been no long-lasting implementations in this area. It just doesn’t work out in practice.

Mayes: There have been many social changes in the United States during the past 30 years. We have become sensitive to the equal employment rights of many groups of citizens. Have these movements affected the way assessment centers operate?

Byham: Assessment centers have been very concerned about social justice and fairness. A great deal of research and effort have gone into these issues. In general, assessment centers have been found to be more fair to protected groups than any other selection methodology.

Mayes: Looking to the future, with the emergence of a global economy and new communication technologies, what changes do you expect to see in assessment centers?

Byham: Assessment centers have targeted two issues: the selection of individuals for overseas assignment, and managers who can think globally. These issues have been attacked through changes in assessment exercises to bring in international issues. Where adaptability to overseas assignments is an issue, assessment center exercises have incorporated exercises that tap sensitivity to cultural differences and the ability to communicate to other cultures. Most importantly, background interview data have been used extensively as a measure of adaptability.

The communication revolution presents some interesting potential applications for assessment methods; for example, assessment over the Internet for personnel development. Assessment centers might also operate through CD-ROM for personnel development. Live assessment for selection or personnel development might be possible through videoconferencing.

Mayes: Thank you very much for your participation in this interview.

A Reassessment of Assessment Centers: Challenges for the 21st Century

Ann Howard
Development Dimensions International and Leadership Research Institute
21 Knoll Road, Tenafly, NJ 07670

During the 40 years since its inauguration, the management assessment center has spread widely and been subjected to continued research. The challenges it faced in its early years are significantly different from those that preoccupy researchers and practitioners at present. This paper reviews evidence relating to both old and new challenges to the assessment center method. Older issues, now mostly put to rest, concerned what the method is (and is not), criterion-related validity, generalizability, user reactions, and whether simpler methods, such as tests and interviews, can substitute. Newer challenges concern construct validity, the changing nature of managerial work, new applications, user-friendliness, and whether multi-rater assessments can substitute. The review concludes that assessment centers have much to offer organizations in the 21st century and makes projections and recommendations for future assessment center practice and research.

Two months prior to this writing, the management assessment center celebrated its 40th birthday. By all indications it is enjoying a strong and healthy middle age. Hundreds of thousands of people have been assessed by the method, and it has been the subject of over 100 research and review articles and several books. Various consultants report that their assessment business is growing (Dobrzynski, 1996). Yet the current popularity of the method could be perceived as somewhat odd. Why, when organizations and the nature of work are changing so dramatically, are people relying on a 40-year-old technique to evaluate managers and other personnel?

This is not the first time that assessment centers have been considered at odds with the times. A review more than twenty years ago noted that they combine with surprising success measures and methods (e.g., personality and projective tests, interviews, situational tests, clinical predictions) then considered of dubious value for personnel selection (Howard, 1974). Today practitioners’ and researchers’ concerns about assessment

TABLE 1  Challenges to Assessment Centers

<table>
<thead>
<tr>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is it?</td>
<td>What makes it work?</td>
</tr>
<tr>
<td>Does it work?</td>
<td>Can it continue to work?</td>
</tr>
<tr>
<td>What organizations can it help?</td>
<td>Where can it help organizations?</td>
</tr>
<tr>
<td>Will users like it?</td>
<td>Can users afford it?</td>
</tr>
<tr>
<td>Can simpler methods substitute?</td>
<td>Can multi-rater assessments substitute?</td>
</tr>
</tbody>
</table>

centers are in many ways different from those expressed in the 1970s. These shifting concerns are partly due to changing times and partly due to progress in assessment center practice and research. Many of the original questions have been answered and new challenges have arisen to take their place.

This paper first considers how old challenges to the assessment center have been resolved and then weighs the evidence relative to new challenges. It ends with projections for the assessment center method in the 21st century and recommendations for practice and research.

The primary questions about the assessment center method that critics have raised across time appear in Table 1. One way to view them is by contrasting pairs.

1. **What is it?** vs. **What makes it work?** It was originally unclear which of many components must be included for a procedure to be called an assessment center. Today the internal workings of assessment centers provoke more concerns. In particular, are assessors rating dimensions or just exercise performance?

2. **Does it work?** vs. **Can it continue to work?** Early research on the criterion-related validity of assessment centers was promising but sparse, and questions lingered about whether it would predict criteria of importance. Changing criteria now raise new questions about the method. Do assessment centers follow a model for staid bureaucracies that won’t fit tomorrow’s flexible departments?

3. **What organizations can it help?** vs. **Where can it help organizations?** Fewer than 100 organizations had implemented assessment centers by the early 1970s (Byham, 1992). Given their wide use today, there is less concern with generalizability across organizations and more concern with generalizability across human resource functions and job levels. In particular, is the assessment center too blunt a tool for executive selection?

4. **Will users like it?** vs. **Can users afford it?** Early writers feared that assessment centers would have unfortunate effects on individuals’ careers and be rejected by assesses and managers. In the fast-paced, globally competitive world of business today, acceptance is more likely to hinge on the method’s costs, time requirements, and logistical problems. Is the assessment center too user-unfriendly?

5. **Can simpler methods substitute?** vs. **Can multi-rater assessments substitute?** Because of the costs of assessment centers, early reviewers suggested substituting cheaper methods like tests, interviews, or personnel records. Today a new competitor is the multi-rater assessment, which has similarities to the assessment center method. Can multiple ratings of work performance eliminate the need for pooled ratings of assessment center performance?

The following section reviews briefly the older challenges to assessment centers. For the most part these early questions have been answered, although two of them gathered new fuel as the years went by.

**OLD CHALLENGES**

**What Is It?**

Some early applications threatened the reputation of the assessment center by borrowing its name for procedures that incorporated only parts of the method. Responsible users felt the need for some ethical guidelines. This need originated at a meeting of the International Congress on the Assessment Center Method, held in 1975. More experience with the method and issuance of federal testing standards prompted a revision of the guidelines in 1979. User innovations and new research stimulated still a third set of guidelines, which were endorsed by the 17th International Congress on the Assessment Center Method in 1989 (Task Force, 1989).

These latest guidelines specify that the following elements must be present for a process to be considered an assessment center.

1. **Dimensions (and other categories).** A job analysis of relevant behaviors should identify the dimensions, attributes, and job performance important to job success and determine what the assessment center should evaluate. Behavioral observations must then be classified into job-relevant categories. Although traditionally called dimensions, the categories may be attributes, characteristics, aptitudes, qualities, skills, abilities, knowledge, or tasks.

2. **Techniques.** The techniques or exercises must be designed to provide information for the dimensions or attributes identified in the job analysis. Multiple techniques must be used; these may include tests, interviews, questionnaires, sociometric devices, and simulations. Job-related simulations, such as in-baskets, group discussions, interaction simulations, and fact-finding exercises, are the hallmark of assessment centers. There must be a sufficient number of these to allow opportunities
to observe candidates' behavior related to the dimensions being assessed.

3. Assessors. Multiple assessors must be used to observe and evaluate each assessees. Assessors must receive thorough training and demonstrate their ability to perform.

4. Gathering and reporting data. Assessors must use a systematic procedure (e.g., handwritten notes, behavioral observation scales, behavioral checklists) to record specific behavior observations as they occur. They must prepare a report or record of observations made during each exercise. Data from assessors and other techniques must be pooled by a meeting of assessors or a validated statistical process.

The guidelines also specify what is not an assessment center. On this list are panel interviews as the sole technique, reliance on a single technique (although one large, complex simulation is acceptable if it includes distinct job-related segments), a paper-and-pencil test battery, a single-assessor evaluation, or simulations that omit pooling of data.

As long as the fundamental characteristics are not violated, the guidelines open assessment center methodology to a number of innovations. This openness, however, is both a strength for innovative practice and a barrier to definitive research.

Does It Work?

Early criterion-related validity data were sparse, came from too few sources, covered too many variations in components, lacked replication, and suffered from criterion contamination (Howard, 1974; Klimoski & Strickland, 1977). During the past twenty years researchers have produced considerably more validity evidence and consolidated findings in several meta-analyses (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984). The average validity coefficient is around .40. Assessment centers were validated against career progress, overall performance ratings, dimensional performance ratings, potential ratings, wages, and training performance. Assessment center performance correlates most highly (r = .53) with potential ratings (Gaugler et al., 1987). There is no question that the assessment center has solid criterion-related validity; it works.

