CHAPTER 4

Adjusting Exercise Design in Assessment Centers: Theory, Practice, and Research

FILIP LIEVENS AND EVELINE SCHOLLAERT

Upon looking at this title, an obvious question is whether assessment center (AC) exercises need to be designed differently from what was done in the past. AC exercises have been around since World War I and their longevity attests to their success. So why propose some changes to this monument in personnel selection and development?

Let us first acknowledge that we concur that ACs are in still in good shape. However, every good brand needs some adjustments (which is a more appropriate term than "changes" in this context) once in a while, based on new theoretical insights and empirical research. Therefore we build our novel exercise design approach on recent insights in person-situation interactionism to make a good tool even better.

The structure of this chapter is as follows: we start with delineating the reasons behind our revised exercise design approach in ACs. Next we explain the theory behind the revised exercise design approach. In a third section, we report on our program of research related to this new approach. The fourth part discusses some other possible applications in the AC domain that are congruent with this revised exercise design approach.

Why Adjust AC Exercise Design?

Generally, ACs show a record of success: first, meta-analytic research confirmed that AC ratings were predictive of a variety of criterion measures. The meta-analysis of AC criterion-related validity studies at the dimension level (Arthur et al. 2003) contained 258 validity coefficients from 34 studies. In each study, the criterion was job-related (for example, job performance ratings, promotion, and salary). On the basis of a set of 6 dimensions Arthur et al. found a range of estimated true criterion-related validities from .25 to .39, indicating the predictive power of ACs. Second, applicants react positively to ACs. The meta-analysis of Hausknecht, Day and Thomas (2004) showed that behavior sample-based selection procedures were perceived more favorably than, for instance, cognitive ability tests and personality inventories. In particular applicants view ACs as more face valid than most

other selection procedures. Third, a recent meta-analysis demonstrated that different assessors tend to agree when they are asked to evaluate candidates (Connelly and Ones 2008). Fourth, the majority of AC studies on adverse impact attest to the widely held view that ACs are reasonably unbiased regarding race and gender. According to the meta-analysis of Dean, Roth and Bobko (2008), positive results were reported, in particular for females and Hispanics. For Blacks, however, their results suggested that ACs may be associated with more adverse impact than was previously thought in the literature, but still have less adverse impact than the typical cognitive ability test.

Besides these benefits leading to the popularity of ACs, one of the main advantages of ACs is that assessors have the opportunity to observe actual behavior in a simulated work setting. This key focus on behavior in ACs is also well reflected in the most recent *Guidelines and Ethical Considerations of Assessment Center Operations*, in which the observation of overt behavioral responses is described as a necessary and fundamental component of ACs (International Task Force on Assessment Center Guidelines 2009). The *Guidelines* further state that AC designers should attempt to design exercises that evoke a large number of job-related behaviors because this should give assessors enough opportunities to observe job-related behavior. Generating more job-related behaviors in an exercise is quintessential for developmental ACs, because these behaviors serve as a basis for providing participants with detailed developmental feedback about their strengths and weaknesses. The more behavioral examples one can assemble, the more convincing the feedback will likely be. In addition, regardless of how that behavior is then captured and evaluated by assessors (for example, in exercise/task-based models, dimension-based models), eliciting and observing behavior is key to effective assessment and development centers.

Although the AC *Guidelines* and the literature emphasize the importance of exercises providing sufficient opportunities for observing job-related behavior, various authors have emphasized that AC exercises score not that well in terms of observability of behavior:

