Assessment Centers: Best Practices for Best Results

A discussion of assessment center use and research findings to help practitioners

Written by Dan Hawthorne

Cogniphany

Table of Contents

Introduction

Walter Burke is a senior instructor in a CIA assessment center (AC) portrayed by Al Pacino in the film, "The Recruit," and his constant response about what the recruits are going to experience is, "Nothing is what it seems. Everything is a test (Donaldson, 2003)." This is a very apt and succinct—albeit overly dramatic—description of the assessment center method. The representation of an AC for the Central Intelligence Agency in the film is also a very appropriate homage considering that ACs were first used to select the right recruits to be OSS (Office of Strategic Services; later to become the CIA) agents for the United States (Thornton & Rupp, 2006). Using this method helped the agency to select new agents for its ranks, and now it has grown beyond the OSS/CIA and into the ranks of private and public business.

What is an AC?

Even though Burke's statement is overly dramatic, it likely is an accurate perception for an applicant or employee who is experiencing an AC for the first time. At a quick glance, an assessment center is a series of exercises an applicant or employee is put through to assess his or her potential abilities for the purpose for promotion, hiring, or development (Thornton, et al., 2006). But, while from an outside perspective, the process may seem to be relatively simple, a lot of careful planning and design work should go into an AC in order to garner the desired results from it. In fact, guidelines for the development and use of ACs are frequently revisited and revised to account for recent findings (Joiner, 2000). Because of this set of guidelines and how an AC is implemented, the AC isn't so much a place, as it is a method. An AC can be conducted in an elaborate corporate headquarters or very simply in a hotel meeting room; the setting is not as important as the method that is followed to design and implement the AC (Thornton, et al., 2006).

Because implementing an AC is a complex, expensive, and intensive process, the following sections will address what research has found are the best practices for designing and implementing an

AC. First, I will present some general information about the AC method, best practices, and general concerns that are standard to almost all ACs, regardless of use or implementation. Second, I will address some of the more specific concerns of each AC type. Finally, I will present some thoughts about where the future of the AC may be going and possible research that might be of benefit to explore.

The Guidelines and the AC Method

The international task force for AC guidelines presents ten features that must be present for a process to be called an "Assessment Center (Joiner, 2000)." These ten features are as follows: (1) Job Analysis, (2) Behavioral Classification, (3) Assessment Techniques, (4) Multiple Assessments, (5) Simulations, (6) Assessors, (7) Assessor Training, (8) Recording Behavior, (9) Reports, and (10) Data Integration. While all of these features are necessary for an AC to bear that moniker, one should not simply checklist these items and move on. An AC is a very expensive engagement and poor planning can be responsible for a lack of return-on-investment (Howard, 1997). Additionally, even though a number of the best practices that will be presented here have been the focus of research, which presents promising evidence of efficacy, many companies have failed to implement them (Spychalski, Quinones, Gaugler, & Pohley, 1997).

"The Guidelines," as they have come to be known, also offer some guidance about how to train assessors, validate the AC, the rights of those who participate in an AC, and what practices may be called ACs, but are not (Thornton, et al., 2006). Unfortunately, the guidelines offer limited information about how to develop specific kinds of ACs, such as developmental ACs, and this (and other reasons) will likely lead to them being revised in the near future. Thus, if one wishes to look for best practices in these applications, it would be best to be familiar with research on those specific areas.

Criticism

A major point of criticism about ACs is that we don't know why they work. They seem to present good criterion-related validity, but poor construct validity (Howard, 1997; Woehr & Arthur, 2003). Some have even found evidence that seems to point to clear results of weak construct validity (Joyce, Thayer, & Pond, 1994) For instance, some researchers have examined exercises and dimensions side-by-side demonstrating that the majority of variance in ratings is attributable to exercises, not dimensions (Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lance, et al., 2000; Schneider & Schmitt, 1992). However, others have found exactly the opposite by using generalizability theory; that is, dimensions, not exercises, are responsible for the majority of variance in ACs (Arthur, Woehr, & Maldegen, 2000). Even others have found evidence that both dimensions and exercises have significant influence on AC ratings (Bowler & Woehr, 2006).

With all of the contention surrounding the issue of construct validity in ACs, researchers have explored alternate solutions to this paradoxical situation. Following this line of research, some have found evidence that supports the possibility that criterion-contamination has caused inflated performance scores of those who succeed during ACs and deflated performance in those who do not succeed during ACs (Klimoski & Brickner, 1987). However, evidence has been presented that poor design, unreliability in assessors (which can be counteracted by assessor training), and inconsistent assessee performance can be powerful influences on why an AC fails (Lievens, 2002). This highlights the importance of using good design when developing an AC, and erring on the side of caution.

Still, other studies have been able to show evidence of construct validity using CFA and MTMM designs (Kolk, Born, & van der Flier, 2004). But, others have used the same methods and discovered evidence that AC ratings do not display construct validity (Sackett & Dreher, 1982) However, it is possible that techniques used to measure the construct validity (MTMM) of ACs are being improperly applied, and this is the reason that researchers have failed to find construct validity (Kleinmann & Koller, 1997). In addition, alternate methods like generalizability theory could provide some more accurate measurements of variance due to construct validity (Arthur, et al., 2000).