Validity coefficients cover a wide range: in one meta-analysis the 90% confidence interval extended from .21 to .53 (Gaugler et al., 1987). This range is partly attributable to the variation in applications; different practices moderate the established validities. Centers producing greater validity had a wider variety of exercises, psychologists in addition to managers as assessors, peer evaluations, and more women and fewer minorities as assesses (Gaugler et al., 1987). Another study of school principals found greater predictive accuracy when more assessors were principals, assessments took place across several locations, and assessors were not close work associates of assesses (Schmitt, Schneider, & Cohen, 1990). Assessment centers work, but how the center is constructed makes a difference in how well it works.

What Organizations Can It Help?

The management assessment center, which began with the Management Development Study research at AT&T, spread first to Bell telephone companies for operational use and then crept out to a few large firms. Early validity data came primarily from AT&T, IBM, and SOHIO (Howard, 1974). The big breakthrough for the dispersion of assessment centers was the founding in the early 1970s of consulting companies that packaged assessment materials for many organizations to use.

Assessment centers now function in a wide variety of organizations, including industrial, educational, military, government, and professional. They have even been used to select clergy (Voyle & Voyle, 1994). It is impossible to get a precise count of the number of people who have been assessed, but surely it is in the hundreds of thousands. Moreover, assessment centers have spread far beyond the United States. Although the U.S.-based “International” Congress had few foreign participants in its early years, representatives from dozens of countries have attended recent meetings. Assessment center conferences also appear regularly in other regions, such as Europe and South Africa. Sample surveys from the United Kingdom indicate nearly a three-fold increase in the use of assessment centers between 1986 and 1991, from 21% to 59% of companies surveyed (Shackleton, 1991). Perhaps the most beautiful assessment center facility resides in Bandung, Indonesia.

At present there seem to be no limits on the kinds or locations of organizations that can make use of assessment centers. How well the method translates to these different sites needs further study, particularly given the number of factors that can moderate validity.

Will Users Like It?

An early concern was whether an assessees did well at a center would be dubbed a "Crown Prince or Princess" and enjoy subsequent promotions as a self-fulfilling prophecy (Kraut & Scott, 1972). Contrarily, those who did not come up to standards might feel that their careers had been given the "Kiss of Death" and leave the corporation (Howard, 1974).

A variety of data dispel these concerns. Validity coefficients are not significantly higher for career progress compared to current performance (Gaugler et al., 1987), which casts doubt on the self-fulfilling prophecy. Organizations often report that they use the assessment center for initial promotions only and it ceases to be considered after that.
Other studies have generally shown positive attitudes toward assessment centers on the part of assesses, assessors, and management (Dodd, 1977; Dulewicz, 1991; Thornton & Byham, 1982; Thornton, 1992). Participants find the exercises difficult and challenging but believe that they measure job-relevant qualities and are fair. They also typically have positive reactions to feedback and its usefulness for their development. As might be expected, those who perform better in assessment centers often have more positive attitudes toward it (Baisden & Roberson, 1993; Jones & Bradley, 1994), although this is not always the case (Dulewicz, 1991).

Studies of applicant reactions have taken on more importance recently because of new attention to meeting the needs of the "customer." Applicants view work samples and the assessment center as more face valid, acceptable, and fair than paper-and-pencil tests (Cascio & Phillips, 1979; Macan, Avedon, Paese, & Smith, 1994; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977). This finding has important implications, because applicants who perceive selection techniques more favorably are also more satisfied with the selection process, the job, and the organization (Macan et al., 1994).

Not all users like the assessment center method. This is not unexpected given that people often dislike being evaluated, no matter who or what is doing the evaluating. Within these constraints, the assessment center usually elicits positive reactions and definitely fares better than paper-and-pencil measures.

**Can Simpler Measures Substitute?**

Because the assessment center is expensive and cumbersome, various writers have questioned whether simpler and cheaper measures could predict important criteria just as well. This challenge was met more easily in earlier years; new data reinvigorate the debate and make the response more complex.

An early study found that personnel records could predict later performance as well as an assessment center and questioned whether the assessment was worth the expense (Hinrichs, 1978). Using the two methods together, however, was even more effective. A study comparing assessment centers and personnel records as predictors of management potential among fire and police candidates replicated these findings (Lowry, 1994). The assessment center also made a considerable contribution to variance in training performance beyond that of supervisors' evaluations (Tziner & Dolan, 1982). All of these studies indicated that personnel evaluations were missing information about interpersonal relationships, which is where the assessment center can add value.

Personnel records pose additional problems. They often aren't standardized and may inadequately represent performance, experience, and training (Lowry, 1994). Personnel and performance records also have limited value when the demands of a more advanced job are not the same as that of the current job (Dulewicz, 1991).

Earlier data suggested that criterion-related validity was higher for assessment centers than for other available selection instruments (Howard, 1974). Ten years later a meta-analysis reported that the average validity coefficient of selection tools was .28 but that of assessment centers was .41 (Schmitt et al., 1984). The only predictors in the same range as assessment centers were work samples and supervisor/peer evaluations. Biodata (.24), mental ability tests (.25), and especially personality tests (.15) had notably lower average validity coefficients.

But other predictors continue to be sharpened. For example, the validity of structured employment-interviews is now reported in the .40s (Landy, Shankster-Cawley, & Moran, 1995). Meta-analyses with the "Big Five" personality characteristics report validities in the .20s (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991), and keyed biodata produce validities ranging from .26 for promotion criteria to .37 for supervisory ratings (Hunter & Hunter, 1984). Researchers have accumulated massive amounts of data on the predictive power of cognitive ability tests (Hunter, 1986; Ree & Earles, 1992; Schmidt & Hunter, 1992); validities for proficiency are in the .50s for highly complex jobs and in the .20s for simple jobs (Hunter, 1986). In terms of pure prediction, then, the competition is heating up. A more intriguing question is how different measuring devices come together in predicting and understanding work behavior (Landy et al., 1995).

To complicate matters, assessment centers can and have included all of the above measures. The prominent measurement psychologist E.L. Thorndike once remarked that anything that exists, exists in some amount, and anything that exists in some amount can be measured. One could add an extension that anything that can be measured, can be measured in an assessment center.

The critical challenge then becomes determining whether separate measurement components make independent contributions to validity. Of particular importance are the simulations, for these are the hallmark of the method and the source of the greatest expense and most tortuous logistics. Data from centers that include a variety of measures are sparse but encouraging. An early study at IBM showed that a personality test scale, two assessment exercises, two assessment ratings, and a biodata inventory made independent contributions to the prediction of an increase in managerial responsibility (Wollowick & McNamara, 1969). In the Man-
WHAT MAKES IT WORK?

Assessment center dimensions represent categories of behavior that can be measured across more than one exercise. Yet, research shows more similarity among dimensions rated within one exercise than among the same dimensions rated across exercises (Archambau, 1979; Bycio, Alvares, & Hahn, 1987; Henderson, Anderson, & Rick, 1995; Joyce, Thayer, & Pond, 1994; Neidig, Martin, & Yates, 1979; Robertson, Gratton, & Sharpley, 1987; Russell, 1987; Sackett & Dreher, 1982; Schneider & Schmitt, 1992; Silverman, Dallessio, Woods, & Johnson, 1986; Turnage & Muchinsky, 1982). These findings suggest that exercises and not dimensions are the currency of assessment centers. Although data often show convergent validity (dimensions do correlate across exercises), discriminant validity is usually wanting (dimensions correlate more highly with other dimensions in the same exercise). This is the so-called construct validity problem which continues to trouble researchers trying to understand just what makes the assessment center method work.

The classic explanation for how the method works is that assessors observe behavior displayed in exercises, classify those behaviors into categories of human attributes, and use those categories to make meaningful predictions about job performance. But critics question this explanation if the dimensional categories aren't rated consistently across exercises.

Alternative explanations for why assessment centers work (Klimoski & Brickner, 1987), have garnered little support. Following are summaries of the arguments and data that confirm or refute them.

1. Criterion contamination. Hypothesis: Criteria used to validate assessment centers are contaminated because the results were used for prior promotions. Yet, meta-analyses show that assessment centers predict current performance as well as career progress (Gaugler et al., 1987). Assessment results were kept confidential in the original AT&T research and still related to advancement twenty years later (Howard & Bray, 1988). Other pure research studies showed about the same level of predictive validity as centers used for decision making (Gaugler et al., 1987).

2. Subtle criterion contamination. Hypothesis: Assessment center ratings and later ratings of performance and progress are nothing more than shared biases. But assessment data related more closely to criteria when psychologists, who presumably lacked the same biases, joined managers as assessors (Gaugler et al., 1987). Assessment centers also have predicted criteria like judgments of third-party observers, ratings by subordinates, and turnover by subordinates; these findings likewise fail to support the hypothesis (Thornton, 1992).
3. Self-fulfilling prophecy. Hypothesis: Expectations about employees' performance, perhaps communicated through the supervisor's behavior, modify employees' feelings of self-efficacy, which later affects their job performance. The evidence disagrees; assessment centers continue to show validity even when no feedback is given to the assessees or the supervisor (Gaugler et al., 1987).

