Journal of Applied Psychology
2015, Vol. 100, No. 4, 1169–1188

0021-9010/15/$12.00 http://dx.doi.org/10.1037/apl0000004

The Interplay of Elicitation and Evaluation of Trait-Expressive Behavior: Evidence in Assessment Center Exercises

Filip Lievens
Ghent University

Eveline Schollaert
University College Ghent

Gert Keen
D*PBO, Amsterdam

In assessment centers (ACs), research on eliciting candidate behavior and evaluating candidate behavior have largely followed independent paths. This study integrates trait activation and trait rating models to posit hypotheses about the effects of behavior elicitation via situational cues on key assessor observation and rating variables. To test the hypotheses, a series of experimental and field studies are conducted. Only when trait-expressive behavior activation and evaluation models work in conjunction, increases in observability are coupled with increases in the interrater reliability, convergent validity, discriminant validity, and accuracy of AC ratings. Implications of these findings for AC theory and practice are formulated.

Keywords: assessment center, behavior elicitation, interpersonal, situational stimuli, role-plays

In many domains, behavioral assessment has emerged as a complement to traditional testing (Lane & Stone, 2006; Sackett, 1998). Examples are objective structured case evaluations in the health professions (e.g., Adamo, 2003), teacher assessments in education (e.g., Pecheone & Chung, 2006), and analogue observation in clinical treatment (e.g., Harvey, Velligan, & Bellack, 2007). In industrial and organizational (I/O) psychology, assessment center (AC) exercises, work samples, and simulations constitute the best-known examples of behavioral assessment (Thornton & Cleveland, 1990).

Generally, the AC method distinguishes itself from traditional testing on the basis of at least two key features. First, the AC method is characterized by a focus on samples of candidate behavior in contextualized situations. To this end, carefully designed simulations (AC exercises) are used as vehicles for eliciting behaviors that are relevant to focal constructs (typically dimensions but also tasks,[1] Borman, 2012; Brummel, Rupp, & Spain, 2009; Jackson, Ahmad, Grace & Yoon, 2011; Thornton & Mueller-Hanson, 2004). Conceptually, the use of simulations for eliciting behavior can be grounded in interactionist theories of human behavior (Mischel & Shoda, 1995; Tett & Burnett, 2003) because it is assumed that the behaviors elicited are a function of candidates' underlying knowledge, skills, abilities, and other c characteristics (KSCOs) and their construal of the job-related situational demands of the simulation (Campion & Ployhart, 2013; Gibbons & Rupp, 2009; Jansen et al., 2013).

As a second distinguishing feature of ACs, human raters are used to observe candidate behavior, classify it, and provide ratings on the focal constructs. To ensure consistency and accuracy in the rating process, raters are required to go through a thorough training and rely on rating aids (e.g., Woehr & Arthur, 2003). In this field, several theoretical models were also developed to better understand the AC rating process (e.g., Lance, Foster, Gentry, & Thoresen, 2004; Lievens & Klimoski, 2001; Zedeck, 1986).

Each of these two key features of ACs and their accompanying research streams has its own focus. In the first research stream, the emphasis is on designing assessment situations for eliciting behavior. In this tradition, candidates' behavioral responses as results of the interaction between their underlying KSAOs and the simulated situations serve as central point of attention. So, this research stream stresses candidates' behavioral responses instead of how assessors subsequently evaluate these responses. The latter is exactly the central focus of the second research stream related to assessor rating processes. As the models in this second research strand focus on the raters, they have paid less attention to questions

This article was published Online First November 3, 2014.

Filip Lievens, Ghent University; Eveline Schollaert, University College Ghent; Gert Keen, D*PBO, Amsterdam, The Netherlands.

We are indebted to Pia Ingold and Francois De Kock for their helpful comments on a prior version of this article and to Paul Sackett and George Thornton for their feedback on the original research ideas behind this study. We would also like to thank Frederik Anseel, Ilse Lievens, Herlinde Pieters, and Ruben De Keyzer for their help in collecting the data for this study. Part of the funding for the studies in this article came from the SIOP Foundation via the Douglas W. Bray and Ann Howard Research Grant Fund.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium. E-mail: filip.lievens@ugent.be

[1] Although the theoretical arguments and empirical studies presented in this article are built around ACs that rely on dimensions as focal constructs, strategies for improving the elicitation and rating of behavior are also relevant for other focal constructs such as tasks (i.e., task-based ACs).

as to how trait-expressive behavior of candidates is activated and how such activation might facilitate rating processes.

The above comparison between these two research streams shows that these strands of research have not been integrated to the extent they should be. Despite the different focus of these research streams through the years, this study's premise is that they essentially deal with different sides of the same coin. Therefore, a first objective of this article consists of integrating trait activation and trait rating models into a more comprehensive model of trait-expressive behavior activation and evaluation. As a second objective, we test hypotheses about the interplay of behavior activation and evaluation through a series of experimental and field studies.

The context of this study is behavioral assessment via role-plays. Nowadays, such one-on-one simulations still constitute the mainstay of interpersonal assessment in I/O psychology and other fields. Although they were traditionally developed as in-person "high-touch" simulations, recent applications have used role-playing in online or telephone-based "low-touch" formats (e.g., Tippins & Adler, 2011).

Theoretical Background

Elicitation of Trait-Expressive Behavior: Theory, Practice, and Research

> After the work had begun, Buster, or occasionally Kippy, might criticize the candidate's plan of operation and suggest other, often incorrect, ways to proceed in order *to test the forcefulness of the man's leadership* [italics added]. Kippy, for instance, might attempt to involve the boss [candidate] in a debate about the relative advantages of the two plans (OSS Assessment Staff, 1948, p. 104).

This quote from the classic book *Assessment of Men* exemplifies that prompting and eliciting candidate behavioral responses has always been a key ingredient in ACs. Not surprisingly, the most recent version of the AC guidelines stipulates that designers should attempt to design exercises that evoke a large number of relevant behaviors because this should give assessors enough opportunities to observe such behavior (International Task Force on Assessment Center Guidelines, 2009).

There exist various ways of eliciting candidate behavior: This can be done on a *general* level in which the whole exercise (its content descriptions and instructions) is assumed to evoke behavior relevant for the focal constructs (McFarland, Yun, Harold, Viera, & Moore, 2005; Schneider & Schmitt, 1992). For instance, a cooperative leaderless group discussion is often regarded as a way of activating leadership emergence and interpersonal behavior, whereas an oral presentation might trigger behavior related to emotional stability and communication. Behavior elicitation can also be done on a *specific* level by planting situational cues in simulations (e.g., Schollaert & Lievens, 2012). To this end, role-players might be trained to provide cues to candidates. For example, a role-player might look slightly distressed as a way of evoking interpersonal sensitive behavior (see Appendix A for other examples). Besides such person-based stimuli, technological cues might also be built into simulations for evoking relevant behavior. Examples are incoming e-mails, telephone messages, or distracter pop-ups (Tippins & Adler, 2011). Apart from eliciting behavior, person-based or technology-based interventions also aim to increase the simulation's realism by inserting a degree of interactivity and reciprocity typical of most work-related situations.

Theoretically, these design characteristics and interventions for eliciting candidate behavior are grounded in interactionist models of human behavior which posit that behavior is a function of the interaction between the person and that person's perception of the situation (Endler & Magnusson, 1976; Jansen et al., 2013; Reis, 2008). In recent years, trait activation theory (TAT, Haaland & Christiansen, 2002; Lievens, Chasteen, Day, & Christiansen, 2006; Tett & Burnett, 2003) has emerged as a dominant interactionist theory in I/O psychology. Figure 1A provides a schematic overview of the main aspects of trait activation theory. As shown, situations serve as moderators that enable the expression of trait-relevant behavior. In this activation process, the moderating effect of situational demands on behavior can be understood through two factors: situation trait relevance and situation strength.