Finally, some have questioned why it is necessary to worry about construct-validity as long as ACs can demonstrate predictive and content-validity (Norton, 1977). This has even erupted into a heated debate about the topic, displaying the need for more empirical evidence (Dreher & Sackett, 1981; Norton, 1981; Sackett & Dreher, 1981). And, even to this day, the debate about what to do continues as a recent issue of Industrial and Organizational Psychology: Perspectives on Science and Practice (SIOP's official publication) highlights the current feelings of the community (Connelly, Ones, Ramesh, & Goff, 2008; Lance, 2008; Lievens, 2008; Melchers & Konig, 2008; Moses, 2008; Rupp, Thornton, & Gibbons, 2008).

Developing an AC

Implementing proper practices can be challenging for many organizations that need to develop an AC. Some have had problems selecting and defining dimensions, designing exercises, as well as training assessors, who may be the very managers who recommended assesses (Lievens & Goemaere, 1999; Spychalski, et al., 1997). Even though some problems exist, there is promising evidence that organizations are using multiple methods to validate data from their ACs, and that this practice is on the rise (Spychalski, et al., 1997). With all of this information, as well as the information from The Guidelines, there are some best practices for development of an AC that come forth.

An AC developer should ensure that all decision-makers are involved, so that there are multiple perspectives, the opportunity to gain buy-in with those involved is present, and to create a stronger sense of corporate involvement in the process (Cochran, Hinckle, & Dusenberry, 1987). This idea is also supported in training and organizational change literature, as gaining executive support for a major project like an AC helps to transmit a signal to the rest of the organization that it is dedicated to the success of the project and those involved in it. (Chiaburu & Marinova, 2005; Cohen & Bradford, 2007; Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008; Velada, Caetano, Michel, Lyons, & Kavanagh, 2007; Warr & Bunce, 1995)

Transparency—sharing of dimensions of interest with assessees—of dimensions appears to increase construct validity without a change in the mean AC ratings (Kolk, et al., 2004). A change in mean AC ratings would indicate some form of "faking" when dimensions are made transparent if this were the case. Thus, especially in developmental ACs, there is likely benefit to allowing assessees to know what dimensions they will be assessed on.

Finally, it is critical that a focused, multi-dimensional job analysis be conducted prior to beginning development of the AC, in order to design the exercises and identify the AC dimensions (Dreher, et al., 1981; Schippmann, Hughes, & Prien, 1987). Detailed job analysis is the objective foundation of any sort of exploration of how a job is constructed, selected for, or how an employee can develop skills to succeed in it. Essentially, this process allows for a scientific dissection of the internal organs of the job, and helps an HR specialist to explore what the critical elements of the job are and what an employee must bring to the job, as opposed to what can be trained. Additionally, without job analysis, the organization is vulnerable to litigation over adverse impact (Dreher, et al., 1981; Norton, 1981; Sackett, 1987; Sackett, et al., 1981).

What's the Purpose of the AC?

When developing an AC, its purpose must be of central focus. The AC's basic purpose should come from an examination of the organization's needs (Thornton, et al., 2006). At a fundamental level, one needs to know if the organization is interested in trying to select new employees, find the right people to be promoted into a new job, or if leadership needs to be developed to confront new upcoming challenges. The purpose of an AC does not need to be limited to just these few examples, but it should be defined early in the development process, because the purpose of the AC will guide many of the decisions about dimensions, exercises, transparency, etc.

When defining the purpose of the AC, the organization must understand that a critical concern to address is the expensive nature of them. Beyond the obvious expenses of development and staffing of the

AC, if existing employees are used as assessees, there will be lost work time for them (Thornton, et al., 2006). Thus, because of this high initial cost, organizations may be tempted to squeeze every bit of utility from the AC as possible. One temptation that should be avoided at all costs is that of "double-dipping" into AC results (Thornton, et al., 2006). The practice of double-dipping is when information from a developmental AC is used to make decisions about promotion or retention in a situation like downsizing. There are numerous reasons to avoid this practice. Using data in this way could open the organization up to legal risks (Thornton, et al., 2006). Since employees in a developmental AC may be more willing to make mistakes, in order to learn from them, the results will be an employee's typical performance, and not maximal performance (Jones & Whitmore, 1995). Finally, if employees were to be made aware of this practice, it could have the possibility of violating beliefs of organizational justice and result in counterproductive work behaviors, which could also include withdrawal behaviors (Rousseau, 2007)

What's Being Rated?

As a part of the necessary job analysis to develop the AC, a major question will become, "What are we trying to measure?" The answer to this is slightly different from other job analyses because of the use of dimensions in ACs. While breaking down tasks that are necessary for the job, one will need to think about what core constructs that have the potential to influence good performance on the job. Meta-analysis has examined the numerous dimensions that have been used over time and found that 6 are the most effective (Arthur, Day, McNelly, & Edens, 2003):

- Consideration/awareness of others
- Communication
- Drive
- Influencing others
- Organizing/Planning,
- Problem solving

Follow-up meta-analysis has supported these findings, and further presented evidence that using these dimensions adds predictive ability above and beyond other assessment and selections tools (Meriac, Hoffman, Woehr, & Fleisher, 2008)

Additionally, the dimensions of drive and consideration/awareness of others were found to not be effective predictors in the presence of the other predictors, which may indicate some overlap between these dimensions (Arthur, et al., 2003). Thus, it may be important to consider the necessity of these dimensions when designing the AC. The question of exactly which dimensions are most necessary for an AC becomes important because of research evidence about the number of dimensions assessors are best able to keep track of when performing behavioral observation. A number of pieces of evidence point to fewer dimensions being better than more, with three dimensions having been identified as the ideal number that assessors are best suited to observe behaviors and be able to have more accuracy in ratings (N. Anderson, Payne, Ferguson, & Smith, 1994; Gaugler & Thornton, 1989; Lievens & Conway, 2001; Schmitt, 1977).