- First, Bycio, Alvares, and Hahn (1987: 472) noted that "assessors within an exercise are sometimes, if not usually, forced to base *all* of their judgments on *four* or *five* behaviors."
- Next, Brannick, Michaels, and Baker (1989) mentioned that assessors often need to rely on *one* particular behavioral reaction ("red hot" item) to score candidates on *several* dimensions.
- Third, Reilly, Henry and Smither (1990) pointed out that assessors sometimes have too few observations on which to base their ratings for some dimensions, when not enough behaviors are evoked.
- Furthermore, Kudisch, Ladd and Dobbins (1997) suggested that consistency of AC ratings across exercises may be enhanced when dimensions are easier to observe. For example, the dimension Communication (which is overt in most exercises) produced more convergent validity, as opposed to Problem Analysis (which is less observable in most exercises).
- Although the meta-analysis of Connelly and Ones (2008) revealed good inter-rater reliability coefficients for AC ratings, reliability was lowest for so-called within-exercise dimension ratings (ratings made on one dimension within a specific exercise). The main argument is that such ratings are often based on rather limited behavioral evidence.

In sum, it seems that it is not always guaranteed that AC exercises enable assessors to collect enough behavioral observations per dimension (see Brannick 2008, Howard 2008, Lievens 2008, Lievens, Tett and Schleicher 2009). An explanation can be found in the traditional AC paradigm which focuses on the exercise as a whole. In this holistic approach, the exercise as a whole is seen as a vehicle for evoking behavior (Howard 2008, Lievens et al. 2009, McFarland et al. 2005). However, as noted above, research has shown that this holistic exercise approach might occasionally be problematic in that an insufficient number of behaviors are elicited. That is the reason why we argue for a more molecular approach by planting situational stimuli within exercises to enhance the observability of dimensions across a variety of dimensions in ACs. Upfront we acknowledge that practitioners might already have intuitively attempted to elicit dimension relevant behavior by building context, content, personnel, problems into exercises or by instructing role players to use prompts in response to paths chosen by participants. However, our point is that theory and research has not shown systematically how the behavioral elicitation process works and has not provided empirical evidence of its effectiveness

Theoretical Background

It is generally acknowledged that behavior of candidates in ACs is determined neither solely by dispositional factors (stable personal characteristics of candidates) nor solely by situational factors (AC exercises) but by the interaction of the person and the situation. Therefore, it is relevant to conceptualize the occurrence of candidate behavior in ACs in terms of a recent interactionist theory such as trait activation theory (Lievens et al. 2009; Tett and Burnett 2003). Trait activation theory focuses on the person-situation interaction to explain behavior based on responses to trait-relevant cues found in situations (Tett and Guterman 2000). These observable responses serve as the basis for behavioral ratings in a variety of assessments such as ACs (Tett and Burnett 2003).

According to trait activation theory, two factors are important to understand in which situations a trait is likely to manifest itself in behavior. First, trait activation theory emphasizes the importance of *situation trait relevance*. A situation is considered relevant to a trait if it provides cues for the expression of trait relevant behavior (Tett and Guterman 2000). Thus, situation trait relevance is a qualitative feature of situations that is essentially trait specific; it is informative with regard to which cues are present to elicit behavior for a given latent trait. Relatedly, trait activation theory states that situations should provide ample opportunities for behavior to be expressed. This idea builds on the well-known principle of aggregation (Epstein 1979) in social psychology, which states that the sum of a set of measurements is more stable than any single measurement from the set. For example, when someone is having an accident and is confronted with an angry driver, this situation provides cues for traits such as Emotional Stability. Conversely, this situation is less relevant to evoke traits such as Imagination (Openness).

Situation strength is the second relevant factor from the trait activation perspective. Situation strength is more of a continuum that refers to how much clarity there is with regard to how the situation is perceived. Strong situations involve unambiguous behavioral demands and are therefore likely to negate almost all individual differences in behavior

without regard to any specific trait. Conversely, weak situations are characterized by more ambiguous expectations, enabling much more variability in behavioral responses to be observed (Meyer, Dalal and Hermida 2010). For instance, at the end of a busy day a shop assistant may be confronted with a messy shop full of odds and ends left by the customers. When the supervisor instructs the shop assistant to clean the mess in the shop, it will be much more difficult to observe individual differences related to the trait order, whereas the opposite might be true in the absence of such clear-cut supervisory instructions (without instructions some shop assistants will immediately start to clean the shop, others will not notice the mess).