Situation trait relevance refers to the qualitative feature of situational demands that increase the likelihood that individuals will demonstrate more of a particular behavior over other behav-

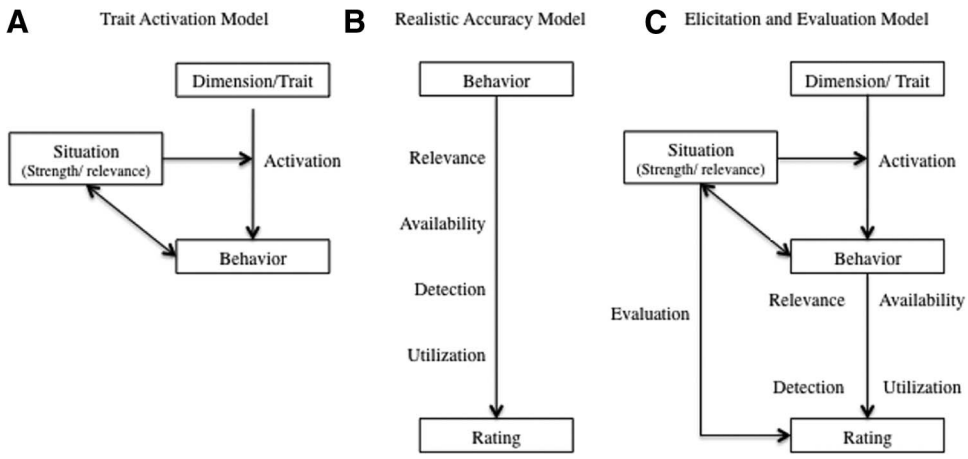

Figure 1. Theoretical models of trait elicitation and evaluation.

iors. For instance, in a computerized in-basket simulation, a new incoming e-mail from a key customer might serve as cue for activating priority setting and decision-making. Some candidates will perceive this as important information, whereas others will simply continue their current activities. The second TAT factor, situation strength, refers to the clarity of a situational demand. A strong situation produces similar behavioral responses across virtually all individuals, whereas the opposite is the case for weak situations (Meyer, Dalal, & Hermida, 2010; Mischel, 1973). For example, cues triggering interpersonal sensitivity might vary from very weak (e.g., showing momentarily a distressed face expression) to very strong (e.g., starting to sob).

Trait activation theory has impacted theoretical conceptualizations underlying ACs as well as spurred on empirical research. The emergence of the mixed-model AC approach which acknowledges trait activation theory as its underlying framework represents an example of its theoretical implications (Borman, 2012; Lievens & Christiansen, 2012; Melchers, Wirz, & Kleinmann, 2012). This mixed AC model is inherently interactionist because AC exercises are then thought to present different situational demands to candidates. Inspired by trait activation theory, recent empirical research has also begun to examine the effects of interventions related to activating trait-expressive behavior (Oliver, Hausdorf, Lievens, & Conlon, in press; Schollaert & Lievens, 2011; Schollaert & Lievens, 2012). Schollaert and Lievens (2011) found that it was possible to teach role-players to use multiple standardized prompts in interpersonal simulations. In particular, prompts were used in about half of the interactions of role-players who were trained to use prompts, whereas role-players without such training used prompts only in 10% of their interactions. No negative effects of the use of prompts on candidates' procedural fairness perceptions of the simulations were reported. Oliver et al. (in press) went one step further and examined how differences in role-player behavior activated different candidate behaviors. In particular, they found that the disposition portrayed by role-players (manifested through verbal statements and prompts, nonverbal communication, and emotional reactions) affected the interpersonal behaviors shown by candidates. For instance, candidates demonstrated fewer relationship building behaviors with role-players with a high affiliation portrayed disposition than with role-players with a low affiliation portrayed disposition (see also the findings on the effects of simulated patients in the health professions field, e.g., Boulet et al., 2009). Whereas the studies discussed above focused on the effects of behavior activation on candidate outcomes, Schollaert and Lievens (2012) also explored the effects on assessor outcomes. They compared role-player prompts with exercise instructions as vehicles for activating candidate behavior. Results attested to significant effects of role-player prompts on the number of behaviors observed. Exercise instructions did not exert any effects.

In short, although the notion of eliciting trait-expressive behavior seems to be firmly rooted in the history of assessment, it is only with the advent of interactionist theories such as trait activation theory that assessment approaches have formally adopted it (see the recent mixed-model AC approach) and empirical research has started to illuminate its effects. Most studies found evidence that the use of situational cues increases the activation of relevant candidate behaviors and influences the type of candidate behaviors exhibited. In other words, prior trait activation research has focused on manipulating situational cues as triggers of underlying candidate traits, with the resulting candidate behaviors and performance serving as the main outcomes.

Evaluation of Trait-Expressive Behavior: Theory, Practice, and Research

Another cornerstone of ACs is that human assessors observe, classify, and rate candidate behavior on focal constructs. There exists a voluminous literature on the effectiveness of approaches for assisting assessors in the complex and demanding rating process. Examples of these interventions include assessor selection, assessor training, and the use of rating aids and video (for reviews, see Lievens, 1998; Woehr & Arthur, 2003).

Alongside the empirical research on assessors and assessor-related factors, several models of the AC rating process have also been developed (see Lance et al., 2004; Zedeck, 1986). For instance, Lievens and Klimoski (2001) made a distinction between rational (i.e., data-driven) assessor, limited capacity, and expert assessor models. However, none of these models were tied to the behavior activation process. Therefore, it might be beneficial to draw from models outside of the AC field to improve our understanding of the interplay between behavior elicitation and evaluation.

In this study, we argue that especially Funder's Realistic Accuracy Model (RAM, Funder, 1995, 1999, 2012) holds promise for connecting elicitation models with assessor rating models for two reasons. First, as a dominant and comprehensive rating model in the personality and social psychological field the RAM presents a theory regarding the circumstances under which and processes by which one might make an accurate appraisal of the psychological characteristics of another person in a social environment (Funder, 1995). Although it was originally developed in the context of personality judgment in interpersonal contexts, it is a metaframework that might also be applied to selection instruments that rely on person perception such as employment interviews (i.e., Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005) or ACs (and especially interpersonal AC exercises such as role-plays). Second, the RAM is a rating model that explicitly recognizes the relevance of behavior, which echoes one of the key factors in trait activation theory. The fact that one of the factors in the RAM is also included in TAT provides a start for "connecting the dots."

As shown in Figure 1B, the RAM makes a distinction between four stages (process variables) that people have to go through to make accurate judgments of others. First, the person has to exhibit behavior relevant to the trait. Indeed, it is difficult to evaluate someone's standing on a trait when that person has not shown relevant trait-expressive behaviors. Clearly, this factor has parallels with the situation relevance factor in TAT. In AC exercises, this factor might be jeopardized when the contextual stimuli have no or limited job-relatedness so that they do not evoke candidate behavior relevant for the focal constructs.

Accessibility constitutes the second process variable. If the focal person shows trait- expressive behavior, that behavior must be available to the rater. This factor refers to the quantity of the trait-expressive information available. At best, numerous behaviors in a wide variety of contexts should be available. In an AC, this factor might be undermined when few relevant behaviors are available for some dimensions.

Third, the trait-expressive, available behavior must also be detected by assessors. To this end, assessors have to scan the complex and fast stream of candidate reactions for relevant behavior. If assessors are cognitively overloaded or distracted, they might not pick up relevant behavior. Video technology might be used to increase assessors' powers of attention.