4. Intelligence. Hypothesis: Assessment ratings reflect only intelligence. Intelligence certainly relates to assessment center performance unless range has been restricted by previous selection processes. But as described earlier, multiple regression analyses have shown that assessment center simulations account for variance in performance beyond that attributed to intelligence (Howard & Bray, 1988; Tziner & Dolan, 1982; Wollowick & McNamara, 1969).

5. Consistency with past performance. Hypothesis: Assessors predict future performance based on knowledge of past performance gathered through a background interview or documented experience. But many assessment centers deliberately hide the names and backgrounds of participants from assessors, undermining this hypothesis (Thornton, 1992).

6. Exercises are job samples. Hypothesis: Simulations are just samples of the job and there is no need to make inferences about underlying human attributes. Overall exercise ratings are no more predictive of later job performance than dimensional ratings, but this neither refutes nor supports the hypothesis (Thornton, 1992). As noted earlier, the assessment center guidelines now permit categorizing behaviors by tasks as well as dimensions (Task Force, 1989). Some of the consequences of doing this are discussed later in this review.

To date, no convincing case has been made that the traditional explanation of why assessment centers work should be replaced by an alternative explanation. Nevertheless, assessment centers' repeated failure to exhibit cross-exercise discriminant validity has inspired a variety of approaches to either fix the construct validity issue or get around it. Following is a summary of these major approaches.

Tweaking Dimension Ratings

Various writers have observed that rating dimensions may involve more cognitive complexity than the average person can handle. Dimensions have always been muddled collections of traits (e.g., energy, learned skills (planning), readily demonstrable behaviors (oral communication), basic abilities (mental ability), attitudes (social objectivity), motives (need for advancement), knowledge (industry knowledge), and other attributes or behaviors, all of which are likely to vary in level of specificity. Add ambiguous definitions to these dimensions and it is not surprising that assessors get confused trying to rate them. Lowry (1995) boldly stated that assessors in the public sector can't do it. Assessors' accuracy is demonstrably poorer as the number of dimensions grows (Gaugler et al., 1987; Maher, 1990).

Users have tweaked assessment center ratings, just as others have tinkered with performance appraisals, to rid them of halo and other inaccuracies. Schneider and Schnitt (1992) made concerted efforts to clarify assessor training, sharpen definitions of dimensions, and standardize a scoring format, but they felt that their efforts didn't accomplish much. Maher (1995) found that beyond a threshold level, additional assessor training didn't help. Two days of training enhanced accuracy considerably more than one day, but adding a third day made no noticeable improvement.

Others successfully used behavior checklists to standardize assessor observations (Reilly, Henry, & Smithir, 1990). Convergent validity improved significantly from .24 to .43, and discriminant coefficients declined slightly from .47 to .41. Checklists, however, can be disconcerting for assessors, who must search through long lists while continuing to observe behavior (Thornton, 1992). Other writers have speculated that checklists could even cause discriminant validity to decline if diverse behaviors describe the same dimension in different exercises (Joyce et al., 1994). Sometimes more precision in the short run creates less in the long run.

Key behaviors (also called key actions) are an alternative to checklists. Approximately three to eight key behaviors characterize a dimension—enough to add structure and reliability but not so many that they distract assessors. Unlike checklists, which focus on frequency, they capture both the quality and quantity of relevant actions. Key behaviors represent models of effectiveness for handling situations. They are like scripts in cognitive psychology—sequences of events or behaviors that lead up to an outcome. For developmental assessments they are learning models for demonstrating competencies that resemble learning points in behavior modeling training (Hauenstein, 1994).

Key behaviors are not, however, independent across dimensions, and using them builds in dimensional intercorrelations. To illustrate, Development Dimensions International (DDI) recommends key principles for interpersonal relationships that include such behaviors as "listen and respond with empathy" and "maintain and enhance self-esteem." These could, and DDI maintains should, be used in a wide variety of interpersonal interactions that might be measured in assessment center exercises. The overlap of key behaviors across dimensions can help clarify which dimensions should be highly correlated.
Key behaviors may offer the best hope for cleaner and more precise dimension ratings. More research on them would be welcomed.

**Dimensional Domains**

Factor analyses have shown that dimension ratings cluster together within similar domains of performance (Thornton & Byham, 1982). The number of factors that emerge and their nature depends on what and how many dimensions are included in the assessment center. Two factors that fairly consistently appear are administrative and interpersonal skills. In more recent centers, leadership and performance management may be set apart from communication and general interpersonal skills. A cognitive factor also emerges in centers that include cognitive ability tests (Bray & Grant, 1966; Huck & Bray, 1976; King & Boehm, 1980; Scherer, Brodziński, & Shore, 1992). When assessment dimensions move beyond skills and abilities into areas related to motivation, personality, or personal style, findings vary widely (Thornton & Byham, 1982).

Some construct validity studies have supported dimension ratings within major domains by correlating them with external measures of related constructs. In one study intellectual ability tests were more highly related to performance or administrative dimensions and personality test measures were more highly related to the interpersonal cluster of dimensions (Shore, Thornton, & Shore, 1990). In another study a self-reported measure of interpersonal behavior was more strongly related to post-exercise ratings from an interview simulation that measured interpersonal behavior than to an in-basket (Russell, 1987).

Dimensional domains are directly or indirectly taken into account in assessment center design. According to Douglas Bray, who originated the first management assessment center, exercises were selected to represent major domains. There was no thought that each technique would accurately measure every dimension. He selected the in-basket primarily to reveal administrative skills, the business game and leaderless group discussion to highlight interpersonal skills, and paper-and-pencil cognitive tests to provide evidence about intellectual ability. Assessors often picked up evidence for a dimension from various techniques, but there was no requirement that they do so. If, for example, an assesseee insulted his subordinates in his in-basket memos, this would be counted against him on leadership skills, but the in-basket was generally not expected to reveal as much about these characteristics as the candidate’s behavior in the business game or group discussion.

In other words, assessment centers include different exercises not simply to provide multiple inputs on constructs but to adequately sample the job content domain. Moreover, using exercises that repeat examples of the same ability would have little practical value and would require sacrificing a more complete sampling of job-related behavior (Neidig & Neidig, 1984).

One simulation is often expected to cover a number of similar dimensions within the same domain. If assessment centers have only a few exercises, then it is not uncommon to find that the exercise is the domain. For example, one study examining convergent and discriminant validity used four exercises— an in-basket, an analysis problem, a coaching exercise, and a written situational test (Joyce et al., 1994). Excluding the situational test as not being a simulation, it could be argued that the three remaining “exercise” factors represented administrative skills (in-basket), interpersonal skills (coaching exercise), and cognitive (case analysis) domains.

The dimensions that cluster together within an exercise are not a random collection of dimensions; they are more similar to each other than a random set because they usually represent one domain. If a dimension rated within a exercise is from a secondary domain, it often has a different connotation. For example, leadership in a group discussion may characterize interpersonal influence; leadership in an in-basket may illustrate guidance. This confounding of exercise and domain diminishes discriminant validity.

Dimensional domains only partially explain low discriminant validity. Several studies have found that dimensions were differentiated more clearly across parallel exercises than dissimilar exercises, but there was still an exercise effect. A study of parallel in-baskets found low convergent validity coefficients that were slightly exceeded by monomethod correlations (Brannick, Michaels, & Baker, 1989). But this study used a highly questionable procedure of scoring every item for each of five dimensions. Sackett and Harris (1988) used confirmatory factor analysis to examine three leaderless group discussions in one center. The best solution was a combination of three method factors and three dimension factors.

A true exercise effect may account for some of the findings of poor discriminant validity; that is, ability may overlap with the exercise (Neidig & Neidig, 1984). Some people simply do better, for example, in one-on-one exercises than in group discussions. In an experiment designed to explore what causes exercise effects, Schneider and Schmitt (1992) found that form (leaderless group discussion and role play) rather than content (competitive and cooperative) accounted for a significant amount of the method variance. But method variance should not necessarily be considered error. Different measurement methods capture different aspects of the total criterion space, and there may be multiple true
performance scores as a function of the situation where the behavior is observed (Lance, Teachout, & Donnelly, 1992).

Assessment center practitioners can take several approaches to get around the problem of domain intercorrelations and exercise effects. Select dimensions to be rated in each exercise carefully, with fewer overlaps across exercises. Differentiate dimensions rated in very different contexts. For example, measure individual leadership and group leadership separately and treat them as distinct dimensions (Byham & Associates, 1992).