Thus, trait activation theory has key implications for AC exercises (see Lievens et al. 2009). That is, the application of trait activation theory involves recognition of the importance of building multiple stimuli into the AC exercises. Accordingly, exercises can be explicitly designed to increase their situation trait relevance. In this respect, Brannick (2008: 132) cogently argued to "deliberately introduce multiple dimension-relevant items or problems within the exercise and to score such items." Apart from increasing the situation trait relevance of AC exercises, trait activation theory also suggests taking situation strength into account when planting stimuli in AC exercises. That is, behavior elicitation should avoid presenting the candidate with a too strong situation (in terms of behavioral demands). For example, role players might create a relatively weak situation by showing for a moment a sad facial expression (prompt to evoke Interpersonal Sensitivity). Some candidates will ask what is bothering the role player, whereas other candidates will ignore the expression or even will not notice the expression, leading to variability in candidate reactions. On the other hand, to evoke Interpersonal Sensitivity the role player might also start to sob. Almost every candidate will notice this and will react on it. This is probably too strong of a situation so that variability in candidate reactions will be masked.

Examples of Situational Stimuli

In the previous section, we outlined some general principles for eliciting candidate behavior in ACs. In this section, we provide five different examples of how to put this general logic into practice.

The first approach entails adapting the content of the exercise. Let us take an oral presentation with challenging questions as an example. Examples of stimuli to elicit behavior relevant to a dimension such as Resistance to Stress (a facet of the broader trait of Emotional Stability) might be the inclusion of a stringent time limit, sudden obstacles, or information overload. In a more systematic way, AC designers might ensure that several content cues are embedded at the task, social, and organizational levels within a given exercise (if job-related, of course).

A second way to elicit job-related behavior is through exercise instructions. In ACs, exercise instructions provide information and set expectations for candidates about what behavior to show or not to show. For example, exercise instructions might be vague (for example, "solve the problem") or more concrete (for example, "motivate the problem subordinate"). Similarly, exercise instructions might be unidimensional (for example, reach consensus) or multidimensional (for example, reach consensus and make the company more profitable). To date, we know little about how such exercise instruction

variations might affect the behavior demonstrated, in terms of either direct effects or in interactions with underlying personality traits.

Thirdly, when interpersonal exercises are used, role-player cues are an additional means of eliciting job-related behavior. In current AC practice, role players are typically given a specific list of things to do and to avoid. Role players are also trained to perform realistically and consistently across candidates. Although these best practices have proven their usefulness over the years, a key function of trained role players consists of evoking behavior from candidates (Thornton and Mueller-Hanson 2004). Trait activation theory can help identify which specific behaviors might be evoked by specific role-player stimuli (prompts). Prompts are defined as predetermined statements that a role player consistently mentions in an AC across candidates to elicit behaviors related to specific job-related dimensions. For example, to arouse behavior related to Interpersonal Sensitivity, the role player might state that he feels bad about a candidate's decision. Similarly, role players might trigger behavior related to Planning and Organizing (deeper trait of Conscientiousness) by asking how the candidate will implement his or her solution.

It is important that these role-player cues should *subtly* elicit assessee behavior because the situations might otherwise become too strong. Indeed, role-player prompts might vary from being very explicit (strong) to being very implicit (weak) in eliciting the dimensions targeted. We illustrate this notion of situation strength in role-player prompts again with the dimension of Interpersonal Sensitivity. That is, role players might react to a decision made by the candidate by showing momentarily a distressed expression on their face (weak situation) or might start to sob (extremely strong situation).

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

Finally, stimuli could also be presented via videotape, PC, or even virtual reality. In the videotape approach, resembling earlier social intelligence measures (Stricker and Rock 1990), candidates are shown short scenes and asked to react to what they saw. Recent applications even enable creation of avatar-based simulation exercises wherein participants take on a virtual identity and are confronted with standardized stimuli in a virtual workplace (Rupp, Gibbons and Snyder 2008).