The fourth and last stage consists of correctly utilizing the trait-expressive, available, and detected behavioral information. In this stage, assessors classify the behavioral information in the correct dimensional category. For example, behaviors such as task division or time keeping should be categorized under planning and organizing instead of under problem solving. This is a complex process because the meaning of behavior might change according to the context and behavior might be affected by more than one trait. Extant assessor training programs such as frame-of-reference training (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002) aim to familiarize assessors with a performance theory so that behaviors are correctly utilized.

In short, the RAM delineates a four-stage process for making accurate ratings. Essentially, these four stages represent four conditions that should ensure that assessors are good (accurate) judges of trait-expressive behavior. Only when each of the four stages is completed successfully, assessors are able to provide accurate ratings of assessees. In the RAM, candidates' trait-expressive behavior is seen as the input to the rating process, whereas we have already noted that behavior is typically considered the output in TAT. So, briefly stated, the RAM starts where the behavior elicitation model ends.

Integrating Behavior Elicitation and Evaluation Models: Development of Hypotheses

Of the two cornerstones of ACs, theory and interventions for improving assessor ratings have received more research attention as compared with the use of situational stimuli as vehicles for eliciting candidate behavior. Importantly, research has not examined how these two aspects influence each other and how they work in tandem. This leaves several pressing questions open. For instance, we do not know how behavior elicitation effects translate into the assessor observation and rating process. That is, we have little understanding about whether and how the elicited behaviors facilitate assessors' ability to detect and utilize them to formulate a rating on the focal AC constructs. In addition, the effects of behavior elicitation on the reliability, discriminant validity, or accuracy of assessor ratings have not been scrutinized. This is because the trait activation model has not been integrated with a trait judgment model. As argued by Lievens, Chasteen, Day, and Christiansen (2006), this is an important limitation as "the flip side of this trait activation model is a trait perception model. This model focuses on assessor judgmental processes and specifies that . . . trait-expressive behavior might be washed out by judgments of assessors (p. 255)".

Therefore, Figure 1C integrates the behavior elicitation and evaluation models, which should enable to ascertain how they work in conjunction by positing hypotheses of the effects of behavior elicitation on key assessor observation and rating variables. As can be seen, the fact that trait-relevant behavior is part of both models makes it possible to map the models into each other and integrate them. In this integrative model, the advantages of the respective models are combined and their limitations reduced. TAT can show its usefulness to the RAM by proposing interventions that influence the elicitation of trait-expressive behavior.[2] Conversely, TAT does not tell us whether the behaviors expressed will be picked up and used correctly by assessors when they provide their ratings. This is the point where the RAM comes in useful because it specifies conditions that should be met for ensuring that trait-expressive candidate behavior is picked up by assessor judgments.

Figure 1C also highlights on which particular issues the integration might be beneficial for advancing the knowledge base in this area. Specifically, when one frames existing research on AC approaches for improving the assessor observation and rating process into the RAM it becomes clear that most of these design interventions have fallen in the behavioral detection (e.g., rating aids, video) and behavioral utilization (e.g., assessor training) categories, whereas there have been few attempts to increase the relevance and accessibility of the behavior in the first place. Hence, calls have been made for more research in this area. For instance, Brannick (2008) proposed "to deliberately introduce multiple dimension-relevant items or problems within the exercise and to score such items" (p. 132). This is in line with TAT as it involves explicit recognition of the importance of building stimuli into the exercises, thereby increasing their situation trait relevance.

Therefore, this study manipulates the relevance and accessibility of trait-expressive behavior by planting situational cues (in the form of standardized dimension-related role-player cues) in AC exercises. Similar to Schollaert and Lievens (2011), we define role-player situational cues as "predetermined statements that a role-player consistently mentions across candidates to elicit behaviors related to specific job-related dimensions" (p. 190). So, we rely on TAT principles for designing interventions to increase the relevance and accessibility of candidate behavior. We then examine how this manipulation produces trickle down effects on key assessor observation and rating variables. To determine these relevant assessor output variables we draw on the RAM. That is, we scrutinize effects in terms of (a) detecting relevant behavior, (b) utilizing that relevant behavior, and (c) rating that behavior. For the latter, this study uses common indices such as the discriminant validity and interrater reliability of assessor ratings. In the following, we posit our hypotheses about the effects of behavior elicitation on each of these dependent variables.

Behavior elicitation effects on detecting dimension-relevant behaviors. As posited by the RAM, assessors must be able to detect trait-relevant, available behavioral information in the stream of candidates' actions. This complex behavioral information is presented to assessors at a fast rate. In other words, when candidates show relevant behavior, it does not imply that assessors will pick it up or will deem it relevant. Evidence as to whether assessors detect behavior and write it down might be obtained from scrutinizing their notes for behavioral statements. These are statements that specifically describe what a candidate says or does (Gaugler & Thornton, 1989). There is evidence that assessors often

[2] It should be acknowledged that research on the RAM identified "good information" as a moderator of the rating process stages. However, the moderators identified typically dealt with differences in information sources, interactions, and media (Funder, 1999).

pick up and write down an insufficient number of behavioral statements per dimension. Bycio, Alvares, and Hahn (1987) even posited that "assessors within an exercise are sometimes, if not usually, forced to base *all* of their judgments on four or five behaviors" (p. 472). This is in line with Lievens and Klimoski's (2001) statement that "unless the exercises provide an opportunity to observe enough behaviors and to do so under (assessor) favorable conditions, it is very difficult to infer traits or dispositions. In this regard, most exercises appear to have been selected or designed more for their face (content) validity, than for their capacity to expose behavior that would reveal the level of specific traits possessed by the participant" (p. 270). Yet, observing a substantial amount of behavior is a vital factor for ACs because behavioral observations serve as basis for providing participants with detailed developmental feedback about their strengths and weaknesses on the constructs of interest. The more of these behavioral examples assessors can provide to candidates, the more useful the feedback might be (Woo, Sims, Rupp, & Gibbons, 2008).

Planting situational stimuli within AC exercises might counter the potential problem of a lack of behavior to be observed (Schollaert & Lievens, 2012). As situational stimuli are explicitly used to enhance situation trait relevance they might increase the amount of relevant behavior available to be observed. Apart from enhancing the quantity of the behavioral information, situational cues might also function as multiple mini situations, thereby increasing the variety of contexts for observing behavior. In turn, this might have beneficial effects on the opportunity for assessors to observe and note down a higher number of behavioral observations. Thus,

> *Hypothesis 1:* Assessors will write down a larger number of behavioral observations per dimension when behavior elicitation is high as opposed to low.

Behavior elicitation effects on correctly utilizing dimension-relevant behaviors. According to the RAM, building in situational cues for increasing the relevance and observability of candidate behavior does not automatically mean that assessors will correctly utilize these behaviors. Once assessors have detected available relevant information, they have to assign it to the correct dimension. Although training (e.g., frame-of-reference training) teach assessors how to do this, classification errors might still occur. In fact, Lievens and Klimoski (2001) refer to this classification process of assigning observed behaviors to dimensions as an unstructured inference and judgment process. This is especially the case for dimensions for which there is little or no information available because assessors might then start wrongly assigning behavior to these dimensions so that they are able to rate them.