Some might argue that it is better to have no overlap of dimensions and exercises. Yet despite the presence of dimensional domains, some general dimensions can and should be measured across several exercises. Oral communication is one example; several confirmatory factor analyses have surfaced this dimension as a cross-exercise factor (Archambau, 1979; Bycio et al., 1987; Sackett & Harris, 1988).

Confirmatory factor analyses frequently show a combination of dimension factors and exercise factors (Bycio et al., 1987; Fletcher & Dulewicz, 1984; Louiselle, 1986; Sackett & Harris, 1988), suggesting that it would be unwise to conclude that dimensions misrepresent what is measured in assessment centers. But there are exercise factors as well, and users should be conscious of the context in which they are observing behavior.

Within-Exercise Dimension Ratings

The construct validity issue surfaced from the practice of rating many dimensions within each exercise. But within-exercise dimension ratings do not accord with the original design of assessment centers. In the Management Progress Study assessors described exercise performance in reports but made dimension ratings only after all the reports and material on each candidate were read aloud in the integration session (Howard & Bray, 1988). Many centers today still use this across-exercise format.

There are several reasons why indiscriminate use of within-exercise ratings creates problems. Many exercises do not sample enough behaviors to provide evidence for multiple dimensions, and assessors end up basing their judgments on only a few indicators (Bycio et al., 1987). Moreover, the same behavior may be used to infer more than one dimension. In addition, one assessor typically rates an individual on a given exercise, which introduces bias. One rater also creates a problem for statistical analyses: error due to the assessor (halo) compounds error due to the exercise (method variance) (Jones, 1992).

All of these reasons assure less discriminant validity when dimensions are rated after each exercise rather than during the integration process. Comparative research has, in fact, confirmed this (Silverman et al., 1986). If assessors wait until the end to rate dimensions, they are more likely to be thinking about similarities in behavior across exercises.

The evidence of low discriminant validity has come from both factor analyses and the multitrait multifactor (MTMM) paradigm. Both require the within-exercise rating of dimensions. To make matters worse, these methods treat ratings of a dimension from each exercise equally even though this does not accord with practice. In the integration session assessors do not simply average within-exercise ratings to produce a final dimension rating. They may, in fact, base their final dimension ratings on only one convincing exercise and discount the rest. Some exercise-by-dimension grids indicate whether an exercise is likely to elicit strong or weak evidence for a dimension, but the statistical treatments don't take this into account.

Confirmatory factor analyses offer more powerful tests than the MTMM or exploratory factor analyses (Bycio et al., 1987; Lance et al., 1992; Thornton, 1992). But even with this method researchers have violated the assumption that errors are uncorrelated because a rater is nested within an exercise. A complete test would partition variance by assessors, exercises, and dimensions, which would require that all assessors view all assesses in every exercise. (Jones, 1992).

Research on convergent and discriminant validity of assessment centers is in a bind. Past research was biased toward showing low discriminant validity. Full mathematical exploration would require violating standards of practice: over exposure of assessors, questionable use of within-exercise dimensions ratings, and equal weighting of dimensions observed across exercises. Perhaps it is time to conclude that traditional construct validation techniques are inappropriate for assessment centers (Joyce et al., 1994).

Tasks vs. Dimensions

Because of the research citing strong exercise effects and low discriminant validity for dimensions, some people have called for using tasks instead of dimensions as the organizing categories in assessment centers (Robertson, 1973; Russell & Domn, 1995; Sackett & Harris, 1988). Operationalization of this idea, however, creates a number of problems.

To illustrate, Russell and Domn (1995) analyzed a center where assessors rated seven tasks that corresponded to job duties. Yet there was no evidence that the assessment center exercises measured tasks such as "maintaining equipment and machinery" or "maintaining safe/clean work areas." Similarly Joyce, Thayer, and Pond (1994) had assessors make ratings from regular assessment exercises on seven common managerial functions, including recruitment and selection. Evidently assessors in
both these studies made inferences about performing particular job functions from behavior they saw in the assessment exercises. But isn’t it a greater leap of faith to go from in-basket performance to “maintaining safe/clean work areas” than to “planning and organizing”? Or from performance on a case analysis to “recruiting/selection” than to “problem detection”? Moreover, how can you give people feedback about being rated low on monitoring equipment when they didn’t monitor equipment in the assessment center? In addition to their practical problems, these applications yielded no significant improvements in criterion-related or construct validity.

Academic experiments have shown some advantages to using tasks instead of dimensions. One study showed that rating tasks was no more accurate than rating dimensions, but it led to increased recall of specific behaviors (Thornton, 1992). In two developmental studies, assessors and assesses liked tasks or attributes equally well as categories and assesses developed after each kind of feedback. But in one of the experiments, the students receiving the task feedback developed to a greater extent (Thornton, Kaman, Layer, & Larsh, 1995). Over time the assessors giving task feedback were more specific, less evaluative, and more original in their comments than those giving attribute feedback. These findings suggest that dimensional feedback can be improved by supplementing it with specifics, such as performance on key behaviors.

There are a number of practical reasons why assessment center users prefer dimensions based on human attributes rather than tasks. Lists of tasks can be long and generalize to fewer situations. Tasks are an unnatural way to describe people and are less meaningful than attributes for developmental feedback (Thornton, 1992). Important to psychologists, task descriptions have little explanatory power. Landy and his colleagues made a similar argument about job analysis; psychologists should be studying human qualities, not tasks (Landy et al., 1995).

To summarize, researchers continue to be troubled about what makes assessment centers work, although the discriminant validity problem is exaggerated by inappropriate methods for measuring it. The alternative of abandoning dimensions to rate tasks has definite disadvantages for practice. A lack of discriminant validity poses no practical problems for selection, but people need accurately rated dimensions if they are to guide development (Archerbeau, 1979; Thornton, 1992).

Tweaking dimension ratings by assessor training or different scoring methods probably won’t improve the situation much. More can be gained by the selection and use of dimensions. Specifically (a) select fewer but more observable dimensions, (b) cover the important domains but don’t expect sharp differentiation of dimensions that are subtle variations on the same theme, (c) rate dimensions only after you have enough behavioral evidence to do so, and (d) define dimensions clearly and unambiguously and add key behaviors for specificity and structure.

**CAN IT CONTINUE TO WORK?**

Critics have attacked assessment centers for using models based on organizations of the past when the role of the manager is changing dramatically (DeVries, 1993). Today managers must handle new initiatives like total quality management, employee involvement, exceptional customer service, international competition, and organizational designs and strategies that call for different competencies. Critics imply that assessment centers don’t capture these qualities.

**Dimensions Then and Now**

The preceding criticism falsely assumes that assessment center dimensions have remained static over the years. To the contrary, DDI’s 1992 compilation of dimensions (Byham & Associates, 1992) differed significantly from that of 10 years previous (Byham, 1982). Counting all types and levels of dimensions, the 1982 list had 49 entries, and the 1992 list had 113. The more recent compilation stemmed from hundreds of job analyses in various organizations and involved more than 120,000 respondents.

Many of the new dimensions came from “future panels” of senior executives, who provided a vision of their organization’s destiny and consequent requirements for human effectiveness in the 1990s and beyond (Byham & Associates, 1992). These dimensions include, for example, customer service orientation, collaboration, strategic planning, and team development. A benchmarking study of human resources practices across 21 leading corporations (Sitzer & Slider, 1994) showed that emerging competencies or dimensions include adaptability, driving change, integrity, team leadership, initiating innovation, cross-functional integration, conceptual thinking, being customer-driven, perseverance, and entrepreneurial skills. These new dimensions keep assessment centers and their organizations focused on the future.

**Evolving Simulations**

As dimensions change, so must exercises. The traditional in-basket, for example, views the manager as a middle person communicating between senior management and employees. Assessees have position power and are expected to make decisions about issues such as delegation, control, and who to send to corporate training. This approach doesn’t fit recent trends like participative problem solving, teams, initiation of major process changes, subordinates’ taking responsibility for their own
development, matrices of reporting relationships, temporary work forces, or telecommuting (Schneider, Huck, Seegers, & Ashworth, 1994). Exercise designers must anticipate such changes to keep simulations focused on the appropriate dimensions.

A major determinant of the type of exercises needed for an assessment center is the level of employee empowerment in an organization. High involvement revolutionizes the way leaders and associates relate to each other and to the outside world. Organizations that emphasize empowerment require exercises and dimensions that take this approach into account.

International Competence

Assessment center practitioners have used several strategies to tackle measurement of international competence, which is increasingly important in a global economy. International personnel may be expatriates with a multinational company or locally-based employees with international responsibilities and involvement. Both need special competencies beyond generic managerial or professional skills.