Ideally, one also needs to address what type of behavior is being examined in the AC and if it is going to translate to the workplace. This has been addressed by discussing whether “signs” or “samples” are being assessed by ACs (Sackett, 1987). *Signs* are essentially maximal performance or the extreme ends of what a participant might display on the job, and *samples* are equivalent to typical performance, or the behavior that one would expect a participant will display on the job most often. Since people will be motivated by the outcome of a situation and goal effects, one would expect that assessees in an administrative (selection, promotion, etc.) context will be more likely to exhibit signs, and those in developmental contexts will be more likely to exhibit samples (Bandura, 1985; Bandura & Locke, 2003; Nadler & Lawler, 2007; Vancouver & Kendall, 2006). Thus, depending on the type of AC that is being developed, one needs to keep in mind whether to expect signs or samples from assessees.

What Exercises will be used?

There are a number of simulation exercises that can be chosen for an AC. While there are variants, the major types of exercises are as follows: written case analyses, oral presentations, leaderless group discussions (assigned or unassigned), role-play exercises, in-basket, oral fact-findings, and business games (Thornton, et al., 2006). One tactic that approximately 20% of organizations have explored is using an integration method with exercises (Spychalski, et al., 1997). A good descriptor of this tactic is "a day in the life (Thornton, et al., 2006)." Essentially, the exercises are pulled together into a cohesive story that is told throughout the whole assessment center, basically immersing the assessee into a day that he or she might experience on the job.

As a general rule it is better to have too many exercises than too few, as this leads to better rating accuracy (Gaugler, Rosenthal, Thornton, & Bentson, 1987). However, too many exercises can be redundant and difficult to manage. Thus, it may be better to use the "Goldilocks principle," and have just the right number of exercises. This basically gives multiple points of reference to an assesssee's performance, and allows an assessee to make up for one instance of poor performance. Past organizational research has shown that companies have used an average of five exercises for assessees (Spychalski, et al., 1997).

However, just as important to consider as the number of exercises is the level of fidelity in each exercise (Thornton, et al., 2006). That is, to what degree could you find a similar situation in the target job? For instance, even though in-basket exercises have been shown to measure management aptitude or potential, differentiating from experience or age, these exercises might be appropriate to an office manager but not a production manager in an industrial setting (Meyer, 1970; Thornton, et al., 2006). By using exercises that exhibit situational-specificity (those that reflect pieces of the target job, such as group exercises and in-baskets) an AC should tend to have more predictive validity (Bray & Grant, 1966). Finally, it is wise to pilot test any exercises with job incumbents and subject-matter-experts (SMEs) to ensure that exercises have sufficient fidelity (Thornton, et al., 2006).

How to Handle Data Integration and Scoring?

Before the AC is actually conducted, another issue that needs to be considered is how data that is gleaned from the exercises will be handled. That is, how will the exercises be scored, and how will the data from multiple assessors be handled. Some specifics that need to be addressed in regard to assessors' raw ratings are the accuracy and reliability of those scores. There is evidence that it is important to have assessors only rate one exercise, and ideally only one dimension within an exercise, as it cuts down on common rater variance (Kolk, Born, & van der Flier, 2002). However, it will rarely be reasonable to have one rater per dimension per assessee in an AC, so this advice is best taken to mean that it is better to reduce the amount of cognitive loading that an assessor has to deal with while rating an exercise.

Additionally, in some situations it may be beneficial to use a behavioral checklist, instead of other methods due to findings that this method is easier to train as well as providing a framework for assessors who may rarely be called on to be assessors (Hennessy, Mabey, & Warr, 1998). Essentially, this process helps these assessors observe and record behaviors without becoming mired in the more esoteric elements of the assessment process.

A decision also must to be made about whether to integrate data using a mechanical method utilizing statistical integration, a method that attempts to gain the consensus of assessors, or a newer idea in the field, a hybrid of the two (Thornton, et al., 2006). This decision is usually driven by the type of AC that is being used (Developmental or Administrative). Mechanical methods are more cold and calculating and may not lend themselves well to situations where feedback to assessees is a necessary part of the AC process, such as developmental ACs. However, consensus methods can be perceived as more subjective and less precise than mechanical methods, and thus may not be best for administrative ACs that are used for selection or promotion.

Evidence does exist that when assessors discuss ratings and come to consensus ratings that the final ratings are more accurate and those whose ratings were initially high are more likely to reduce their ratings, rather than increase them (Herriot, Chalmers, & Wingrove, 1985). Additionally, Group exercises show improved inter-rater reliability when consensus scores are used (.65 - .73), instead of simply using the initial scores (.67 - .84) that raters assign (A. Jones, 1981).

There are some pieces of criticism about consensus scores that cite the subjectivity of those scores and the possibility of rater biases being introduced into the scores. However, evidence supports the equality of a mechanical combination of scores, compared to having assessors arrive at a consensus (Pynes & Bernardin, 1992). Essentially, the two methods show no significant differences in predictive validity or disparate impact.