Eliciting a higher frequency of relevant behavior via planting situational cues in AC exercises might produce a reduction in such classification errors because each situational cue aims to evoke a specific dimension. Prior theory and research has found that the availability of such situational information facilitates drawing inferences about people's characteristics (Christiansen et al., 2005; Gilbert, 1989; Trope, 1986). That is, the elicitation of behavior via situational cues seems to make situation-behavior linkages more apparent to assessors, thereby increasing the probability that they interpret the behavioral information correctly as to its meaning for the underlying dimension. Thus,

> *Hypothesis 2:* Assessors will make fewer errors when classifying behavioral observations into dimensions when behavior elicitation is high as opposed to low.

Behavior elicitation effects on discriminant validity of assessor ratings. Upon detecting and correctly utilizing the behavioral information, assessors formulate a rating for each dimension. A recurring theme in the literature deals with the challenge of obtaining distinct ratings from assessors on dimensions (Bowler & Woehr, 2006; Kuncel & Sackett, 2014; Lance, 2008; Lance Lambert, Gewin, Lievens, & Conway, 2004; Putka & Hoffman, 2013). That is, ratings on various dimensions within a specific exercise correlate highly. Lack of distinct measurement of dimensions is troublesome if one wants to obtain a fine-grained and differentiated portrayal of candidates' strengths and weaknesses.

Various explanations have been put forward for this lack of distinct measurement. As one possible explanation, it has been argued that exercises do not give candidates enough opportunity to demonstrate dimension-related behavior. Consequently, assessors do not have enough behavioral observations to make distinct ratings on the dimensions (Brannick, 2008; Reilly, Henry, & Smither, 1990). Brannick (2008) further posited that exercises are not always developed to separately rate the dimensions. That is, exercises are often not designed to tap the dimensions as exclusively as possible. So, assessors often need to base their ratings on a few "red hot" items, which means they rely on one particular behavioral reaction to rate several dimensions (Brannick, Michaels, & Baker, 1989).

In light of this recurring problem, it may be useful to use a behavior elicitation strategy for making the different dimensions more overt in the exercises and therefore easier to observe. When multiple situational cues that each aim to evoke a specific dimension are planted in AC exercises more behavior is activated that can be perceived by assessors as falling within the construct domain of a specific dimension so that each dimension can be assessed more exclusively. In turn, this should lead to less spillover in rating candidate behavior for different dimensions and therefore to better discrimination among dimensions. Moreover, it follows from the RAM that the anticipated benefits of behavior elicitation on the detection and correct utilization of a larger quantity of relevant behavioral information should trickle down on the discriminant validity of assessor ratings. That is, once assessors have classified this larger amount of dimension-related behavioral information into the correct dimension, it might follow that they provide more distinct ratings. Thus,

> *Hypothesis 3:* Assessors will make more distinct ratings on the dimensions when behavior elicitation is high as opposed to low.

Behavior elicitation effects on interrater reliability of assessor ratings. When different assessors are asked to evaluate a candidate on a given dimension meta-analytical results show that they agree moderately (i.e., average interrater reliability of .73, Connelly & Ones, 2008). This might be especially the case for ratings on specific dimensions within an exercise because such ratings might be based on rather limited behavioral evidence (Gibbons & Rupp, 2009). When an exercise fails to elicit enough behaviors relevant to a dimension, the representativeness of the

observed behavior for the construct domain may be insufficient to guarantee high consistency in dimension ratings (Epstein, 1979).

As noted above, we posit that increasing the situational relevance and availability of candidate behavior via the inclusion of standardized stimuli in AC exercises will affect the amount of relevant candidate behavior shown in exercises, which might subsequently have beneficial effects on assessors' ability to detect and correctly utilize that behavioral information when rating. Moreover, we argue that these effects might trickle down in terms of positively affecting the consistency of measurement in ACs (Brannick, 2008) because evoking more candidate behavior might increase the standardization of those exercises. That is, the situational stimuli used for evoking candidate behavior might create a common structure for the assessors while observing and rating candidates. Along these lines, behavior elicitation through planting situational stimuli in ACs can be compared with the use of standardized dimension-related questions in interviews. Research in the interview domain has shown that the interrater reliability of high structured interviews is higher than that one of low structured interviews (Conway, Jako, & Goodman, 1995; Huffcutt, Culbertson, & Weyhrauch, 2013). Using multiple situational stimuli for "structuring" AC exercises might exert similar beneficial effects on the interrater reliability of ratings in an exercise. Thus,

> *Hypothesis 4:* Assessors' interrater reliability of dimensional ratings will be higher when behavior elicitation is high as opposed to low.

Overview of Studies

This article presents four studies. Study 1 and Study 2 take place in a simulated assessor environment with psychology student assessors observing and evaluating video recorded candidates in AC exercises. These candidates took part in the exercises as preparation for job applications. Study 3 tests most of the hypotheses in a field AC. Finally, Study 4 expanded our examination with hypotheses related to rating accuracy.

Study 1

Method

Sample. The sample consisted of 177 I/O Psychology students (72.9% women; mean age = 20.4 years, *SD* = 1.16 years) who participated in the study to receive credit for a course in I/O psychology at a large Belgian university. All participants had been in college for 2 years. They had no prior experience as assessors.

Simulated assessor environment. In line with recommendations to incorporate into laboratory research representations of real life by faithfully reconstructing important elements and sources of information that are present in actual situations (e.g., Greenberg & Tomlinson, 2004), we did our utmost best to simulate both the rating task and rating context of assessors. So, participants were told that they would participate as assessors in a simulated assessor environment, and observe and rate video recorded AC candidates. In particular, they were asked to evaluate candidates applying for an entry-level managerial job on four dimensions (i.e., interpersonal sensitivity, organizing and planning, problem solving, and tolerance for stress). Assessors knew that afterward they would be expected to explain their ratings to one another. This common AC practice served as an incentive to take the assessor task seriously.

Prior to serving as assessors, participants were given a half-day assessor training. The trainer was an experienced assessor with a graduate I/O psychology degree. The training program was composed of three main parts: (a) an introduction (2.5 hr) about the basics of ACs; (b) a portrayal (0.5 hr) of the content of the exercise and the four dimensions (using principles underlying frame-of-reference training, Lievens, 2001; Schleicher et al., 2002); and (c) a workshop (1 hr) on the observation and rating process. In that last part, assessors were taught the process of observing, classifying, and rating candidate behavior. The trainer instructed the assessors to make behavioral descriptions instead of nonbehavioral interpretations (Gaugler & Thornton, 1989). All assessors also observed and rated practice videotapes. Afterward, the trainer discussed the observations and ratings made. Discrepancies were clarified and the trainer provided the assessors with feedback.

Design. Trained assessors were randomly assigned to one of two conditions. In one condition, they watched videotapes of candidates with role-players who had attended a role-player training that taught them to use situational cues (high-behavior elicitation condition). In the other condition, they watched videotapes of candidates with role-players who had attended a role-player training that had not taught them such cues (low-behavior elicitation condition).

Video recorded assessee performances. We video recorded the AC performances of 54 final-year students who were pursuing a major in law or sciences (58.2% women, mean age = 22.8 years, *SD* = 1.2 years). These students were recruited by an e-mail to participate in an AC exercise for increasing their experience with selection procedures and to receive feedback on their performance. The AC exercise was a commercially available role-play exercise that was targeted to applicants who pursued entry-level managerial jobs. Students were randomly assigned to either a role-play with a role-player who was not taught to use situational cues (low-behavior elicitation) or a role-play with a role-player who was taught to use situational cues (high-behavior elicitation).

There was anecdotal evidence that these students were motivated and perceived the AC simulation in a similar way as in actual selection practice. For instance, they decided themselves to take part in the role-play as preparation for job applications. In addition, they wore business attire and reported to be nervous. To measure their motivation, candidates completed a test motivation scale (see Arvey, Strickland, Drauden, & Martin, 1990) after the role-play. This scale consisted of five items (e.g., "Doing well on this exercise was important to me") with a Likert-type scale ranging from 1 (*very inaccurate*) to 5 (*very accurate*). The internal consistency reliability of the ratings was .82 and the mean score was 3.8 (*SD* = .53).