As one example, Twisk (1994) identified international interpersonal competencies as managing international teams, relationship-building skills, and international negotiation skills. The cognitive domain incorporated strategic awareness, political astuteness, understanding of foreign languages, understanding of international marketing and finance, and technical/professional competence. Personality-related dimensions included interest in the local culture, open-mindedness, willingness to participate in local activities, sensitivity to different cultures, nonjudgmental attitudes, and awareness of one's own cultural background.

The assessment center process for expatriates can include assessors from both the sending and receiving culture. A network of seven European countries constructed exercises and dimensions in different languages that reflected what is typical to each country. Videotapes of candidates allow assessors in both the sending and receiving countries who are trained in the appropriate cultural model to compare their ratings of the candidate (Twisk, 1994).

Cultural adaptations can be surprisingly subtle and complex. The World Bank, with employees from 130 countries, struggled to set up an assessment center that would be fair and acceptable to all users. Designers of the centers wrestled with definitions of standards, culturally neutral names, and issues like the meaning of being "on time." Some cultures they had to consider even rejected the word "development" if applied to people rather than, say, water systems. The designers ultimately defined dimensions in terms of actions taken, ignoring the manner in which they were executed. For example, "influencing others" emphasized clarity in getting ideas across, not being confident or assertive. Assessors excluded from their dimensional ratings consideration of nonverbal behavior, such as smiling and eye contact (Tyler & Butler, 1993).

Assessment centers also need to take into account varying norms across cultures. To illustrate, researchers compared dimension ratings of applicants for team member positions in high-involved assessment centers in the U.S., Mexico, the United Kingdom, Portugal, and Canada. Assessors rated U.S. workers more favorably on problem identification, problem solution, and team influence but more poorly on quality than those in the UK and Portugal. Mexican workers exceeded U.S. candidates in work pace (Cosentino & Lehman, 1995). Assessors across six companies rated Japanese middle managers as strong in listening, energy, and judgment but weaker in planning and organizing, leadership, and persuasiveness (Umeshima, 1992).

Cultural norms create problems for multinational organizations that want to live by one set of standards. Texas Instruments trains assessors to provide culturally appropriate feedback while still maintaining overall company norms. Feedback reports focus on skills and the development needs of participants rather than normative comparisons (Tuggle, 1996).

Assessment Centers to Lead Change

Increasingly, organizations use assessment centers to lead change. Needless to say, this practice is irreconcilable with critics' conceptualizations of assessment centers as behind the times. To illustrate, a Dutch hospital used assessment centers to facilitate a turnaround after a merger (Van Woerkom & Feltmann, 1993). A customer service organization employed the method to revamp the position of customer service manager as more marketing focused, responsive, and proactive (Bobrow, 1996). A jet engine manufacturer used three developmental assessment centers as a follow-on to a reengineering effort (Fleisch, 1995), and a Silicon Valley high technology firm and a telecommunications company used assessments to implement strategic organizational change (Adler, 1995).

Assessment centers lead change by providing a way to define and measure the competencies and supporting behaviors that people need to meet future challenges. If employees are expected to operate cross-functionally, think in terms of process, make continuous improvements, act in an empowered fashion or empower others, broaden their job responsibilities, focus more on customers, or change the thrust of their jobs, then they need to understand how their behavior must change. The assessment center helps organizations to either select people to match new requirements or provide incumbents with diagnostic information that will guide their development.
To summarize, assessment centers are fundamentally plastic. They can adapt to new requirements, and they have. Users should modify dimensions and exercises to reflect current and future needs, including international competence. Assessment centers not only don’t lag change, they can help organizations lead change.

WHERE CAN IT HELP ORGANIZATIONS?

Organizations traditionally used assessment centers for selecting people with potential for lower- and middle-management jobs. Some writers argue that the method is inappropriate for executive positions, which have only a few candidates. Organizations are said to need a more precise tool for senior jobs, which is why the assessment center has had little impact on the executive suite (DeVries, 1993).

The following review of information about executive selection provides the first response to this challenge. But the criticism implies that assessment centers have little use beyond lower- and middle-level management selection; later sections address this issue.

Executive Assessment

It is incorrect to say that assessment centers were never used for executive selection; one such center appeared 35 years ago. In 1961, Southern Bell ran a Higher Level Management assessment center to evaluate the potential of upper-middle managers for positions at the officer level using the 26 dimensions created for the Management Progress Study. But in past years executive centers have not prevailed. A survey of assessment center users showed that fewer than 5% used their centers to select the chief executive officer (CEO) (Bentson, Gaugler, & Pohley, 1990).

Recently, however, more companies report using assessment centers and tests to select people from top executives on down (Dobrzynski, 1996; Silzer & Slider, 1994). DDI assessed over 350 executives last year, ranging up to the CEO (R. Warrenfeltz, personal communication, September 13, 1996).

Organizations are currently turning to more professional methods for executive selection for several reasons. Applied behavioral science has in the past advanced in crises, and there is a new sense of economic emergency (Howard & Bray, 1992). Restructuring of the economy and failures in organizational performance have shaken senior managers’ complacency and aroused their interest in sound management selection methods. Because the organizational environment is changing so rapidly, companies are not satisfied with reviewing a senior manager’s past experience. They want to know how an executive will perform in emerging situations, not how well he or she performed in the past. Downsizing has exacerbated the potential damage from a poor choice, which makes a good choice decidedly cost effective. Searching for and training a top executive can cost as much as $250,000 (Dobrzynski, 1996).

Major organizational changes can also sweep senior managers into assessment centers. Plant start-ups, for example, often involve assessing all levels of the organization from the CEO down (Rogers, 1992).

Executive assessment centers are not identical to those for lower and middle levels. It is true that managers who were adept at making lower-level decisions are more likely to render effective executive judgments than are those who fumbled supervisory decision making. This is why performance in Management Progress Study middle-management in-baskets was related to promotion to the executive level 20 years later (Howard & Bray, 1990). But successful middle-management decision making represents insufficient skill for handling executive-level decisions, which is why new dimensions and exercises are called for.

Executive assessment centers reflect the scope of senior-level positions and the special competencies that set executives apart from other managers. The same dimensional domains apply across management levels, but the dimensions themselves vary qualitatively. In an executive center, for instance, professional knowledge might include global awareness, customer service might encompass a marketplace focus, and planning might require strategic vision (Kinnard & Morrison, 1994).

One description of managerial scope ranges from managing tasks (individual contributor), managing through others (team leader), managing accountabilities (function manager), and managing multiple functions (general manager), to managing multiple businesses (corporate executive). Many dimensions or competencies link to this same continuum. For example, working with others expands from teamwork to team leadership, team development, talent pool development, and global development (Sloan, 1993).

Exercises for executive-level assessment centers reflect the strategic and momentous decisions that senior managers confront. In one program candidates discuss business ventures in a management team meeting and prepare and present to their boss a strategic analysis of a turnaround situation in another department (Sloan, 1992). In another center, role players interrupt candidates and challenge their judgment in dealing with a unionization threat, a sexual harassment lawsuit, and an upset senator’s plea about a plant start-up (Rogers, 1992).

Past indifference to assessment centers for selecting senior managers stemmed from the times, not the method’s inadequacies. Organizations today seek reassurance that they are placing the right people in positions that are make-or-break in a fast-moving global economy. Advantaged by
their comprehensiveness and plasticity, assessment centers are sweeping into the executive selection spotlight.

Nonmanagement Levels
Assessment centers have also moved into the lower end of the employee hierarchy. The nonmanagement employee now assumes many more responsibilities due to the changing nature of work (Howard, 1995) and the prevalence of teams (Wellins, Byham, & Wilson, 1991). Use of assessment centers for plant start-ups has escalated rapidly. DDI alone assessed about 10,000 team member applicants over the past year (E. Elder, personal communication, September 12, 1996).

Dimensions for such centers often include problem assessment and solution, teamwork, and meeting participation. Team member simulations elicit behaviors to measure these dimensions. To illustrate, entry-level brewer workers at Adolph Coors company participate in a leaderless group discussion about problems commonly faced by work teams (a team member overlooking errors, abuse of sick leave, reluctance to adopt new procedures, and violation of a safety regulation). In a problem-solving exercise the team makes selection and placement decisions among 14 candidates. Participants in a production exercise analyze problems in a mock power transfer unit and correct problems in the apparatus like incorrect gear ratios, misaligned belt wheels, or missing nuts and bolts. The task requires no technical knowledge, but assesses must share information they have been given about the apparatus and safety rules (Thornton, 1993).

New Purposes
Assessment centers have expanded not only across new organizational levels, but also into new human resource functions. Beyond traditional selection and promotion, organizations use assessment centers in the following ways.

