002, April). *The moderating effect of assessors' judgment processes on the criterion-related validity of judgmental ratings in an operational assessment center*. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.
Bobrow, W., & Leonards, L. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality*, *12*, 217—236.
Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology*, *67*, 3—9.
Bray, D. W., & Campbell, R. J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology*, *52*, 36—41.
Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives.* New York: John Wiley & Sons.
Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied*, *80*(17), 1—27.
Brink, K. E., Lance, C. E., Bellenger, B. L., Morrison, A., Scharlau, E., & Crenshaw, J. L. (2008, April). *Discriminant validity of a "next generation" assessment center*. Paper presented at the 23rd annual meeting of the Society for Industrial and Organizational Psychology. San Francisco, CA.
Bush, M. A. (2004). Assessment center construct validity: Establishing expectations based on the dimension activation theory (Doctoral dissertation, University of Tennessee, Knoxville). *Dissertation Abstracts International: Section B: The Sciences & Engineering, 51*(1-B), 469.
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.
Byham, W. C. (1970). Assessment centers for spotting future managers. *Harvard Business Review*, *48*, 150—168.
Byham, W. C., Smith, A. B., & Paese, M. J. (2000). *Grow your own leaders. Acceleration pools: A new method of succession management.* Pittsburgh, PA: DDI Press.
Caldwell, C., Gruys, M. L., & Thornton, G. C., III (2003). Public safety assessment centers: A steward's perspective. *Public Personnel Management*, *32*, 229—249.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81—105.
Campbell, J. P., McCloy, R., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35—70). San Francisco: Jossey-Bass.
Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.
Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology*, *69*, 167—181.
Clapham, M. M., & Fulford, M. D. (1997). Age bias in assessment center ratings. *Journal of Managerial Issues*, *9*, 373—387.
Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment*, *11*, 17—29.
Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology*, *63*, 211—216.
Coulton, G. F., & Feild, H. S. (1995). Using assessment centers in selecting entry-level police officers: Extravagance or justified expense? *Public Personnel Management*, *24*, 223—254.
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.
Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: Efficient tool or a hammer to kill a fly? *Personnel Psychology*, *55*, 827—849.
Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685—691.
Denning, D., & Grant, D. (1979). Knowledge of assessment process: Does it affect candidate ratings? *Journal of Assessment Center Technology*, *2*, 7—12.
Do, B.-R., & Brummel, B. J. (2006, April). *Item preknowledge on test performance and item confidence*. Paper presented at the 21st Annual Meeting of the Society for Industrial and Organizational Psychology. Dallas, TX.
Dodd, W. E. (1975). Attitudes toward assessment center programs. Unpublished manuscript; IBM Corporation.
Dodd, W. E. (1977). Attitudes toward assessment center programs. In J. I. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 161—183). New York: Pergamon Press.
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.
Feltham, R. (1988). Assessment centre decision making: Judgemental vs. mechanical. *Journal of Occupational Psychology*, *61*, 237—241.
Fiske, D. W., Hanfmann, E., MacKinnon, D. W., Miller, J. G., & Murray, H. A. (1948). *Selection of personnel for clandestine operations: Assessment of men.*Walnut Creek, CA: Aegean Park Press (Reprinted).
Fletcher, C. (1991). Candidates' reactions to assessment centres and their outcomes: A longitudinal study. *Journal of Occupational Psychology*, *64*, 117—127.
Franks, D., Ferguson, E., Rolls, S., & Henderson, F. (1999). Self-assessments in HRM: An example of an assessment centre. *Personnel Review*, *28*, 124—133.
Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*, 493—511.
Gaugler, B. B., & Thornton, G. C., III (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, *74*, 611—618.
Gavin, J. F., & Hamilton, J. W. (1975). Selecting police using assessment center methodology. *Journal of Police Science and Administration*, *3*, 166—176.
Gibbons, A. M., & Rupp, D. E. (2004, April). *Developmental assessment centers as training tools for the aging workforce*. Paper presented at the 19th Annual Meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.
Gibbons, A. M., & Rupp, D. E. (2007, April). *Inconsistency in assessment center performance: A meaningful individual difference?* In Rupp, D.E. (Chair), Assessment center (modern) validity: Forty years since Bray and Grant. Presented at the 22nd Annual Meeting of the Society for Industrial Organizational Psychology, New York.
Gibbons, A. M., & Rupp, D. E. (in press). Performance consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*.
Goffin, R. D., Rothstein, M. G., & Johnston, N. G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology*, *81*, 746—756.
Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology*, *51*, 357—374.
Gowing, M. (2006, September). *Small simulations, big results!* Paper presented at the 33rd International Congress on Assessment Center Methods, London, UK.
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.
Hardison, C. M., (2005). Construct validity of assessment center overall ratings: An investigation of relationships with and incremental criterion validity over Big 5 personality traits and cognitive ability. Unpublished doctoral dissertation: University of Minnesota.
Hardison, C. M., & Sackett, P. R. (2004, April). *Assessment center criterion–related validity: A meta–analytic update*. Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
Henry, S. E. (1988). *Nontraditional applications of assessment centers. Assessment in staffing plant start-ups*. Paper presented at the meeting of the American Psychological Association, Atlanta, GA.
Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance. *International Journal of Selection and Assessment*, *15*, 405—411.