Generation of role-player prompts. We conducted a prestudy to generate situational cues that could be used for evoking behavior to the four relevant dimensions. First, to ensure a collection of situational cues that were actually used in AC practice seven experienced assessors (mean age = 38.6 years, *SD* = 7.87, 57% males, mean experience in selection = 13.3 years, *SD* = 8.80) were asked to report possible situational cues that role-players could use to evoke relevant behavior in the role-play. During this phase, 198 situational cues were reported. Second, we refined this list by dropping situational cues that were (a) inappropriate, (b) too

vague, (c) too concrete, and (d) redundant. After this procedure, 84 situational cues were left. Third, these 84 cues were presented to two other groups of assessors: eight graduate students in I/O psychology (62% males, mean age = 27.1 years, *SD* = 2.17) and 12 actual experienced assessors (42% males, mean selection experience = 5.71, *SD* = 7.35). These assessors were asked to retranslate the situational cues to the dimensions. This was done to ensure that the situational cues indeed evoked behavior relevant to their purported dimension. If there was agreement of at least 70% we considered the cue to be adequate for activating candidates' propensities related to the respective dimension, while at the same time representing not too strong of a situation. This led to a final set of 21 situational cues (see Table A1 in Appendix). An example of a cue (to trigger behavior related to interpersonal sensitivity) was "I feel offended by the fact that I had to come here."

Role-player trainings. Nineteen role-players (58% women; mean age = 22.9 years, *SD* = 1.5 years) were randomly assigned to either a role-player training without situational cues or a training with such cues. The trainer was a consultant with a graduate I/O psychology degree and 15 years of assessment experience. Both trainings took half a day and had an identical format. The first 1.5 hr consisted of a lecture wherein role-players learned the content of the AC exercise and their specific role. Next, a videotape of a role-player was presented (1.5 hr). In the role-player training without situational cues, the trainer introduced the videotape by explaining that role-players play their role consistently across candidates, following the International Task Force on Assessment Center Guidelines (2009). In the role-player training with situational cues, the trainer added to all of this a demonstration of the use of situational cues for evoking behavior. In both trainings, the third and last part of the training (2 hr) comprised of exercises, feedback, and discussion. Given that the second part was shorter in the role-player training without situational cues than in the other training, this last part was deliberately a bit longer in this training. At the end, the trainer spent more time discussing the importance of role-players in the AC methodology. Accordingly, the total duration did not differ across the two trainings.

Check of the availability of situational cues. To check to what extent role-players used situational cues in the role-plays with the candidates, four master I/O psychology students (100% women; mean age = 21.8 years, *SD* = .96 years) coded the role-players' behavior. To this end, the coders received a half-day training. They independently wrote down the verbatim behavior of the role-players. Next, they counted the number of times a role-player used situational cues for evoking dimension-related behavior. They also counted the number of interventions that could not be considered relevant for evoking dimension-related behavior. Interrater agreement ($\kappa > .70$) was satisfactory for all dimensions. Discrepancies were resolved through discussion.

Inspection of the number of situational cues per dimension showed that in the low-behavior elicitation condition, on average two situational cues were used (i.e., 7% of all role-player interventions). Conversely, in the high-behavior elicitation condition, on average 13 situational cues were used (i.e., 48% of all role-player interventions). Thus, this manipulation check confirmed that the two conditions are appropriately labeled low versus high in behavior elicitation and that the role-players actually used the cues taught in the role-player training. So, there was more relevant behavioral information evoked and available to assessors when they observed and rated candidates in the high-behavior elicitation condition as compared with the low-behavior elicitation condition.

Measures. During the role-play, assessors completed an observation form on which they noted down and classified the observed behaviors. Immediately after watching an assessee performing in the role-play, they independently rated the assessee on the four dimensions (interpersonal sensitivity, organizing and planning, problem solving, and tolerance for stress) via a 5-point BARS, ranging from *poor* (1) to *excellent* (5). On average, assessors observed and rated three candidates. This process took about 45 min. The following measures were obtained.

Assessors' behavioral observations. The number[3] of behavioral observations written down by assessors served as dependent variable. Two independent and trained coders (100% women; mean age = 22.5, *SD* = .71) with a graduate I/O psychology degree examined the individual notes of the assessors. In a preliminary phase, they independently coded the notes of 20 assessors randomly selected from the assessor pool of Study 1. They counted the number of behavioral observations per assessor. Cohen's (1960) kappa equaled .92. Given this high level of interrater agreement and the fact that the observation forms yielded a total of 5,411 notes, we divided the observation forms in two piles. Each coder was assigned one pile and coded the notes of that pile.

Assessors' incorrect classifications. Assessors classified their behavioral notes into the various dimensions. Accordingly, it was clear which behaviors they had used for making dimensional ratings. The aforementioned coders again scrutinized the notes and counted the number of incorrect classifications. Coding agreement was high ($\kappa > .90$).

Assessors' dimensional ratings. Ratings on each of the four dimensions served as input to test the hypotheses about discriminant validity and interrater reliability.

Results and Discussion

Hypothesis 1 (H1) concerned the influence of behavior elicitation on the amount of behavioral observations written down by assessors. Descriptive statistics of the number of behavioral observations broken down by behavior elicitation condition are presented in Table 1. To Test H1 we conducted a MANOVA with the behavior elicitation condition as independent variable and the number of behavioral observations for problem solving, interpersonal sensitivity, planning, and tolerance for stress as a set of four dependent variables. There was a multivariate main effect for behavior elicitation $F(4, 390) = 9.153$, $p < .001$ (partial $\eta^2 = .09$). Overall, assessors in the high-behavior elicitation condition wrote down 12% more behavioral observations than those in the low-behavior elicitation, ($M_{high} = 11.52$, $M_{low} = 10.30$, $d = .25$). Thus, on the overall level, there was support for H1.

Follow-up univariate analyses revealed that the main effect of behavior elicitation was significant and in the expected direction for

[3] We also ran the analyses with the proportion of behavioral observations (i.e. the ratio of the number of behavioral observations to the total number of observations) because more prolific assessors might produce more "raw" behavioral observations (Gaugler & Thornton, 1989). We reran the analyses for incorrect classifications in the same way. Analyses with proportions instead of raw numbers produced similar results.

Table 1
Means, Standard Deviations, and Effect Sizes of the Number of Behavioral Observations Broken Down by Condition in Study 1

	Low behavior elicitation (N = 199)		High behavior elicitation (N = 196)			
	M	*SD*	*M*	*SD*	*d*	*p*
Problem solving	3.13	2.27	2.85	1.73	−.14	.16
Organizing and planning	2.22	1.65	2.62	1.47	.25	.01
Interpersonal sensitivity	3.48	1.86	3.92	2.26	.21	.03
Tolerance for stress	1.47	1.16	2.13	1.60	.48	.00
Total	10.30	4.67	11.52	4.93	.25	.01

Note. Positive effect sizes (*d* values) mean that the number of behavioral observations was higher in the high behavior elicitation condition.

the dimensions planning and organizing (M_{high} = 2.62, M_{low} = 2.22, d = .25), interpersonal sensitivity (M_{high} = 3.92, M_{low} = 3.48, d = .21), and tolerance for stress (M_{high} = 2.13, M_{low} = 1.47, d = .48), but not for problem solving, (M_{high} = 2.85, M_{low} = 3.13, d = −.14). Thus, on the dimensional level, our results supported H1 for three of the four dimensions.