Overview of Empirical Research

Let us start by acknowledging that some earlier studies have already scrutinized the effects of specific characteristics of AC exercises (Highhouse and Harris 1993, Schneider and Schmitt 1992). For instance, Schneider and Schmitt (1992) experimentally manipulated the effects of exercise content (competitive vs. cooperative demands) and exercise form (for example, role play vs. group discussion) on candidate ratings. The form of the exercise emerged as the most important exercise factor in leading candidates to perform differently across exercises.

So, the focus of this limited number of prior studies on AC exercise characteristics has typically been the *exercise as a whole*. Specific situational stimuli *within exercises* were not investigated. To fill this gap in empirical research and to put our aforementioned theory to the test we set up a research program. So far, the following questions have been addressed:

1. Is it possible to build situational stimuli in AC exercises?
2. What are the effects of building situational stimuli in AC exercises on observability?
3. What are the effects of building situational stimuli in AC exercises on inter-rater reliability?
4. What are the effects of building situational stimuli in AC exercises on construct-related validity?
5. What are the effects of building situational stimuli in AC exercises on applicant reactions?

The remainder of this section provides an overview of the available empirical research evidence. Note that most of the studies are still ongoing and that only preliminary research evidence is presented.

IS IT POSSIBLE TO BUILD SITUATIONAL STIMULI IN AC EXERCISES?

Some situational stimuli (for example, exercise instructions, video-based stimuli or stimuli in virtual reality) can be used and implemented independently of the candidate. In that case, it is relatively straightforward that the answer to this question is "Yes." However, other situational stimuli (role-player prompts) are given in a constantly changing situation. In that case, the answer to this question might be more complex. Prompts create a situational stimulus for evoking job-relevant behavior. Role players are taught to use multiple standardized prompts for each dimension in a consistent fashion across candidates. These prompts provide a framework for responding as every conversation is different. By using prompts a situational stimulus for evoking job-relevant behavior is created. In role-player training, role players are then taught to use multiple standardized prompts per dimension in a consistent fashion across candidates. These prompts are framed in a script as every conversation is different. On the one hand role players need to follow this script as strictly as possible, thereby being expected to use enough prompts per dimension. Yet, on the other hand they also need to play their role in a credible way. Each candidate reacts differently so that the role player often has to pursue several dissimilar strategies to maintain the script. These opposing demands might put some pressure on role players. Consequently, the question is whether role players are actually able to use prompts in a standardized way despite those opposing demands.

Schollaert and Lievens (in press) sought to examine this. Their focal question was whether role players were able to use prompts. A sample of role players was randomly assigned to one of the following two conditions: role-player training without prompts and role-player training with prompts. Generally, the results indicated that attending training with prompts substantially increased the number of prompts used by role players during the assessment exercises. Effect sizes were large. In the role-play exercise, the average proportion of prompts increased fivefold. In the presentation exercise, the proportion of prompts quadrupled. Thus, these results support the view that role players

might serve as a practical means of structuring AC exercises to consistently evoke job-relevant behavior. Despite this positive evidence, using prompts is not straightforward as results also showed that after attending role-player training with prompts, half of the interactions still failed to show any prompts.

WHAT ARE THE EFFECTS OF BUILDING SITUATIONAL STIMULI INTO AC EXERCISES ON OBSERVABILITY?

As noted above, the AC *Guidelines* have underscored the importance of the observability of behavior. Hence, an important argument for using situational stimuli is based on the key assumption that they should increase the number of observations per dimension noted down by assessors.

Schollaert and Lievens (2009) tested this assumption of the increased situational relevance of AC exercises by contrasting two vehicles for increasing behavior observability, namely instructions to role players prior to a role play and role-player prompts during the role play. No main effect was found for exercise instructions. Apparently, providing specific exercise instructions did not influence the number of good observations. However, results showed a main effect for prompt-training, with the use of role-player prompts leading to greater observability of dimension-relevant behavior. Thus, the inclusion of situational stimuli and especially the use of prompts might serve as a practical vehicle to avoid that assessors need to rely on too few behavioral reactions to score candidates (Brannick et al. 1989).