Hiatt, J. (2000, May). *Selection for positions in a manufacturing startup*. Paper presented at the 28th International Congress on Assessment Center Methods. San Francisco, CA.
Hoffman, B. J. (Chair), Arthur, W., Lance, C. E., Lievens, F., Russell, C. J., & Woehr, D. J., (2008, April). *Assessment center validity: Where do we go from here?* Paper presented at the 23rd annual meeting of the Society for Industrial and Organizational Psychology. San Francisco, CA.
Hoffman, B. J., & Meade, A. W. (2007, May). *Invariance tests as assessment center construct validity evidence*. Paper presented at the 22nd annual meeting of the Society for Industrial and Organizational Psychology. New York, NY.
Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*, 455—470.
Hogan, J., & Zenke, L. L. (1986). Dollar-value of alternative procedures for selecting school principles. *Educational and Psychological Measurement*, *46*, 935—945.
Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality*, *12*, 13—52.
Howard, A. (2001). Identifying, assessing, and selecting senior leaders. In S. J. Zaccaro & R. Klimoski (Eds.), *The nature of organizational leadership* (pp. 305—346). San Francisco, CA: Jossey-Bass.
Howard, A., & Bray, D. W. (1988). *Managerial lives in transition: Advancing age and changing times*. New York: Guilford Press.
Howard, A. & Metzger, J. (2002, October). *Assessment of complex, consultative sales performance*. Paper presented at the 30th International Congress on Assessment Center Methods, Pittsburgh, PA.
Howard, L. & McNelly, T. (2000, May). *Assessment center for team member level and supervisory development*. Paper presented at the 28th International Congress on Assessment Center Methods. San Francisco, CA.
Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of White and Black females. *Personnel Psychology*, *29*, 13—30.
International Task Force on Assessment Center Guidelines (2008). *Guidelines and ethical considerations for assessment center operations*. Available at www.assessmentcenters.org
Jacobson, L. (2000, May). *Portfolio assessment: Off the drawing board into the fire*. Paper presented at the 28th International Congress on Assessment Center Methods, San Francisco, CA.
Jansen, P. G. W., & Stoop, B. A. M. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology*, *86*, 741—753.
Jensen, A. R. (1992). Commentary: vehicles of g. *Psychological Science*, *3*, 275—278.
Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*(3), 376—407.
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., Kuptsch, C., & Koller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review*, *45*, 67—84.
Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, *40*, 243—260.
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—338.
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.
Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment*, *14*, 360—371.
Krause, D. E., & Thornton, G. C. III (in press). A cross-cultural look at assessment center practices: A survey in Western Europe and North America. *Applied Psychology: An International Review*.
Kuptsch, C., Kleinmann, M., & Koller, O. (1998). The chameleon effect in assessment centers: The influence of cross-situational behavioral consistency on the convergent validity of assessment centers. *Journal of Social Behavior and Personality*, *13*, 102—116.
Lance, C. E. (2008a). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 84—97.
Lance, C. E. (2008b). Where have we been, how did we get there, and where shall we go? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 140—146.
Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*, 22—35.
Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*, 377—385.
Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, *13*, 323—353.
Landy, F. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*, 1183—1192.
Larsh, S. L. (2001). *Effects of feedback format on self-efficacy and subsequent performance: A comparison of attribute-based and situation-based developmental feedback*. US: ProQuest Information & Learning.
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. (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. (2007, May). *Assesssment centers: A tale of exercises, dimensions, and dancing bears*. Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York, New York.
Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2004, April). *Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity*. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.
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., & Thornton, G. C., III (2005). Assessment centers: Recent developments in practice and research. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *Handbook of Personnel Selection* (pp. 243—264). Malden, MA: Blackwell.
Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*, 715—738.
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.
Maurer, T. J., Solamon, J. M., & Lippstreu, M. (2008). How does coaching interviewees affect the validity of a structured interview? *Journal of Organizational Behavior*, *29*, 355—371.
McFarland, L. A., Yun, G., Harold, C. M., Viera, L., Jr., & 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.
Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology*, *80*, 517—524.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741—749.
Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35—44.
Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*, 252—283.
Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology*, *60*, 527—529.