To test Hypothesis 2 (H2) we conducted a MANOVA with the behavior elicitation condition as independent variable and the number of incorrect classifications for problem solving, interpersonal sensitivity, planning, and tolerance for stress as a set of four dependent variables. There was a multivariate main effect for behavior elicitation $F(4, 390) = 6.521$, $p < .001$ (partial $\eta^2 = .06$). However, as shown in Table 2, the effect was in the opposite direction as hypothesized because across all dimensions there was a 20% increase in incorrect classifications in the high-behavior elicitation condition, (M_{high} = 1.73, M_{low} = 1.38, d = −.20). Thus, on the overall level, no support was found for H2. Follow-up univariate analyses showed that this multivariate effect was driven by the dimension planning because there were significantly more incorrect classifications for this dimension in the high-behavior elicitation condition, (M_{high} = 1.21, M_{low} = .67, d = −.43). So, there was also no support for H2 on the dimensional level.

Table 2
Means, Standard Deviations, and Effect Sizes of the Number of Incorrect Classifications Broken Down by Condition in Study 1

	Low behavior elicitation (N = 199)		High behavior elicitation (N = 196)			
	M	*SD*	*M*	*SD*	*d*	*p*
Problem solving	.49	.95	.35	.73	.17	.10
Organizing and planning	.67	1.10	1.21	1.41	−.43	.00
Interpersonal sensitivity	.17	.47	.09	.35	.19	.08
Tolerance for stress	.09	.34	.13	.47	−.11	.25
Total	1.38	1.61	1.73	1.82	−.20	.04

Note. Positive effect sizes (*d* values) mean that the number of incorrect classifications was lower in the high behavior elicitation condition.

To Test Hypothesis 3 (H3) and Hypothesis 4 (H4), we used generalizability analysis (Brennan, 2001). As opposed to classical test theory, generalizability analysis permits the simultaneous estimation of many sources of variance inherent in ratings. In this study, generalizability analysis partitioned the sources of variance in AC scores in three sources: candidates, dimensions, and assessors. Candidates (C) served as the object of measurement. Assessors (A) and dimensions (D) were the facets. As assessors rated each candidate on all dimensions, these two facets were crossed with each other. Assessors were nested in candidates because assessors did not rate all candidates. In addition, in some cases there were missing values. This creates an unbalanced design that was not fully crossed. As noted by Putka and Hoffman (2013), modern variance component estimation procedures (e.g., restricted maximum likelihood estimation) enable dealing with such sparseness in the data resulting from assessors not being fully crossed with candidates (see also Searle, Casella, & McCulloch, 2006). Table 3 presents the estimated variance components across the two conditions. Variance components reflect each facet's contribution to the total variance. As they depend on the scale of measurement (here a 5-point scale), we also present their relative magnitudes (Shavelson & Webb, 1991), which is done by the percent contribution of each variance component.

H3 proposed that the discriminant validity of assessor ratings would be higher in the high-behavior elicitation condition. Evidence of discriminant validity is derived from the variance component associated with the Candidates × Dimensions (C×D) interaction. A higher value of this variance component indicates more distinct candidate ratings across dimensions (Putka & Hoffman, 2013). There was only a slight increase in the explained variance associated with the C×D interaction between the low-behavior elicitation condition (5.6%) and the high-behavior elicitation condition (6.5%). So, H3 was not supported.

H4 posited that the interrater reliability of assessor ratings would be higher in the high-behavior elicitation condition. Table 3 shows that variance components related to unreliability (i.e., Assessors, Assessors × Candidates (A×C) interaction, Assessors × Dimension (A×D) interaction, see Putka & Hoffman, 2013) were not consistently lower in the high-behavior elicitation condition

Table 3
Results of Generalizability Analyses Broken Down by Condition in Study 1

	Low behavior elicitation		High behavior elicitation	
Effect	*VC*	Explained variance (%)	*VC*	Explained variance (%)
C(andidates)	.07	5.9	.05	4.0
A(ssessors)	.00	0.0	.05	3.7
D(imensions)	.01	0.9	.03	2.4
C × A	.37	30.4	.31	24.7
C × D	.07	5.6	.08	6.5
A × D	.00	0.0	.06	4.8
Error	.70	57.2	.68	54.0

Note. *VC* = estimated variance components. The explained variance is the percentage of the sum of the variance components (i.e., the total variance) that each variance component accounts for.

than in the low-behavior elicitation condition. Whereas the variance component related to the A×C interaction was lower, the variance components related to Assessors and to the A×D interaction were higher. In both conditions, the large variance component of the A×C interaction is noteworthy, implying that in both conditions assessors differed a lot in their ratings of the candidates (regardless of the dimension assessed). Hence, planting situational cues for evoking behavior did not seem to lead to more reliable assessor ratings. So, H4 was not supported.

In short, the findings of Study 1 are mixed. The positive news is that behavior elicitation via situational cues enhanced the observability of dimensions because assessors detected more relevant behavior related to three of four dimensions (see also Schollaert & Lievens, 2012). These results are important because the developmental feedback given to candidates after an AC is among others contingent upon the number of behavioral observations gathered. The lack of an effect for the problem solving dimension might indicate that using cues works better for evoking behavior for dimensions with a personality loading (tolerance for stress, interpersonal sensitivity, planning and organizing) than for dimensions with a *g* loading (problem solving). A related explanation is that behaviors can be more readily elicited and observed for social dimensions than for problem solving dimensions. In any case, there seems to be a class of dimensions (cognitively oriented dimensions) for which including situational stimuli for activating behavior does not work.

The negative news, however, is that no support for the other hypotheses was found: The increased detection of relevant behavior did not enhance assessors' ability to correctly classify behavior and make distinct ratings. In the high-behavior elicitation condition, interrater reliability was sometimes even lower.

How can these mixed results be explained? We believe that that the classification results help to shed light onto this. That is, the increased observability did not lead to improved discriminability and interrater reliability because assessors might have been overwhelmed by the increased number of behavioral observations and did not always classify those behaviors into the correct dimensions. In fact, the RAM posits that behavior exhibited leads to "good" ratings only when assessors proceed successfully through all four proposed stages. Although the relevance, availability, and detection conditions were met, our results show that assessors did not correctly utilize (classify) the detected behavioral information. So, one of the essential chains in the RAM process might have been broken.

More generally, Study 1 manipulated factors related to behavior elicitation (increasing the availability of relevant behavior) but did not influence the assessor process itself. In particular, role-players learned the situational cues in the role-player training but these cues were not included in the assessor training so that assessors were not familiarized with these cues. Hence, the lack of support for the effects of behavior elicitation on the discriminant validity and reliability of assessor ratings might be due to the fact that assessors received too limited guidance for correctly utilizing and rating the available, relevant, and detected behavioral information. This issue might be solved by making assessors familiar with the behavior eliciting situational cues, which was done in Study 2.

Study 2

In Study 2, we aim to impact *both* the behavior elicitation process and the assessor rating process. That is, we build situational stimuli not only in the exercises via role-players, but we also want to make sure that those situational stimuli influence the rating process by including them in the assessor training. So, assessors are also familiarized with the behavior evoking cues and with which dimensions they activate.

In Study 2, two conditions are distinguished. The first condition is the same as the high-behavior elicitation condition of Study 1. So, role-players are taught the behavior evoking cues, whereas assessors are given a traditional assessor training. In the other condition, role-players again learn how to use situational cues for evoking behavior. However, the assessor training is also adjusted. Besides the traditional part of the training, assessors are also familiarized with the situational cues associated with the dimensions.