1. Recruitment. Assessment centers provide realistic previews of jobs, as to acquaint engineers and scientists with the role of manager (Thornton, 1992).

2. Placement. After a diagnosis by assessment center, new recruits are placed in areas that will capitalize on their greatest strength or shore up developmental needs. Placement decisions may also concern assignments to temporary task forces and virtual teams.

3. Development. The assessment center diagnoses strengths and skill deficiencies to help plan appropriate training and development.

4. Proficiency assessment. A center diagnoses proficiency in skills for accreditation, such as for science teaching certificates (Jacobson & Pecheone, 1993).

5. Organization development and human resource planning. Assessment centers help managers work more cooperatively as a team, support organizational change efforts (as mentioned earlier), and identify general skill deficiencies that may require hiring replacements or global training and development.

6. Career and succession planning. Assessment centers identify competencies associated with key positions, and the individual or organization uses the information to plan development and movement.

One use of assessment centers that providers often shun is downsizing. Aside from the fact that it might earn the assessment center a nasty reputation as an "assassination center," such a practice is ethically irresponsible if organizations use a few exercises for termination decisions without attention to past contributions of the individual and other relevant factors. Given that 90% of discrimination lawsuits now concern termination decisions (Sharf, Jones, & Copus, 1996), this usage is particularly unwise. But assessment centers can contribute to the positive side of organizational change by identifying the most appropriate candidates for new organizational ventures.

To summarize, assessment centers have come a long way since their initial use for selection and promotion decisions about lower- and middle-level managers. They have spread across organization levels from team member to CEO and into new human resource applications. Organizations increasingly focus their attentions on human competencies, sometimes organization-wide competencies. These initiatives speak the language of assessment centers. As the need grows to evaluate people against competencies, organizations undoubtedly will call upon assessment centers to help.

CAN USERS AFFORD IT?
Assessment centers are costly and cumbersome. It takes a lot of time and funds to train the assessors, operate the centers, and turn around the results. These problems intensify as the pace of business quickens and organizations downsize, shedding potential assessors and personnel staffs. Even assessment center proponents acknowledge its practical problems—expensive, time-consuming, labor intensive, and difficult to manage (Dulewicz, 1991; Thornton & Byham, 1982).

Assessor training is one time-consuming process that users have attempted to streamline. Better forms and more efficient procedures, such as computerized programs to teach the rating of dimensions, have shortened and improved the process. Having assessors specialize in particular exercises likewise reduces training time.
Technology offers the most promising means for making assessment centers more user-friendly. Designers use audiotapes, videotapes, videodisks, and computers to save time and expense. They often find that they gain other benefits, such as greater realism for assessees and enhanced reliability among assessors. According to a recent survey of 33 assessment center users in the U.S., UK, and New Zealand, 30% of respondents used technological aids for recording behaviors; 21% had computer-aided scoring, and 33% had technologically-based simulations (Van Katwyk, 1996). Following are some examples of how users apply these technologies.

Scoring and Reporting Data

Scoring technologies include computerized forms, computerized behavior checklists, and group decision support systems (Cowart, 1992; Van Katwyk, 1996). Designers have streamlined the in-basket—a major usurper of assessors’ time—by defining response alternatives and providing systematic analyses of each item. These efforts both shortened the time commitment and added reliability (Jones, Herriot, Long, & Drakeley, 1991). Some have tried multiple-choice in-baskets (Morris, 1991), but this approach changes the in-basket into another type of instrument with unpredictable consequences. It is no longer a sample of spontaneous behavior.

Assessors use computer software to help them record, classify, and rate behaviors by dimension. The computer then uses this information to write drafts of exercise and final reports (Cosenzino, 1991). A comparative study using video protocols found no significant differences between handwritten and computer-generated reports in numbers of questions, key facts, or recommendations. Format also did not affect dimension ratings. The computer program cut in half the time required to produce the report. Assessors reported lower levels of stress, fatigue, and difficulty as well as increased confidence in their work (Lepard & Cowart, 1993).

The computer can store other data for easier computation and collection of norms. Statistical rather than clinical integration of data is another time-saving approach. This idea is not new (Howard, 1974), but users have seldom tried it because of their fear that assessors’ important insights would be lost in the routine application of statistical formulae. The extent to which an integration session motivates greater conscientiousness among assessors is unknown (Byham, 1992) but is a possible concern. Many assessors report a potent sense of satisfaction from putting the evidence together and creating a holistic view of the assesees.

Statistical integration requires high reliability in the assessment process, so it is not without risks, especially under rapidly changing organizational conditions. Nevertheless, the latest assessment center guide-
his or her next move. These kinds of exercises are not real behavioral simulations. According to the assessment center guidelines, a simulation should require participants to respond behaviorally to situational stimuli (Task Force, 1989, p. 7). Video exercises of the “what would you do?” variety normally do not require assessees to demonstrate behavior, only to indicate their behavioral intentions. Though it is quite legitimate to use such exercises in an assessment center, if no other exercises require actual behavior, the process cannot legitimately be called an assessment center.

The “what would you do?” or behavioral intention type of exercise is not new to assessment centers. In fact, the Management Progress Study contained such an exercise, later dropped, called Personnel Problems. This written exercise briefly described situations and asked assessees to indicate their probable response in a multiple-choice format, bringing the exercise to the level of a paper-and-pencil test. Motowidlo, Dunnette, and Carter (1990) established the validity of such a low-fidelity exercise, but it was a unique application, carefully developed in a time-consuming process of gathering managers’ input and ratings of possible actions. Such exercises lack the face validity of behavioral simulations.

Other exercise designs measure spontaneous candidate responses. In a safety simulation delivered by videodisk to candidates for certification in teaching science, examinees observe a classroom science activity on a computer monitor. Using a mouse pointer, microphone, and keyboard, they identify and verbally respond to a series of safety violations in “real time.” When they observe a violation they stop the activity, focus on the hazard, and offer an appropriate preventive or corrective action as if they were teaching the lab. The computer records their responses, but a human must later review, analyze, and score them (Jacobson & Pecheone, 1993).

A simulation used to train police officers in the judgmental use of force comes one step closer to natural behavior. It not only lets the assessees respond naturally, but also simulates a counterresponse (Van Horn, 1996). The officer views a videotape of out-of-control behavioral situations and responds to people on the screen with verbal commands and actions. The assessees can employ laser-equipped weapons, including firearms, mace and pepper sprays, and batons, and a computer measures response accuracy. The instructor selects a response from several scenarios that might have resulted from the officer’s action and displays a video of that outcome. After several simulated interactions, the person on the screen may calm down or may draw a hidden firearm and shoot the assessees.

Disassembling the “Center”

Recorded responses to simulations lend themselves to taking the “center” out of assessment centers. Assesors need not be present if simulations can be videotaped, and they can rerun a tape to verify hypotheses. The assessment process can be disassembled into parts with assessees and assessors scheduling activities at their own convenience. Final reports using videotaped behaviors were shown to be comparable to those with live observations. An added benefit is that assessors using the video format reported lower levels of stress, fatigue, and difficulty (Leopard, Edgemon, & Burns, 1990).

If assessment activities can be recorded and transmitted easily, assessors don’t have to be located near the assessees. This practice facilitates the use of professional assessors, who are increasingly in demand as organizations outsource staff functions. DDI, for example, used to train staffs in client organizations to run their own centers. Today the company additionally employs over 100 professional assessors who evaluated approximately 18,000 people during the past year (E. Elder, personal communication, September 12, 1996). Experienced external staff require less training for each new center and are more reliable, although they lack the same job understanding that comes naturally to assessors within an organization.

CAN MULTIRATER ASSESSMENTS SUBSTITUTE?

The popular multirater or 360° assessments also diagnose individual strengths and development needs, and some writers have suggested that they are cost-effective alternatives to assessment centers (Nowack, 1993). The two methods have other elements in common: both use dimensions or competencies, have multiple raters, and try to capture concrete job behaviors. Despite these similarities, they are neither equivalent nor interchangeable. The following sections summarize a more complete treatment (Howard, Byham, & Haunstein, 1994) of why the two methods cannot replace but can complement one another.

Comparisons of Multirater Assessments and Assessment Centers

Multirater assessments and assessment centers are contrasted in Table 2. The two methods differ substantially in behaviors observed, characteristics of observers, and types of observations.

Behaviors. Assessment center exercises measure proficiency to execute a prescribed task. Assessees are usually highly motivated, and, if not unduly stressed, can exhibit their best or optimal performance. Multirater assessments measure perceptions of skill usage in the natural work environment. Targets’ motivation may not be as strong and external constraints may dampen their job behavior, so observers see their typical rather than optimal performance.