There are dangers associated with using a pure consensus method of data integration, such as the dilution of information through group discussion, group polarization, groupthink, and a failure to discuss all information (Thornton, et al., 2006). Essentially, while group discussion has benefits, in that more eyes can capture more information and that members can work to provide checks and balances on one another, it is important to consider the dark side of pure group consensus. In situations of group consensus it is very possible for the group to fall victim to the tyranny of the majority.

Some of these problems can be alleviated with how this stage of the process is addressed in training, and with the tools that are given to assessors (Thornton, et al., 2006). Tools and training should direct assessors to focus on their task as behavior recorders and move them away from making inferences about behavior or letting their personal biases influence ratings. Additionally, people have been shown to make better decisions and be better able to be self-critical when their accountability in the situation is made salient to them (Tetlock, 1983). Thus, making each assessor aware of the impact of their actions in training may be an important piece to include. Finally, it may be useful to consult literature on team design and processes for tips on how to best train teams of assessors to work together and address some of the above concerns with positive group dynamics (Jex & Britt, 2008).

When using a mechanical method, it will be necessary to derive weights for each dimension so that differential importance of dimensions can be addressed in the integration equation (Thornton, et al., 2006). Thus, a good process for deriving these weights could come from pilot-testing and validation of data through multiple regression and using regression weights (β) as weights in the subsequent equation for calculating OARs. Specifically, weights should be cross-validated if differential.

How Should Assessors be Trained?

Being an assessor is a cognitively demanding task, so it is important to prepare assessors for this, in addition to having the right number of dimensions for assessors (Gaugler, et al., 1989). Research has found that assessors who received insufficient training to perform assessor duties were more likely to provide inaccurate ratings and ratings that were influenced by personal biases (Spychalski, et al., 1997). Exercise-ratings by assessors are vulnerable to biases and rating errors, thus training is important to try to counteract any biases that assessors may bring to the AC (Shore, Thornton, & Shore, 1990). Thus, in order for assessors to be able to accurately and reliably rate the behaviors that assessees exhibit in exercises, they require training regardless of whether assessees are psychologists or managers from the organization with the target job.

The ideal form of training for assessors is known as frame-of-reference (FOR) training, because it works to gives assessors a shared point of reference for behaviors that is grounded in examining observed behaviors without inferential leaps (Thornton, et al., 2006). Frame of Reference (FOR) training has good efficacy in teaching assessors to more accurately rate behaviors, and helps them to avoid some of the biases and rating errors that can plague an AC (Gaugler & Rudolph, 1992; Goodstone & Lopez, 2001; Jackson, Stillman, & Atkins, 2005; Schleicher, Day, Mayes, & Riggio, 2002). Without training, raters have been shown to implement their own version of FOR which is built individual schema (N. Anderson, et al., 1994). Essentially, raters have their own higher-order factor that they use to define behaviors. Additionally, there are certain personality dimensions that influence raters to be more “tender-hearted” which may cause them to introduce excessive leniency into their overall ratings of assessees (Bartels & Doverspike, 1997). Using FOR training helps these assessors record and rate behaviors more accurately and reliably, and to ignore their own individual schema and personal biases.

Finally, there are some general rules that can help in training assessors and making sure that the best people are selected to be assessors in an AC. In general, psychologists make better assessors, having more valid and reliable ratings of assessees (Gaugler, et al., 1987). When managers have been used as assessors, they seem to be less apt to distinguish between multiple dimensions (Lievens, 2001b). This could be because of outside knowledge of assessees or pre-defined feelings about what characteristics are needed for a job that managers are bringing into the AC (Lievens, 2001a; Moser, Schuler, & Funke, 1999).

Strengths and Limitations

As with any HR process there are inherent strengths and weaknesses to the AC method. While there are items that need to be considered with specific uses of the AC method—and will be addressed in later sections—the following are some general considerations for the AC method as a whole.

One of the biggest strengths of the AC method is its predictive validity. While there are concerns with construct validity many pieces of research have found that ACs utilizing The Guidelines have good predictive validity (.37 to .41) in comparison to other selections devices (Arthur, et al., 2003; Howard, 1997; Meriac, et al., 2008; Moser, et al., 1999; Schmitt, Gooding, Noe, & Kirsch, 1984; Spychalski, et al., 1997). Additionally, the AC method has been found to be linked to job retention in both male and female applicants (L. R. Anderson & Thacker, 1985).

However, some research has found that AC exercises can be moderated by cognitive ability, which may increase sub-group differences (Goldstein, Yusko, Braverman, Smith, & Chung, 1998). Thus, it may be important to examine all exercises to make sure that if there are cognitive components involved in their design that they are appropriate for the job that is being assessed for. African-American assessees have been shown to receive systematically lower AC ratings (d=.52), but women (d=-.19) appear to receive systematically higher AC ratings than male assessees (Dean, Roth, & Bobko, 2008). However, ACs have been shown to cause less adverse impact than other comparable assessment devices such as self-report measures and cognitive ability tests (C. C. Hoffman & Thornton, 1997).

There is a bright side to any of the concerns about differential results for assessees though based on race or gender. Research has found that applicants tend to find that ACs are more face valid than other assessment methods (Macan, Avedon, Paese, & Smith, 1994). Thus, applicants believe that they have been given a fair opportunity to display their skills and find that the exercises in an AC are more representative of the target job than a pencil and paper test or assessment.