We compare these two conditions as a means of retesting the hypotheses that did not receive support in Study 1. When we familiarize assessors with the situational cues, we expect that the beneficial effects on behavioral classification, interrater reliability, and discriminant validity will occur this time. The main reason is that making assessors familiar with the cues should ensure they correctly utilize the larger quantity of relevant behavioral information. Accordingly, there would no longer be a break down in the RAM process.

Specifically, when assessors are knowledgeable about the dimension-related cues used by role-players, they are also "prompted" to direct their attention to the potential occurrence of candidate behavior related to a particular dimension as a response to the situational cue. In interactionist terms, acquainting assessors with role-player cues alerts them to the situation ("if"), and the following candidate behavior ("then," see Mischel & Shoda, 1995), which may help them to see the potential situation-behavior relations. As this provides guidance to assessors as to which behavior is evoked by which dimension, we expect that it will result in fewer classification errors in assigning behaviors to dimensions. It will also lead to better discriminations between dimensions given the increased probability that assessors interpret the behavioral information correctly as to its meaning for the underlying dimension. Finally, it might also provide assessors with more structure when observing and rating the candidates because they know when a particular dimension is evoked by a situational cue, thereby increasing the interrater reliability of their ratings. This leads to the following hypotheses.

> *Hypothesis 5:* Assessors familiar with behavior evoking cues will correctly classify a larger number of behavioral observations than assessors unfamiliar with them.
>
> *Hypothesis 6:* Assessors familiar with behavior evoking cues will make more distinct ratings on the dimensions than assessors unfamiliar with them.
>
> *Hypothesis 7:* The interrater reliability of dimensional ratings of assessors familiar with behavior evoking cues will be higher than that one of assessors unfamiliar with them.

Apart from testing these hypotheses, we also examine in Study 2 whether these effects generalize to another exercise (oral pre-

sentation). Moreover, use of two exercises in Study 2 permits investigating the effects of behavior elicitation on another long-standing AC issue, namely the convergent validity of assessor ratings. Prior research has documented that assessor ratings of the same dimension across different exercises typically correlate lowly. As cogently summarized by Speer, Christiansen, Goffin, and Goff (2014), candidate-related and assessor-related reasons serve as the two complimentary explanations for this finding. On the candidate side, it has been posited that AC exercises cannot be considered parallel exercises as they provide different demands to candidates. Hence, it is not realistic to expect that candidates behave and perform similarly across them. Cross-situationally consistent behavior of candidates can be expected only when exercises contain similar trait relevant cues (Lievens et al., 2006). On the assessor side, it has been argued that assessors often have too little behavioral evidence to provide consistent ratings on the dimensions across exercises (e.g., Bowler & Woehr, 2006; Bycio et al., 1987; Speer et al., 2014). Research on the dimension of oral communication attests to the importance of collecting sufficient observations. That is, oral communication has emerged as one of the "best" dimensions in terms of convergent validity (Bowler & Woehr, 2006; Kauffman, Jex, Love, & Libkuman, 1993). This is understandable because assessors have ample opportunity for observing oral communication across exercises.

We argue that familiarizing assessors with the situational cues for evoking behavior might impact on both of these explanations. First, planting similar cues across exercises for evoking behavior might increase the cross-situational consistency of candidate behavior. For instance, in both a role-play and presentation exercise, role-players might use a cue such as "What now?" to evoke the dimension of planning and organizing. In other words, use of multiple situational cues for a specific dimension across different exercises might increase the correspondence of the different exercises in triggering behavior related to a particular dimension. Second, making assessors aware of these situational cues should ensure that this increased cross-situational consistency is not "washed away" at the assessor level (Tett & Burnett, 2003). That is, when similar situational cues are used across two exercises and when assessors discern these situational cues as relevant hints of a respective dimension, they should experience fewer problems in correctly utilizing the detected behavior. Thus,

> *Hypothesis 8:* The convergent validity of dimensional ratings of assessors familiar with behavior evoking cues will be higher than that one of assessors unfamiliar with them.

Method

Sample. A similar sample as in Study 1 was used. A group of 186 psychology students (69.1% women; mean age = 20.9 years, *SD* = 1.32 years) participated in the study to receive credit for an I/O psychology course at the same large Belgian university as in Study 1. All students had been in college for 2 years. They had no prior experience as assessors. Similar to Study 1, they were first thoroughly trained and next placed in a simulated assessor environment to perform as assessors and evaluate video recorded assessees.

Design. Participants were randomly assigned to one of two conditions: an assessor training without familiarizing them with the behavior evoking cues (see Study's 1 high-behavior elicitation condition) and an assessor training that familiarized them with these cues. Note that assessors of both conditions were asked to observe and rate videotapes of the *same* candidates interacting with role-players using situational cues.

Assessor training. Both trainings lasted half a day. In both trainings, the trainer was an experienced assessor with an I/O psychology graduate degree. The *assessor training without situational cues* followed the same procedure as described in Study 1. The *assessor training with situational cues* had the same format but assessors were also made familiar with the cues. To this end, they received per dimension and exercise a list of the cues. The filler information used in the assessor training without situational cues was the same as the one used in Study 1 (see description of role-player trainings in Study 1).

Video recorded assessee performances. Similar to Study 1, we invited final-year students to participate in AC exercises as preparation for job applications. This time 26 students (56.1% women, mean age = 22.2 years, *SD* = 1.7 years) were video recorded while participating in exercises with role-players who used behavior evoking situational cues. Besides the role-play of Study 1, they also participated in an oral presentation.

Generation of role-player prompts. In the role-play, the same situational cues were used as in Study 1. For the oral presentation, situational cues were generated using the same procedure of Study 1. The final list of situational cues for the presentation included 18 cues (see Table A2 in Appendix A). An example was "Ask how the budget will be allocated" (to elicit planning and organizing).

Role-player training. The same approach was used as in the high-behavior elicitation condition of Study 1. The training lasted longer than in Study 1 because Study 2 role-players were taught to use cues related to two exercises (role-play and oral presentation).

Check of the availability of situational cues. This manipulation check was conducted in the same way as in Study 1. The manipulation check confirmed that role-players actually used a substantial number of the cues taught in the training.

Measures. The observation and rating process was similar to Study 1 but now assessors watched video recorded assessee performances in two exercises (role-play and presentation). To control for order effects, we varied the sequence wherein the exercises were presented to assessors.

This observation and rating process lead to similar measures as in Study 1. Two trained graduate I/O psychologists (100% women; mean age = 22.5 years, *SD* = .71 years) counted the number of behavioral observations and incorrect classifications in the notes of the assessors (N = 4,293). Coding agreement for both of these measures was high ($\kappa > .90$ for both exercises). In addition, ratings on the dimensions served as input to test hypotheses about discriminant/convergent validity and interrater reliability.

Results and Discussion

We started by inspecting the number of behavioral observations noted by assessors. In Study 2, assessors recorded an average of 11.46 behavioral observations per role-play performance and an average of 11.73 behavioral observations per presentation performance. So, assessors noted down about the same number of behavioral observations as in the high-behavior elicitation condition of Study 1 (M = 11.52), replicating that result. We did not

expect that making assessors familiar with the behavior eliciting cues would exert additional effects on the detection of behavior. Results confirmed this because there were negligible differences in the number of behavioral observations across conditions in Study 2.