Assessment centers can simulate future jobs and working conditions, providing another distinction. Multirater assessments normally focus on
### TABLE 2  Comparison of Multirater Assessments with Assessment Centers

<table>
<thead>
<tr>
<th>Multirater Assessment</th>
<th>Assessment Center</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Behaviors</strong></td>
<td></td>
</tr>
<tr>
<td>On-the-job</td>
<td>Elicited by simulations</td>
</tr>
<tr>
<td>Performance</td>
<td>Proficiency</td>
</tr>
<tr>
<td>Typical</td>
<td>Optimal</td>
</tr>
<tr>
<td>Current assignment</td>
<td>Future assignment</td>
</tr>
<tr>
<td><strong>Observers</strong></td>
<td></td>
</tr>
<tr>
<td>Untrained</td>
<td>Trained</td>
</tr>
<tr>
<td>Different standards</td>
<td>Common standards</td>
</tr>
<tr>
<td>Low interrater agreement</td>
<td>High interrater agreement</td>
</tr>
<tr>
<td>Familiar with target</td>
<td>Unfamiliar with assesse</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td></td>
</tr>
<tr>
<td>Much information irrelevant</td>
<td>More information relevant</td>
</tr>
<tr>
<td>Competing demands</td>
<td>Focused on appraisal task</td>
</tr>
<tr>
<td>Diffuse</td>
<td>Concentrated</td>
</tr>
<tr>
<td>Extended time</td>
<td>Single time</td>
</tr>
<tr>
<td>From memory</td>
<td>Immediate</td>
</tr>
<tr>
<td>More halo</td>
<td>Less halo</td>
</tr>
</tbody>
</table>


The current assignment. Even if raters project behavior on future jobs, they must base their observations on the person’s current job performance.

**Observes.** Multirater observers usually receive no special training for their task. They are apt to use different standards and often don’t agree with each other about the same target. Assessors receive several days of concentrated training on how to rate dimensions and categorize behaviors. This training drives them toward common standards and they generally agree well.

Multirater observers are susceptible to distortion of their ratings because they are familiar with the person being rated. They may be influenced, for example, by the popularity of the target. Assessors are disqualified from rating those they know too well, which preserves objectivity.

**Observations.** Assessors focus on observing situations designed to elicit behaviors related to a few known dimensions. Multirater observers are less focused; they must consider a wide variety of frequently irrelevant behaviors extending over long periods of time. Assessors make their observations in concentrated periods where there are no other tasks to be considered; multirater observers must rely on their memories of behaviors they might have observed when there were many competing demands on their time. The combination of a lack of training, susceptibility to bias, and unfocused observations results in multirater respondents giving similar ratings across dimensions. Assessors are better able to differentiate among behavioral characteristics and their ratings have less halo.

Clearly, assessment centers are more precise than multirater methods for evaluating individual strengths and developmental needs. This results from their concentrated focus, common standards, and constructed opportunities to demonstrate skills and competencies. Multirater assessments can accumulate observations about how behaviors are exhibited in natural situations, but the tool is not as sharp.

**Using Both Methods**

Even though multirater assessments are less precise than assessment centers for measuring competencies under optimal conditions, there are some advantages to using the two methods together. One reason is to provide a validity check. For assessment centers it is one approach for establishing the meaningfulness of dimension ratings. Moreover, higher management and individual assesses will have more confidence in the findings if both methods agree on relative strengths and development needs.

But disagreement between the two methods doesn’t necessarily mean that one is inaccurate. Users who combine the two methods have an opportunity to explore conditions that interfere with optimal performance. If, for example, an assessee exhibits strong organizing skills in an assessment center but is rated poorly on this quality on the job, there may be chaotic job conditions that need attention. Contrarily, if the same individual is considered acceptable at organizing on the job but is sloppy and careless with the assessment center’s in-basket, the organization might want to explore compensatory mechanisms on the job, such as an efficient assistant. In both cases of disagreement between the two methods, the organization learns more about the individual and develops hypotheses to test.

Individuals who undergo both multirater assessments and assessment centers should acquire strong motivation to change. Assessment center feedback appears objective, detailed, and professional. When assesses learn they have development needs, they have rational motives to change. Multirater assessments may be less accurate but people react viscerally to "seeing myself as others see me." Those who receive
consistent developmental messages from these two sources of feedback have both a rational and emotional impetus to change.

Still a fourth reason why users gain from using assessment centers and multirater assessments together is that some dimensions are better evaluated by one method than the other. Assessment centers have an advantage for situations like the following.

1. **Unobservable processes.** Assessment center procedures can compensate for assessors’ inability to observe what is going on in people’s heads, as in activities measuring analysis, judgment, or planning and organizing. For example, assessors often interview assesses about their reasoning in an in-basket. Computerized in-baskets or problem analyses can also capture problem-solving strategies.

2. **Skills that can be isolated.** Competencies relevant to specific, bounded situations are readily simulated in assessment centers and can be measured in their optimal state. These include dimensions like meeting leadership, oral presentation, or scheduling.

3. **Processes that extend over time before completion.** An assessment center can build in long-range perspectives for dimensions such as strategic planning or developing organizational talent that would be difficult for observers to capture in the natural environment.

4. **Integrative processes.** Processes that integrate various activities or parts of the organization can be simulated. For example, assessment center exercises can call upon strategic leadership or extra-organizational awareness that independent natural observers would see only in part.

On the other hand, multirater observers have an advantage over assessors in evaluating attributes that are more difficult to elicit in simulations, such as natural levels of motivation or day-to-day behaviors. These include sustained effort (e.g., work standards, energy), types of work motivation (desire to advance or be recognized), ongoing relationships (collaboration, behavioral flexibility), and behaviors that go beyond specified job requirements (sensitivity, integrity).

### ASSESSMENT CENTERS FOR THE 21ST CENTURY

Assessment centers have taken many turns since their inauguration in business in the 1950s. First recognized as techniques for management selection, they have evolved into tools for organizational change and multiple human resource functions.

Over the years the assessment center has responded vigorously to many challenges, summarized in Table 3. The progression of challenges affirms that new answers bring new questions, as illustrated in the third column of Table 3. As assessment centers move forward, they offer many

#### TABLE 3 Responses to Assessment Center Challenges and New Questions

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Response</th>
<th>New Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is it?</td>
<td>Defined by 1989 guidelines.</td>
<td>Will assessment centers remain intact if the selection process has multiple hurdles?</td>
</tr>
<tr>
<td></td>
<td>Must have job analysis, behavioral observations put in categories, multiple techniques including simulations, multiple assessors, and systematic data gathering and reporting.</td>
<td></td>
</tr>
<tr>
<td>Does it work?</td>
<td>Yes. Solid evidence of criterion-related validity.</td>
<td>Can it work even better if constructed optimally?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What organizations does it help?</td>
<td>Apparently unlimited. Used in industrial, education, military, government, professional, and religious settings. Spreading globally.</td>
<td>Do all of these applications work equally well? What situations benefit most and least?</td>
</tr>
<tr>
<td>Will users like it?</td>
<td>Yes, especially compared to paper-and-pencil tests. Applicants find it face valid and fair; assessors, assesses, and management usually positive.</td>
<td>Will high technology make it even more attractive to management and assesses?</td>
</tr>
<tr>
<td>Can simpler methods substitute?</td>
<td>Depends on the use. For selection, more research is needed on independent contributions of components, but simulations appear to add value. For diagnosis for development, assessment centers are hard to beat.</td>
<td>Can we be more precise about what different tools measure and where to use them best?</td>
</tr>
<tr>
<td>What makes it work?</td>
<td>No compelling replacement for traditional explanation (observing behaviors and classifying them into human attributes related to jobs in question). Discriminant validity is weak if dimensions are rated within exercises, but results confounded with single rater error. There may be an exercise effect that is not error.</td>
<td>Do dimension ratings relate meaningfully to similar aspects of performance? Can key behaviors enhance the accuracy of dimension ratings and boost convergent and discriminant validity?</td>
</tr>
<tr>
<td>Can it continue to work?</td>
<td>Yes. It is fundamentally plastic and can be readily adapted to new dimensions, exercises, and international applications. Now used to lead change.</td>
<td>How do the various innovations affect reliability and validity?</td>
</tr>
</tbody>
</table>

(continued)
TABLE 3 (continued)

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Response</th>
<th>New Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where can it help organizations?</td>
<td>Across all levels from non-management to executives; in any human resource function that needs diagnostic information on individual or group performance.</td>
<td>Is the method more valid at some organization levels compared to others? How do different components contribute to prediction at different organization levels?</td>
</tr>
<tr>
<td>Can users afford it?</td>
<td>Looking better all the time. New technologies are bringing down costs, time, and effort, but measuring interpersonal skills electronically is still a challenge.</td>
<td>What effects do high-technology solutions (scoring, simulations, integration) have on outcomes? What is the impact of professional assessors?</td>
</tr>
<tr>
<td>Can multirater assessments substitute?</td>
<td>No. Assessment centers are much more precise for measuring optimal performance, but the two methods can complement each other.</td>
<td>Does using both methods better motivate development? Where are the two methods most likely to disagree and why?</td>
</tr>
</tbody>
</table>

opportunities for innovative practice and informative research. Following are some projections and recommendations that expand on Table 3.