WHAT ARE THE EFFECTS OF BUILDING SITUATIONAL STIMULI IN AC EXERCISES ON INTER-RATER RELIABILITY?

If situational stimuli are built into AC exercises, one might expect that this has beneficial effects on inter-rater reliability for several reasons. First, evoking more candidate behavior should increase the standardization of those exercises. Second, the opportunity to observe and take notes on dimension-related behavior should also increase the reliability of the ratings made in light of the aforementioned principle of aggregation (Epstein 1979). Just as the reliability and content representation of a test increases with the addition of items from the same domain, assessing a given dimension in an AC exercise might improve with the addition of dimension-specific cues. Third, the use of standardized situational cues in ACs can be compared to the use of standardized questions among interviewers. Research in the interview domain has shown that the inter-rater reliability of structured interviews is higher than that of unstructured interviews (Conway, Jako and Goodman 1995). Thus the use of standardized dimension-related stimuli across candidates might increase the standardization, the structure, and the consistency of AC ratings.

In our research, we found empirical support for these hypotheses only when assessors were also familiar with the situational stimuli built into the AC exercises. For example, Lievens, Keen, and Schollaert (2010) compared three conditions. In the low behavior elicitation condition, no formal attempts were implemented to evoke dimension-related behavior. In the medium behavior elicitation condition, role players were trained to use specific prompts for evoking candidate behavior. The high behavior elicitation condition was similarly designed as the medium behavior elicitation condition, with the addition that assessors were also familiarized with the prompts for eliciting behavior. In that

condition, assessors knew which prompts were related to which dimensions. Inter-rater reliability was highest in the third condition, where role players used prompts for evoking behavior and where assessors were also familiar with the prompts used by role players.

Taken together, these results make sense as they indicate that it is not enough that exercises generate more behavior. In addition, it seems important that assessors receive information about the cues that elicit behavior in order to consistently observe, classify, and rate that behavior.

WHAT ARE THE EFFECTS OF BUILDING SITUATIONAL STIMULI IN AC EXERCISES ON CONSTRUCT-RELATED VALIDITY?

When situational stimuli are used to elicit a higher number of behaviors in AC exercises, it can be expected that dimensions are also better measured in AC exercises. Thus, one might anticipate beneficial effects on the construct-related validity of AC exercises. Given the higher number of behaviors available per dimension, assessors should be less prone to using a couple of behavioral items per dimension to rate these dimensions. In other words, their ratings should be less susceptible to halo. The convergence of their ratings across exercises might also increase due to the higher observability of the dimensions. This might be an important advantage of the use of situational stimuli as the construct-related validity of AC ratings has traditionally been identified as one of the weaker points of AC technology, especially when within-exercise dimension ratings are used (Bowler and Woehr 2006; Lance 2008; Lievens and Conway 2001).

In recent years, some studies have put these expectations to the test. Schollaert and Lievens (in press) found that Problem-solving and Interpersonal Sensitivity dimensions were better measured in AC exercises when role players used prompts for evoking these dimensions. This was evidenced by higher correlations between Problem-solving and Interpersonal Sensitivity ratings and Cognitive Ability and Agreeableness, respectively. Lievens et al. (2010) focused on behavior elicitation via role-player prompts in a sample of actual candidates for a managerial job. As noted above, they distinguished between three levels of behavior elicitation (high, medium and low). Results showed that construct-related validity (convergent and discriminant correlations) was highest in the high behavior elicitation condition. That is, significantly more evidence for convergent and discriminant validity was established when role players used prompts for eliciting behavior and when assessors were familiar with these prompts.