Finally, while ACs have a significant cost associated with them and require the dedication of human capital to implement an AC, they do show significant, reproducible results that are worth their associated costs (Thornton, et al., 2006).

Specific Concerns and Recommendations by AC Type

Because of its adaptability, there are numerous applications of the AC method that organizations and researchers have explored. Organizations have used ACs for recruitment, selection, placement, performance appraisal, organization development, HR planning, promotions, and even layoffs (Thornton, et al., 2006). In a meta-analysis of AC usage by organizations Spychalski, et al. found that the three most popular reasons for implementing an AC were selection, promotion, and development planning (1997). In this regard, selection and promotion can be grouped together for the purposes of examining specific best practices. This is because in the case of selection and promotion, assessees in ACs used for these two purposes are applicants for a new position, and will be more likely to present *signs* of behavior, instead of *samples* (Sackett, 1987). An effective way to present best practices for ACs would be to delineate between administrative applications (selection and promotion) and developmental applications. Thus, the following sections will be presented with these two major headings, and some smaller sub-headings with information for specific applications in either administrative or developmental ACs.

Administrative

In administrative ACs it is generally useful to have assessors give a final overall assessment rating (OAR) which serves as a final recommendation of selection for hire or promotion (Thornton, et al., 2006). If an OAR is used, then it will be important to validate this score with job performance data, as the OAR should be predicting that the potential new hire or promotee will be successful or unsuccessful in his or her new job. So, failure to validate these OARs could result in poor choices or disparate impact (Dean, et al., 2008).

Layoffs or "Assassination Centers"

One possible use of ACs is to select which employees will be retained when a downsizing is necessary (Howard, 1997; Thornton, et al., 2006). While this is one application of the AC method that is very useful, it also needs to be handled carefully. In fact, when this application is handled poorly it has been infamously dubbed an "Assassination Center" inside some organizations (Moses, 2008). Thus, even though such places can be saddled with a bad name, some have stated the wisdom of using them because of their ability to decrease survivor's guilt and that they have more face validity with those who go through them (Howard, 1997). That is, more applicants and assessees have reported that they felt as though they were given a good opportunity to properly present their skills and abilities (Macan, et al., 1994). Thus, in this context, application of the AC method could help to reduce legal concerns of layoffs, because of this higher face-validity than other selection tools.

Promotion

When using an AC to select employees for promotion, properly delivering feedback is one of the most critical pieces of the AC. Those who have successful performances in an AC are likely to see increased job performance, but those who have unsuccessful AC performance may have decreased work performance (Fletcher, 1991). Because of this, it may be necessary to provide follow-up with those who did not succeed at being promoted to avoid these otherwise good employees from turning over or exhibiting counter-productive work behaviors.

Additionally, during development of an AC that has the purpose of attempting to predict leadership, a 6-step recursive process has been used that has effectively predicted leadership (Brownell, 2005). This process (see Figure 1) which takes the development team through all of the necessary stages of AC construction and implementation and then advises that information from use of the AC be used to evaluate its effectiveness and retool the AC for its next implementation.

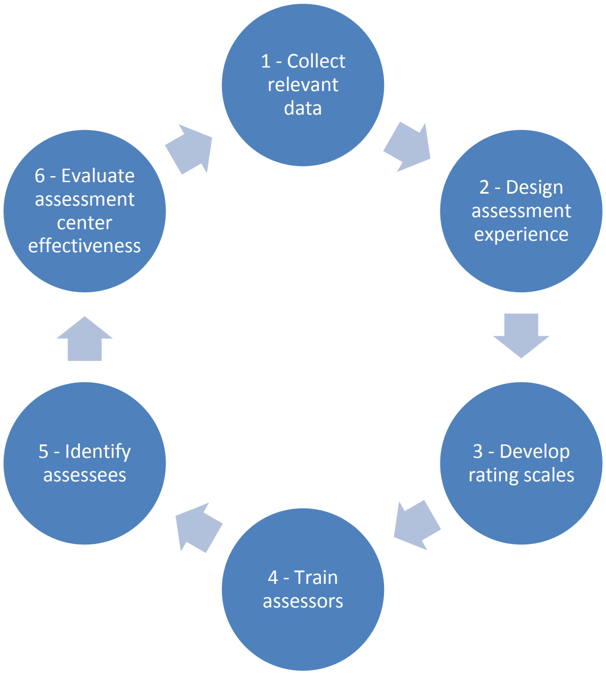
1 - Collect relevant data
2 - Design assessment experience
3 - Develop rating scales
4 - Train assessors
5 - Identify assessees
6 - Evaluate assessment center effectiveness

Figure 1. Assessment center development process (adapted from Brownell, 2005)

Developmental

A developmental AC is an AC that is more training-focused and constructed to help an organization’s employees improve on skills that are needed perform their jobs (Thornton, et al., 2006). Thus, developmental ACs serve a dual-purpose to organizations, in that they are able to assess weaknesses of participants as well as help them to develop those weaknesses into strengths (Thornton, et al., 2006). Developmental ACs have been used effectively in a number of different settings including graduate training programs (Kottke & Shultz, 1997)

A developmental AC uses a blocked design to allow for multiple stages that serve specific purposes in the AC (Rupp, et al., 2006). The first block is used to find a baseline of performance for the assessee (Thornton, et al., 2006). Then successive blocks are used to let assessees have the opportunity to practice skills that they have received feedback about and coaching on between each block. Additionally, it is important to follow-up with assessees multiple times after the end of the AC to provide more coaching and mentoring. It may even be important to provide another block of exercises 6-9 months after the completion of the developmental AC to reassess them for improvements on dimensions.