Neidig, R. D., Martin, J. C., & Yates, R. E. (1979). The contribution of exercise skill ratings to final assessment center evaluations. *Journal of Assessment Center Technology*, *2*, 21−23.
Noe, R. A., & Steffy, B. D. (1987). The influence of individual characteristics and assessment center evaluation on career exploration behavior and job involvement. *Journal of Vocational Behavior*, *30*, 187−202.
Norton, S. D. (1981). The assessment center process and content validity: A reply to Dreher and Sackett. *Academy of Management Review*, *6*, 561−566.
Page, B. T. (1995). *Assessment center handbook*. Longwood, FL: Gould Publications.
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.
PoliceCareer.com, http://www.policecareer.com, retrieved June 15, 2008.
PolicePromotion.com, http://www.policepromotion.com, retrieved June 15, 2008.
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.
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.
Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*, 1009−1037.
Roth, P., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black–white differences in overall and exercise scores. *Personnel Psychology*, *61*, 637−662.
Rothstein, H., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta analysis: Prevention, assessment and adjustments* Chichester, UK: Wiley.
Rowe, T. L. (2006). *A preparation guide for the assessment center method.* Springfield, IL: Charles Thomas Publisher.
Rupp, D. E. (2006). The future is here: Recent advances in assessment center methodology. *Invited talk presented at the Society for Industrial and Organizational Psychology Fall Consortium. Charlotte, NC.*
Rupp, D. E., Gibbons, A. M., Baldwin, A. M., Snyder, L. A., Spain, S. M., Woo, S. E., et al. (2006). An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *Psychologist Manager Journal*, *9*, 171−200.
Rupp, D. E., & Thornton, G. C. III. (2003). *Development of simulations for certification of competence of IT consultants*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 116−120.
Russell, C. J. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal*, *30*, 817−826.
Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., et al. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology*, *80*, 664−670.
Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, *26*, 565−606.
Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology*, *40*, 13−25.
Sackett, P. R. (1998). Performance measurement in education and professional certification: Lessons for personnel selection? In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 113−129). Mahwah, NJ: Lawrence Erlbaum.
Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401−410.
Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology*, *69*, 187−190.
Sackett, P. R., & Tuzinski, K. A. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.), *How people evaluate others in organizations* (pp. 111−129). Mahwah, NJ: Lawrence Erlbaum.
Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482−486.
Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*, 735−746.
Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations.* Monterey, CA: Brooks/Cole.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262−274.
Schmitt, N. (1993). Group composition, gender, and race effects on assessment center ratings. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 315−332). Hillsdale: Lawrence Erlbaum Associates, Inc.
Schmitt, N., & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology*, *63*, 261−264.
Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology*, *86*, 451−458.
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.
Schuler, H., & Fruhner, R. (1993). Effects of assessment center participation on self-esteem and on evaluation of the selection situation. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 109−124). Hillsdale: Lawrence Erlbaum Associates, Inc.
Shore, T. H., Thornton, G. C., III, & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology*, *43*, 101−116.
Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions tranparent on relations with typical performance predictors. *Human Performance*, *20*, 187−203.
Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201−293.
Spearman, C. (1923/1973). *The nature of "intelligence" and the principles of cognition.* New York, NY: Arno Press.
Spearman, C. (1927). *The abilities of man.* New York, NY: Macmillian.
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 Applied Social Psychology*, *30*, 1474−1491.
Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology*, *50*, 71−90.
Stark, S., Chernyshenko, O. S., Chan, K. -Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, *86*, 943−953.
Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance*, *13*, 205−251.
Thoresen, J. D. (2002). *Do we need dimensions?* Paper presented at the meeting of the International Congress on Assessment Center Methods, Pittsburgh, PA.
Thornton, G. C., III, (1992). *Assessment centers in human resource management.* Reading, MA: Addison-Wesley.
Thornton, G. C. III, (1993, March). *Selecting entry-level brewery workers at Adolph Coors Company.* Paper presented at the 21st International Congress on the Assessment Center Method, Atlanta, GA.
Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance.* New York: Academic Press.
Thornton, G. C. III, Hollenbeck, G. P., & Johnson, S. K. (in press). Selecting leaders: Executives and high potentials. In J.L. Farr & N. Tippins (Eds.), *Handbook of employee selection*. Mahwah, NJ: Lawrence Erbaum.

Thornton, G. C. III, & Johnson, R. (2006, May). *Employment discrimination litigation involving assessment center practices*. Presentation in M.M. Harris, Recent Developments in Employment Discrimination Law and I–O Psychology at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Thornton, G. C. III, & Krause, D. E. (in press). Comparison of practices in selection vs. development assessment centers: An international survey. *International Journal of HumanResource Management*.

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

Thornton, G. C., Murphy, K. R., Everest, T. M., & Hoffman, C. C. (2000). Higher cost, lower validity and higher utility: Comparing the utilities of two tests that differ in validity, costs and selectivity. *International Journal of Selection and Assessment*, *8*, 61—75.

Thornton, G. C. III, & Potemra, M. (in press). Utility of assessment center for promotion of police sergeants. *Public Personnel Management*.

Thornton, G. C., III, & Rupp, D. R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development.* Mahwah, NJ: Lawrence Erlbaum.

Thornton, G. C., III, & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology*, *65*, 351—354.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology*, *87*, 1020—1031.

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

Tziner, A., Meir, E. I., Dahan, M., & Birati, A. (1994). An investigation of the predictive validity and economic utility of the assessment center for the high-management level. *Canadian Journal of Behavioural Science*, *26*, 228—245.

Tziner, A., Ronen, S., & Hacohen, D. (1993). A four-year validation study of an assessment center in a financial corporation. *Journal of Organizational Behavior*, *14*, 225—237.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*, 197—210.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment*, *8*, 216—226.