Two other studies experimented with the use of video-based vignettes for eliciting dimension-related behavior. In Lievens (2009), candidates for police-officer jobs watched video-based scenes. Each of these scenes triggered a specific dimension. At the end of each scene, the character in the video spoke directly into the camera. Candidates were next required to answer the character directly, with their verbal and non-verbal reply being captured by a webcam. These reactions were then coded by trained assessors. One set of analyses examined the consistency of assessors' dimensional ratings across scenes (convergent validity). That is, did scenes that triggered a similar dimension provide a consistent measurement of that specific dimension? In line with expectations, the consistency in assessor ratings was acceptable (only a more ambiguous dimension such as Integrity scored a bit lower), confirming that the use of multiple videotaped scenes for measuring one dimension might serve as a good vehicle for obtaining a more consistent measurement of the targeted dimension. Brink et al. (2008) also showed candidates short

video scenes and asked them to react to what they saw. They focused on the discriminant validity of assessor ratings and found that assessors were able to make good differentiations among the various dimensions.

WHAT ARE THE EFFECTS OF BUILDING SITUATIONAL STIMULI IN AC EXERCISES ON APPLICANT REACTIONS?

Traditionally, AC exercises are held in high regard by participants. However, the use of situational stimuli might also have an impact on how participants perceive AC exercises. On the one hand, the use of situational stimuli might lead to more favorable applicant perceptions because candidates might appreciate that prompts aim to elicit job-related behavior. That is, the use of stimuli that evoke relevant behavior might be perceived as increasing the overlap with behavior on the job. As candidates prefer job-related selection procedures (Hausknecht et al. 2004) the inclusion of situational stimuli might lead to higher perceptions of job-relatedness.

On the other hand, the use of situational stimuli might also have some negative effects on applicant perceptions. Possibly it reduces the realism and interpersonal warmth of AC exercises because it might detract from the natural flow of the exercise. In fact, prior research in the interview domain has shown that the use of structure in interviews led to less favorable candidate perceptions as compared to unstructured interviews (Conway and Peneno 1999).

So far, only one study has examined the impact of the use of situational stimuli (in this case role-player prompts) on applicant perceptions. Schollaert and Lievens (in press) examined the effects of situational stimuli in the form of role-player prompts on perceptions of job-relatedness, two-way communication, and interpersonal treatment. They hypothesized that the use of prompts led to higher perceptions of job-relatedness and to decreases in the perceptions of two-way communication and interpersonal treatment. Half of the candidates were confronted with role players not using prompts and half of the candidates were confronted with role players using prompts. For interpersonal treatment and job-relatedness, no significant effects were found. This could be due to a ceiling effect as previous research showed that candidates already react highly favorable to AC exercises. For two-way communication, a significant effect was reported. Contrary to expectations, candidates had the perception of having more opportunities to give their opinion in the condition with prompts. Candidates might have considered prompts as providing them with more opportunity to converse with the role player. In any case, these results suggest that the use of prompts does not have a negative influence on candidate reactions. However, a caveat is warranted as this study was not conducted in a real selection setting and the sample consisted of final year university students without AC experience. So, future research with experienced applicants is needed. In addition, the effects of the use of situational cues on participants' acceptance of AC feedback should be scrutinized.

SUMMARY AND FUTURE RESEARCH

So far, researchers have experimented with three types of situational stimuli: exercise instructions, role-player prompts, and videotaped stimuli. Generally, this overview of the growing empirical evidence shows that it is possible to build situational stimuli into AC exercises. Incorporating multiple situational stimuli is also found to generate the

anticipated effects in terms of increasing the number of behaviors to be observed. That is, assessors noted down more behaviors on their observation forms in the case where situational stimuli were built into the exercises. However, situational stimuli do not guarantee that assessors are taking them into account in their ratings. Only if assessors are familiarized with the situational cues they have effects on inter-rater reliability, convergent validity, and discriminant validity. Finally, applicants do not seem to notice any negative effects of the use of situational cues in AC exercises.