There are five major differences between a traditional AC and a developmental AC (Thornton, et al., 2006). First, there are the dimensions that are chosen for the AC. It is important to select dimensions that are developable. “Developable” means that dimensions are under the control of assessees and can be changed through feedback, coaching, and training. Surveying management in organizations where a developmental AC is being considered may be particularly effective at determining which dimensions managers believe to be developable (Rupp, et al., 2006).

The second major difference relates to assessor roles and how assessors are trained (Thornton, et al., 2006). This difference comes in because assessors have to fill the role of raters of behavior as well as the role of trainers and coaches. Because of the dual role of assessor and facilitator, they need to be trained differently. Specifically, they also need to be able to provide feedback to assessees and coach them on how to make personal improvements in deficiencies that the developmental AC identifies.

The third major difference is in the types of exercises developmental ACs need for assessees to be focused on learning and development, instead of assessment (Thornton, et al., 2006). This means that exercises need to be designed with a great deal of reality and simulate the job accurately. Because of the blocked design of DACs, assessees should be allowed the opportunity to make mistakes in simulations that they can learn from. Error-based learning has shown great efficacy in creating transfer of trained concepts into the workplace (Keith & Frese, 2008; Smith-Jentsch, Jentsch, Payne, & Salas, 1996). That is, by allowing assessees the chance to make errors in a safe environment, where they can receive timely feedback, they have a better chance of retaining new skills and bringing them into the workplace.

How feedback is handled is the fourth major difference between traditional ACs and developmental ACs (Thornton, et al., 2006). Feedback in a developmental AC is not simply for the purpose of delivering information about the assesssee's performance, but also to help him or her improve any deficiencies. Because of the emphasis on improvement, handling feedback with assessees in developmental ACs is particularly important because of how those who have performed poorly may respond to feedback. Assessees who have poor performance have shown a tendency of failure to follow-up with mentors after the AC has completed (Abraham, Morrison, & Burnett, 2006). However, research has shown evidence that supports the view that individuals who attend DACs show higher motivation (β=.13, p<.001) to advance in their careers (R. Jones, et al., 1995). Thus, with proper feedback and management of assessees after the completion of the AC, even poor performers can show dramatic improvement.

Finally, one needs to consider the development plan used in DACs, which is the foundation for assessee improvement (Thornton, et al., 2006). This development plan sets goals and gives a path to the assessee for how to improve the areas that the DAC has identified as deficient. Thus, the development plan should be comprehensive and include goals that are maintained and monitored after the DAC has completed, so that the assessee can see improvement over time (Goodge, 1994).

Training-focused

While it could be argued that all developmental ACs are “training-focused, “ a truly training-focused AC differs by having even more emphasis placed on the training of assessees in improving their performance on dimensions in the AC (Thornton, et al., 2006). The AC developer has to be even more careful about ensuring the exercises that are selected for the training-focused AC are job-relevant and trainable. For instance, in-basket exercises have been shown to be trainable, improving performance on perceptiveness, delegation, and overall performance (Brannick, Michaels, & Baker, 1989).

An additional concern in this type of AC is in not only incorporating best practices from the field of ACs, but one must also consider examining other implementation procedures from the training literature (Thornton, et al., 2006). For example, literature reviews from meta-analyses have found that many ACs simply use anecdotal reaction-based measures to gauge how well assessees have learned the presented material (Gibbons, Rupp, Snyder, Holub, & Woo, 2006). However, information from training literature has presented evidence that that these measures may be inadequate for measuring training transfer, and it may be more effective to measure self-efficacy as this has been shown to be significantly predictive ($\beta=.24$, $p<.05$) of future transfer (Sitzmann, et al., 2008).

Where are We Going?

In any field of applied research an important question to ask when one is thinking about best practices, is “Where are we going?” That is, are we making important steps, or are we just spinning our wheels in contentious debate. When considering where our next steps are, we have to consider where we are now, and think about what areas of the field need attention to advance our understanding of the important constructs. This topic is what the upcoming sections will address; where are we currently, and how can we move ahead from that point?

Where are We Now?

As mentioned previously, in a recent issue of SIOP's premiere journal, top researchers in the field presented differing viewpoints on the degree to which ACs represent dimensions or whether they are essentially a group of work sample tests (Connelly, et al., 2008; Lievens, 2008; Melchers, et al., 2008; Moses, 2008; Rupp, et al., 2008). This series of articles serves as a good stake in the ground of where we are right now. Charles Lance—who wrote the focal article—summarizes the series by stating that we should not jettison the idea that dimensions are the underlying constructs guiding employee behavior (2008). We should also seek to look for validity using more than just the MTMM model. Essentially, we need to expand how we think about validity and look at this paradox in different ways; instead of the simple checklisting of requirements (Binning & Barrett, 1989; Landy, 1986).