**Disaggregation**

For personnel selection, different measurement methods have unique advantages. Paper-and-pencil tests, including mental ability and personality-tests, are applicable to vast numbers of people at once and are economical to administer and score. But they are signs and not samples of behavior. Determining if people can really perform the functions of a job or role begs for examples of behavior. Examples from the past are useful because past behavior so often predicts future behavior. Thus interviews, experience-based biodata instruments, and multirater assessments add to the pool of knowledge about past behavior. Simulations provide a more precise analysis of behavior in current tasks. Even more important, they offer a preview of how people will behave in future situations. They are particularly useful for diagnosing interpersonal and administrative skills. Research may show that each measurement method can add value in predicting future performance.

Selection systems that must discriminate among large numbers of candidates frequently use several measurement methods in multiple hurdles, allowing only the most promising candidates to reach the more expensive techniques at the end. Prescreening with tests, interviews, or video-based, computer-scored measures whittles down the number of candidates for assessment center simulations.

Whether or not a multiple-hurdle selection process in its entirety is identified as an assessment center depends on whether assessors take into account all the information it produces when they rate dimensions. If, for example, a cognitive ability test is a prescreen but assessors never review the scores when rating dimensions, the test is not part of the assessment center. On the other hand, if assessors use the test results to rate a dimension such as applied learning, then the test is part of the assessment center.

Users may choose to have the assessment center include only simulations, but there must be enough of them to rate the dimensions and the process must have enough features to qualify it as an assessment center. Otherwise, the selection system is just borrowing a piece of assessment center methodology. There are, then, at least three possible multiple-hurdle models:

**Model 1: Selection with simulations.** One of the hurdles includes one or more simulations. Assessors view behaviors and give an overall exercise rating or rate a few dimensions. There is insufficient coverage of key dimensions for the position and/or no pooling of data. A decision-making committee brings together information from the various hurdles, including the simulations, and makes a go-no go decision. This process borrows assessment center methodology but is not a true assessment center.

**Model 2: Selection with an assessment center as one hurdle.** One of the hurdles is a bona fide assessment center. It may be based on simulations alone or simulations with some other tools, such as behavioral interviews. Assessors rate key dimensions, pool their data, and otherwise follow the assessment center guidelines. A decision-making committee brings together information from the various hurdles, including the assessment center, and makes a go-no go decision.

**Model 3: Assessment center within an assessment center.** Model 2 is nested within a larger selection process that also follows the assessment center method. The decision-making committee comprises assessors who may or may not have served as assessors in the embedded center. The committee brings together information from the various hurdles, including the embedded assessment center, and uses all measures to rate a final set of dimensions. The dimension ratings form the basis for a go-no go decision.

These variations on the multiple-hurdle model thus cycle back to the first challenge to assessment centers: What is it?
Plasticity

Assessment centers are explicitly open to innovation; their definition is broad. Users (and critics) should treat them as modular living space, not Victorian parlors (Howard, 1993). Practitioners must constantly keep their eyes open for organizational changes that will impact assessment. As jobs, roles, and organizational strategies change, so should dimensions and exercises.

Modifying the assessment center model has its pitfalls. Designers can generate an infinite number of dimensions, but assessors can reliably rate only a few. Centers will function poorly without proper attention to fundamentals such as proper assessor training and quality checks (Schmitt et al., 1990). Flexibility should not be an excuse for carelessness.

Assessment center practice is changing so rapidly that the amount of research needed to support these changes is staggering. Yet published assessment center research is concentrated in academic studies, with no assurance that findings will generalize to organizational realities (Byham, 1992). Whenever significant modifications are made to an assessment center, users should initiate research to (1) examine the reliability and validity of the new center, and (2) compare the old with the new.

Technology

Automated methods will continue to advance assessment centers. Personnel selection may take place from kiosks that permit audio, video, and data transmissions by computer. The Internet is an obvious carrier.

Assessment exercises will also benefit from innovative automation. Computerized simulations for cognitive and administrative tasks hold much promise as aids to collecting, scoring, and transmitting data. Simulations can be more like adaptive testing with branches based on prior decisions and actions.

The ultimate interpersonal simulation may be virtual reality (Howard & Bray, 1992). Candidates can behave in a simulated world and their movements carefully recorded. There is still the problem of how to respond to the assessees' behavior without human intervention or a multiple-choice format. It will require much better voice recognition systems than we have today, high-powered interpretive protocols, and extensive branching and storage capabilities. They may be possible in the 21st century.

Individual Development

One consequence of the changing nature of work is greater individual responsibility, from making business decisions at lower levels to managing one's own development and career (Howard, 1995). This emphasis will inspire individuals to seek assessments with direct ties to development. Emerging self-development programs can provide guidance and learning aids that can be delivered on platforms such as intranets.

Meaningful diagnosis for development requires assessment feedback that is quite specific. This means that such assessment centers will rely heavily on simulations because of the behavioral detail they can exhibit. Key behaviors can capture this specificity and serve as learning models for a dimension. Granular training exercises at the key behavior level help people focus more precisely on areas that need development.

People responsible for their own self-development will need coaches and feedback givers. This necessity builds a natural bridge from assessment centers to multi-rater assessments. It also embeds the assessment center within a larger development process.

CONCLUSION

This review uncovered no compelling reasons why assessment centers cannot continue to prosper in the next century. The upper range of criterion-related validity coefficients indicates potential for a very high payoff if assessment centers are well constructed. And their plasticity assures that they can be molded to address new needs that arise over time.

The assessment center of the next century may expand into different forms from the traditional model. Technological innovations and outsourcing of personnel staffs are already making it less of a center and more of a circumscribed process that can be embedded in other systems. Practitioners have borrowed parts of assessment center methodology and applied them to different needs. For example, some selection systems use simulations as another test in a multiple-hurdle model, other selection procedures add mini-simulations to behavioral interviewing. Competency-based systems draw from the idea of assessment center dimensions, and multi-rater assessments use multiple raters and dimensions. Symbolic of these expansions, the participants at the 1996 International Congress on the Assessment Center Method voted to change the name of the conference to International Congress on Assessment Center Methods. Other spinoffs may emerge and contribute to the legacy of assessment centers.

If practitioners and researchers proceed with creativity, caution, and common sense, assessment centers and their progeny will remain invaluable tools for organization and individual development in the 21st century.
REFERENCES


The Assessment Center Process: New Directions

Phillip E. Lowry
Public Administration, University of Nevada, Las Vegas, NV 89154

The assessment center process is one of the most widely used selection processes. It is particularly important in the public sector where a recent survey found 62% of cities of 50,000 or more population use the process for fire and police promotion purposes. As the process matures, additional research shows there are problems, especially with validation of the current dimension-specific model. These concerns have begun to drive the process in new directions. This paper addresses one of these new directions—the measurement of performance in work-sample exercises (task-specific centers) rather than attempting to measure complex constructs such as leadership, judgment, etc. (dimension-specific centers). The paper describes the task-specific model and compares the two processes. It also offers suggestions for additional research.

The assessment center process for personnel selection is becoming increasingly important, especially in the public sector. Over 62% of the respondents in a recent survey of police and fire chiefs reported that they use assessment centers, especially for promotion (Lowry, 1996). The reason for this popularity can probably be best summarized by the statement, "they work," provided by an anonymous respondent to the cited survey.

The primary assessment center model in use today is generally based on the pioneering research by Douglas Bray at AT&T in the 1950s (Bray, Campbell, & Grant, 1974). Bray's model measures knowledge, skills, and abilities (KSA) derived from a job analysis. The KSAs typically include such traits and complex constructs as leadership, judgment, problem solving, decision making, etc. These are called performance dimensions. Hence I use the term "dimension-specific" assessment centers when referring to this model of the assessment center process.

Although the dimension-specific assessment center has received strong support in terms of its validity, there have been some troublesome findings that suggest all may not be well. For example, as early as 1982, Sackett and Dreher suggested that "there is virtually no support for the view that the assessment center technique generates dimension scores that can be interpreted as representing complex constructs." (p. 409).