Future research should extend these results. A key issue consists of investigating whether the use of situational cues also increases the criterion-related validity of the AC exercise; the rationale being that as more dimension-relevant behavior will be elicited, the use of situational stimuli might increase the overlap with the criterion.

Another issue is whether the provision of cues changes the dimension that is actually being measured. That is, does it change from a measure of maximum performance to typical performance, or vice versa? The provision of cues in AC exercises may make the situation stronger, and suggest to participants what they should do, rather than allow them to choose what to do. Interestingly, the effects might differ depending on the type of dimensions (personality like dimensions versus ability like dimensions). For example, consider a role play where the candidate is a manager and the role player is a supervisor having a problem with an employee. With no cues, the candidate may or may not engage in coaching behaviors; with cues from the role player (for example, "Well, what can you do to help me with my problem?"), the candidate may start giving suggestions. The exercise then provides behavior relevant to coaching (maximum performance or "can do"), but at the cost of denying the candidate the opportunity of proactively displaying any inclination to provide coaching to the subordinate (typical performance or "will do"). So, giving cues might change the dimensions being measured from a measure of tendency to coach to a measure of ability to coach (see also McDaniel, Hartman, Whetzel and Grubb, 2007).

The same sort of analysis should be applied to dimensions like problem-solving. With minimal cues, the candidate may do very little systematic problem analysis and decision analysis. But the role player might give a series of cues in follow-up questions after the presentation: "What led you to say you would do … ?" "What other solutions did you consider … ?" The situation then may become stronger, and the candidate would give answers that he did not really consider in his initial preparation. Thus, one could evaluate the ability to do problem-solving because more of those behaviors would be displayed. In summary, to measure "personality-like" traits, providing cues might make the situation stronger and reduce individual differences in the behavior one wants to observe. To measure "cognitive ability – like" traits, providing cues may ensure that relevant behaviors are displayed, and thus enhance measurement accuracy. In future studies, we plan to test these ideas.

Other Implications

So far, we have focused on the use of situational cues in AC exercises. However, the inclusion of situational cues not only has implications for AC exercises, but also for at least the following three other components of AC technology, namely the design of behavioral checklists, assessor training, and alternate AC exercises.

BEHAVIORAL CHECKLISTS

As noted above, initial research with the inclusion of predetermined situational cues in AC exercises suggests that it is important that assessors are familiar with the cues used. To accomplish this, a practical approach might consist of including the situational cues that were designed to elicit candidate behavior (for example, the role-player statements) in the behavioral checklists provided to the assessors. In an even more structured format, these cues could be presented in their anticipated chronological order, with the candidate behaviors to be observed arranged around them. Accordingly, assessors are reminded and prompted by the situational cues when attending to candidate behavior. It might help them to "see the forest for the trees" in the complex stimuli triggered by AC exercises. Brannick (2008) refers to this approach as aligning the stimulus content of the exercises with the scoring rubric (as is sometimes done in in-baskets, wherein they are provided behaviorally anchored rating scales which show, for each in-basket item, the types of responses that would be considered high/medium/low performance).

ASSESSOR TRAINING

A related implication consists of familiarizing assessors with the situational cues in assessor training. In current assessor training practice, the focus is placed on imposing a consistent frame-of-reference on assessors (Lievens 2001). In such training programs, the dimensions and the accompanying behaviors play logically a crucial role. However, it is equally important that assessors know when specific behavior is potentially being activated by various situational stimuli. We are not aware of studies that have examined such a more comprehensive assessor training approach. Apart from teaching assessors the cues in a lecture, other options are possible. For example, when the same individuals serve as assessors and role players in the AC, they also learn to use the cues. This might be especially helpful in cross cultural settings where candidates and assessors come from different backgrounds.

ALTERNATE AC EXERCISES

One potential benefit of incorporating situational cues in AC exercises is that the AC exercise is no longer a "black box"; that is, the development and inclusion of situational cues within an AC exercise breaks it down into different parts and components. Thus, the identification of situational cues might guide the determination of the deeper structural aspects (the so-called radicals, to use a term from item generation theory; Irvine, Dann and Anderson 1990; Lievens and Sackett 2007) of an AC exercise, in terms of providing a template of what aspects of the exercise map onto which dimensions, and should be kept constant across exercises.