Lance (2008) stated that the field may have to "go back to the basics," suggesting that we need to reexamine how we think of ACs and how to accurately assess how they work. It appears that this is the direction that researchers may be going in. In a new look at ACs, researchers have begun to look at how dimensions and exercises may interplay in AC design (B. Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2009). In this paper, Hoffman, et al. presents evidence that both dimensions **and** exercises represent variance that is significant and important to addressing the construct-validity question of ACs. Thus, it would seem that they have worked in the spirit of Lance's call for research to go back to the basics, although he primarily advocated for a task-based approach (Lance, 2008).

Future Areas of Research

Some important areas of future research that need to be explored stem from the debate and discussions outlined above. Following the line of research that Hoffman, et al. have started, the field needs to look at the validity paradox with new eyes, instead of simply rehashing old methods that seem to produce contentious results that cloud understanding (B. Hoffman, et al., 2009; Lance, 2008).

In addition, with increasing globalization, it is very likely that doing research into cross-cultural applications of ACs will become increasingly important (Thornton, et al., 2006). As countries begin to interact more and more, we need to think about how ACs will work with communalistic cultures as opposed to individualistic cultures (Hampden-Turner & Trompenaars, 2000). Specifically, we need to explore if the same dimensions will work across broad cultural differences, or do different cultures think about job performance in ways that we have not considered with ACs.

References

Abraham, J. D., Morrison, J. D., Jr., & Burnett, D. D. (2006). Feedback seeking among developmental assessment center participants. *Journal of Business and Psychology, 20*(3), 383-394.

Anderson, L. R., & Thacker, J. (1985). Self-monitoring and sex as related to assessment center ratings and job performance. *Basic and Applied Social Psychology, 6*(4), 345-361.

Anderson, N., Payne, T., Ferguson, E., & Smith, T. (1994). Assessor decision making information processing and assessor. *Personnel Review, 23*(1), 2.

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*(1), 125-154.

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*(4), 813-835.

Bandura, A. (1985). *Social Foundations of Thought and Action: A Social Cognitive Theory*: Prentice-Hall.

Bandura, A., & Locke, E. A. (2003). Negative Self-Efficacy and Goal Effects Revisited. *Journal of Applied Psychology, 88*, 87-99.

Bartels, L. K., & Doverspike, D. (1997). Assessing the assessor: The relationship of assessor personality to leniency in assessment center ratings. *Journal of Social Behavior & Personality, 12*(5), 179-190.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3), 478-494.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*(5), 1114-1124.

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*(6), 957-963.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General & Applied, 80*(17), 27.

Brownell, J. (2005). Predicting leadership: The assessment center's extended role. *International Journal of Contemporary Hospitality Management, 17*(1), 7.

Chiaburu, D. S., & Marinova, S. V. (2005). What predicts skill transfer? An exploratory study of goal orientation, training self-efficacy and organizational supports. *International Journal of Training & Development, 9*, 110-123.

Cochran, D. S., Hinckle, T. W., & Dusenberry, D. (1987). Designing a developmental assessment center in a government agency: A case study. *Public Personnel Management, 16*(2), 145-152.

Cohen, A. R., & Bradford, D. L. (2007). Influence without authority: The use of alliances, reciprocity, and exchange to accomplish work. In J. Osland, M. Turner, D. Kolb & I. Rubin (Eds.), *The Organizational Behavior Reader* (8th ed., pp. 569-579). Upper Saddle River, NJ: Pearson Prentice Hall.

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(1), 121-124.

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*(3), 685-691.

Donaldson, R. (Writer) (2003). The Recruit. In J. Glickman & R. Kidney (Producer). USA: Touchstone Pictures.

Dreher, G. F., & Sackett, P. R. (1981). Some problems with applying content validity evidence to assessment center procedures. *Academy of Management Review, 6*(4), 551-560.

Fletcher, C. (1991). Candidates' reactions to assessment centres and their outcomes: A longitudinal study. *Journal of Occupational Psychology, 64*(2), 117-127.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*(3), 493-511.

Gaugler, B. B., & Rudolph, A. S. (1992). The influence of assessee performance variation on assessors' judgments. *Personnel Psychology, 45*(1), 77-98.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*(4), 611-618.

Gibbons, A. M., Rupp, D. E., Snyder, L. A., Holub, A., & Woo, S. E. (2006). A preliminary investigation of developable dimensions. *Psychologist-Manager Journal, 9*(2), 99-123.

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*(2), 357-374.

Goodge, P. (1994). Development centres: Design generation and effectiveness. *The Journal of Management Development, 13*(4), 16.

Goodstone, M. S., & Lopez, F. E. (2001). The frame of reference approach as a solution to an assessment center dilemma. *Consulting Psychology Journal: Practice and Research, 53*(2), 96-107.

Hampden-Turner, C., & Trompenaars, F. (2000). *Building Cross-Culture Competence*. London: Yale University Press.

Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment centre observation procedures: An experimental comparison of traditional, checklist and coding methods. *International Journal of Selection and Assessment, 6*(4), 222-231.

Herriot, P., Chalmers, C., & Wingrove, J. (1985). Group decision making in an assessment centre. *Journal of Occupational Psychology, 58*(4), 309-312.

Hoffman, B., Melchers, K., Blair, C., Kleinmann, M., & Ladd, R. (2009). *Exercises and dimensions are the currency of assessment centers*. Paper presented at the 24th Annual SIOP Conference.

Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*(2), 455-470.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21-super(st) century. *Journal of Social Behavior & Personality, 12*(5), 13-52.

Jackson, D. J., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18*(3), 213-241.

Jex, S., & Britt, T. (2008). *Organizational Psychology: A Scientist-Practitioner Approach* (2nd ed.). Hoboken, NJ: John Wiley and Sons, Inc.

Joiner, D. A. (2000). Guidelines and ethical considerations for assessment center operations: International task force on assessment center guidelines. *Public Personnel Management, 29*(3), 315.

Jones, A. (1981). Inter-rater reliability in the assessment of group exercises at a UK assessment centre. *Journal of Occupational Psychology, 54*(2), 79-86.

Jones, R., & Whitmore, M. (1995). Evaluating developmental assessment centers as interventions. *Personnel Psychology, 48*(2), 377-388.

Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology, 47*(1), 109-121.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*(1), 59-69.

Kleinmann, M., & Koller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior & Personality, 12*(5), 65-84.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*(2), 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*(4), 325-338.

Kolk, N. J., Born, M. P., & van der Flier, H. (2004). A triadic approach to the construct validity of the assessment center: The effect of categorizing dimensions into a feeling, thinking, and power taxonomy. *European Journal of Psychological Assessment, 20*(3), 149-156.

Kottke, J. L., & Shultz, K. S. (1997). Using an assessment center as a developmental tool for graduate students: A demonstration. *Journal of Social Behavior & Personality, 12*(5), 289-302.

Lance, C. E. (2008). Where have we been, how did we get there, and where shall we go? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(1), 140-146.

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*(2), 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*(4), 323-353.

Landy, F. J. (1986). Stamp Collecting Versus Science: Validation as Hypothesis Testing. *American Psychologist, 41*(11), 1183-1192.

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264.

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*(3), 203-221.

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*(4), 675-686.

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

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*(6), 1202-1222.

Lievens, F., & Goemaere, H. (1999). A different look at assessment centers: Views of assessment center users. *International Journal of Selection and Assessment, 7*(4), 215-219.

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*(4), 715-738.

Melchers, K. G., & Konig, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(1), 125-127.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*(5), 1042-1052.

Meyer, H. H. (1970). The validity of the In-Basket Test as a measure of managerial performance. *Personnel Psychology, 23*(3), 297-307.

Moser, K., Schuler, H., & Funke, U. (1999). The moderating effect of raters' opportunities to observe ratees' job performance on the validity of an assessment centre. *International Journal of Selection and Assessment, 7*(3), 133-141.

Moses, J. (2008). Assessment centers work, but for different reasons. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(1), 134-136.

Nadler, D. A., & Lawler, E. E. (2007). Motivation: A diagnostic approach. In J. Osland, M. Turner, D. Kolb & I. Rubin (Eds.), *The Organizational Behavior Reader* (8th ed., pp. 171-180). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Norton, S. D. (1977). The empirical and content validity of assessment centers vs. traditional methods for predicting managerial success. *Academy of Management Review, 2*(3), 442-453.

Norton, S. D. (1981). The assessment center process and content validity: A reply to Dreher and Sackett. *Academy of Management Review, 6*(4), 561-566.

Pynes, J., & Bernardin, H. (1992). Mechanical vs consensus-derived assessment center ratings: A comparison of job performance validities. *Public Personnel Management, 21*(1), 17-28.

Rousseau, D. M. (2007). The psychological contract: Violations and modifications. In J. Osland, M. Turner, D. Kolb & I. Rubin (Eds.), *The Organizational Behavior Reader* (8th ed., pp. 41-48). Upper Saddle River, New Jersey: Pearson Prentice Hall.

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. *The Psychologist-Manager Journal, 9*(2), 171 - 200.

Rupp, D. E., Thornton, G. C., & 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*(1), 116-120.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*(1), 13-25.

Sackett, P. R., & Dreher, G. F. (1981). Some misconceptions about content-oriented validation: A rejoinder to Norton. *Academy of Management Review, 6*(4), 567-568.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*(4), 401-410.

Schippmann, J. S., Hughes, G. L., & Prien, E. P. (1987). The use of structured multi-domain job analysis for the construction of assessment center methods and procedures. *Journal of Business and Psychology, 1*(4), 353-366.

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*(4), 735-746.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology, 62*(2), 171-176.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*(3), 407-422.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*(1), 32-41.

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*(1), 101-116.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93*(2), 280-295.

Smith-Jentsch, K., Jentsch, F., Payne, S., & Salas, E. (1996). Can pretraining experiences explain individual differences in learning? *Journal of Applied Psychology, 81*(1), 110-116.

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*(1), 71-90.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74-83.

Thornton, G., & Rupp, D. (2006). *Assessment centers in human resources management: Strategies for prediction, diagnosis, and development*. New York: Psychology Press.

Vancouver, J. B., & Kendall, L. N. (Writer) (2006). When Self-Efficacy Negatively Relates to Motivation and Performance in a Learning Context [Article], *Journal of Applied Psychology*.

Velada, R., Caetano, A., Michel, J., Lyons, B., & Kavanagh, M. (2007). The effects of training design, individual characteristics and work environment on transfer of training. *International Journal of Training and Development, 11*(4), 282-294.

Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology, 48*(2), 347-375.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*(2), 231-258.