Such a more molecular approach to AC exercise design might make it easier to develop alternate forms of AC exercises (Brummel, Rupp and Spain 2009). For example, one might develop several role-player cues to evoke behavior related to the dimension of Interpersonal Sensitivity in a series of role plays. Superficial differences among the cues would be incidental to their deeper similarities (as radicals) in targeting the same dimension. The same might be done for other dimensions. Thus, more generally, to

construct alternate forms of AC exercises, we suggest changing the surface features of an AC exercise, while keeping the deep structure of the exercise intact.

Epilogue

In this paper, we proposed to build multiple situational stimuli in AC exercises. We also reported on various studies that have implemented this approach. We want to emphasize that we do not suggest that current best practices of exercise development (and AC design in general) should be abandoned. Rather, we argue that our approach should also play a more prominent role in such development, with the goal of making a good tool even better. Whereas current practices typically simulate key task, social, and organizational demands of the job, we see untapped potential in planting multiple stimuli within exercises as a systematic and structured means of increasing the frequency and variability of job-related behavior in AC exercises.

References

Arthur, W. Jr., Day, E.A., McNelly, T.L. and Edens, P.S. 2003. A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology,* 56, 125–154.

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

Brannick, M.T. 2008. Back to basics of test construction and scoring. *Industrial and Organizational Psychology,* 1, 131–133.

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

Brink, K.E., Lance, C.E., Bellenger, B.L., Morrison, M.A., Scharlau, E.A. and Crenshaw, J.L. 2008, April. Discriminant validity of a "next generation" assessment center. In: B.J. Hoffman (Chair), *Reexamining assessment centers: Alternate approaches*. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Brummel, B.J., Rupp, D.E. and Spain, S.M. 2009. Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology*, 62, 137–170.

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

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

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

Conway, J.M. and Peneno, G.M. 1999. Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology,* 13, 485–506.

Dean, M.A., Roth, P.L. and Bobko, P. 2008. Ethnic and gender subgroup differences in assessment center ratings: a meta-analysis. *Journal of Applied Psychology,* 93, 685–691.

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

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

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

Howard, A. 2008. Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology,* 1, 98–104.

International Taskforce on Assessment Centers Guidelines. 2009. Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment,* 17, 243–253.

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

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

Lance, C.E. 2008. Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology,* 1, 84–97.

Lievens, F. 2009, February. Effects of response fidelity on test performance and validity. Paper presented at the 23rd Meeting of the Personnel and Human Resources Research Group, College Station, TX.

Lievens, F. 2008. What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice,* 1, 117–120.

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

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

Lievens, F., Keen, G. and Schollaert, E. 2010, April. *A novel look at behavior elicitation in assessment center exercises*. Poster session presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, United States.

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

Lievens, F., Tett, R.P. and Schleicher, D.J. 2009. Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In *Research in Personnel and Human Resources Management,* edited by J.J. Martocchio and H. Liao. Bingley: JAI Press, 99–152.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L. and Grubb, W.L. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology,* 60, 63–91.

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

Meyer R.D., Dalal, R.S. and Hermida, R. 2010. A review and synthesis of situational strength in the organizational sciences. *Journal of Management,* 36, 121–140.

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

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

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

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

Schollaert, E. and Lievens, F. in press. The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment.*

Schollaert, E. and Lievens, F. 2009, October. *Using trait activation theory to increase the amount of candidate behaviors elicited in assessment center exercises?* Paper session presented at the 4th Dutch-Flemish Research Meeting on Personnel Recruitment and Selection, Amsterdam, The Netherlands.

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

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

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

Thornton, G.C. III and Mueller-Hanson, R.A. 2004. *Developing Organizational Simulations: A Guide for Practitioners and Students*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.