In Hypothesis 5 (H5), we posited that familiarizing assessors with behavior elicitation cues would produce fewer incorrect behavioral classifications. Table 4 presents descriptive statistics of the number of wrong behavioral classifications per condition. To test H5 we conducted a MANOVA with behavior elicitation as independent variable and the number of incorrect classifications for problem solving, interpersonal sensitivity, planning, and tolerance for stress in both exercises as dependent variables. Exercise was a within-subjects variable. There was a multivariate main effect for behavior elicitation, $F(1, 154) = 16.027$, $p < .001$ (partial $\eta^2 = .09$). If we translate the significant effects in practical terms, there were 58% and 41% reductions in incorrect classifications across dimensions in the role-play and presentation, respectively when assessors were familiarized with the behavior elicitation cues. So, on the overall level, we found support for H5.

Follow-up univariate analyses showed that assessors who were familiar with the behavior elicitation cues made significantly fewer incorrect classifications for problem solving and organizing and planning in the role-play and for organizing and planning in the presentation. So, although all the differences were in the hypothesized direction, they reached significance for two of the four dimensions in the role-play and for one dimension in the oral presentation. Thus, on the dimension level these results partially supported H5.

Similar to Study 1, generalizability analysis was used to test the other hypotheses. As we used two exercises in Study 2, the generalizability analysis had three facets: assessors (A), exercises (E), and dimensions (D). Candidates (C) served as object of measurement. Assessors were nested within candidates because assessors did not rate all candidates. Table 5 presents results of the generalizability analysis of both conditions. Hypothesis 6 (H6) posited the discriminant validity of assessor ratings to be higher in the condition in which they are familiarized with the behavior elicitation cues. The explained variance associated with the C×D interaction in the condition without situational cue familiarization was 2.9%, whereas it was 4.3% in the condition with situational cue familiarization. Although both percentages are still small, the explained variance of the C×D interaction nearly doubled when assessors were familiar with the cues. So, there was support for H6.

Hypothesis 7 (H7) stated that the interrater reliability of assessor ratings would be higher in the condition in which assessors are familiarized with the behavior elicitation cues. To this end, we inspected the variance components contributing to unreliability (Putka & Hoffman, 2013) Although some of these variance components (i.e., assessors and all two-way interactions between assessors and other facets) were similar across conditions, it was striking that the Assessors × Candidates (A×C) interaction explained twice as much variance in the condition when assessors were unfamiliar with the cues (14.9%) than in the condition when they were familiar with them (7.0%). This suggests that familiarizing assessors with the cues had a positive impact on the interrater reliability of their ratings. So, H7 was supported.

Hypothesis 8 (H8) stated that the convergent validity of assessors' ratings would be higher in the condition in which assessors were familiarized with the behavior elicitation cues. Convergent validity evidence can be inferred from a low variance component associated with the Candidates × Exercises (C×E) interaction. In support of H8, the C×E interaction explained lower variance in the condition wherein assessors were familiarized with the cues (1.2%) than it did in the condition without such familiarization (10.6%).

Table 4
Means, Standard Deviations, and Effect Sizes of the Number of Incorrect Classifications Broken Down by Condition in Study 2

	Assessors unfamiliar with cues ($N = 76$)		Assessors familiar with cues ($N = 80$)			
	M	*SD*	*M*	*SD*	*d*	*p*
Role-play						
Problem solving	0.49	0.79	0.05	0.22	.86	.00
Organizing and planning	1.41	1.57	0.75	1.29	.46	.00
Interpersonal sensitivity	0.13	0.41	0.11	0.39	.05	.75
Tolerance for stress	0.13	0.41	0.05	0.22	.26	.12
Total	2.12	2.12	0.89	1.40	.70	.00
Presentation						
Problem solving	0.33	0.64	0.28	0.71	.08	.51
Organizing and planning	1.59	1.85	0.88	1.52	.43	.00
Interpersonal sensitivity	0.16	0.43	0.06	0.24	.28	.06
Tolerance for stress	0.14	0.60	0.10	0.41	.09	.49
Total	2.12	2.13	1.24	2.07	.42	.01

Note. Positive effect sizes (*d* values) mean that the number of incorrect classifications was lower in the condition in which assessors were familiar with the behavior evoking cues.

Table 5
Results of Generalizability Analyses Broken Down by Condition in Study 2

	Assessors unfamiliar with cues		Assessors familiar with cues	
Effect	*VC*	Explained variance (%)	*VC*	Explained variance (%)
C(andidates)	.10	7.5	.21	14.6
A(ssessors)	.04	2.6	.04	3.1
D(imensions)	.02	1.8	.03	1.8
E(xercises)	.01	0.5	.00	0.0
C × A	.20	14.9	.10	7.0
C × D	.04	2.9	.06	4.3
C × E	.14	10.6	.02	1.2
A × D	.02	1.6	.00	0.0
A × E	.01	0.6	.01	1.1
D × E	.02	1.3	.00	0.2
C × A × D	.04	2.8	.12	8.8
C × A × E	.08	5.7	.10	6.8
C × D × E	.01	1.0	.03	1.9
A × D × E	.00	0.0	.02	1.5
Error	.62	46.3	.67	47.8

Note. *VC* = estimated variance components. The explained variance is the percentage of the sum of the variance components (i.e., the total variance) that each variance component accounts for.

Generally, Study 2 results lend support to our hypotheses. Assessors who are familiar with the situational cues for evoking behavior make fewer classification errors for some dimensions,[4] provide somewhat more distinct ratings, use the dimensions in a more consistent fashion across the exercises, and agree more with each other. This underscores the crucial importance of the interplay between behavior activation and evaluation approaches. Only when these two systems are linked to each other and work in tandem, beneficial effects are found. A limitation, however, is that these results were obtained in a lab setting. Therefore, in Study 3 we set up a quasi-experiment in an operational AC to examine whether these behavior elicitation and rating effects generalize to the field.

Study 3

Study 3 tested the hypotheses related to discriminant validity (H6), interrater reliability (H7), and convergent validity (H8) in a field setting.[5] At the backdrop of the conceptual arguments posited in Study 1 and 2, we expect support for the hypotheses only when assessors are made aware of the behavior evoking situational stimuli of the role player.

Method

Sample. The sample consisted of 498 candidates who participated in an AC for either selection or development purposes in the Netherlands. All candidates applied for managerial/supervisory jobs. AC participants came from different organizations and had all gone through a screening stage (on the basis of traditional tests and inventories). Candidates worked in the service sector (either the health care or transportation industry). On average, 90% of the candidates were between 30 and 45 years and 60% of them were males. All candidates had prior work experience. However, specific information related to their prior work experience and education was not available.

Description of assessment center. On the basis of job analyses the following dimensions were measured: decisiveness, reducing resistance, clarifying objections, removing objections, and giving instructions. Each AC consisted of three role-plays (e.g., with a problem subordinate, colleague). Candidates were expected to achieve a specific goal (e.g., staying in the team vs. leaving it, attending a workshop).

In each simulation, two assessors rated the dimensions. The assessor pair always consisted of one professional assessor and one assessor from the commissioning organization. Professional assessors had a background in human resources, management, or psychology (representing approximately 40% of the professional assessors). All assessors were provided with the same 1-day training that conformed to the International Task Force on Assessment Center Guidelines (2009). This training consisted of explaining the dimensions, exercises and rating process, and of several practice videos. Per exercise, assessors provided dimensional ratings on a 7-point scale (from 1 = *poor* to 7 = *excellent*). Similar to Study 1 and 2, no behavioral checklists were used. Aggregation of the dimension ratings made per exercise occurred after each simulation through averaging (without discussion). A senior assessor with a graduate I/O psychology degree managed each AC.

Design. In this field setting, it would have been unethical to randomly assign candidates of the same selection procedure to different behavior elicitation conditions, thereby potentially giving some candidates more opportunity to demonstrate behavior than others. Therefore, we used a quasi-experimental design and implemented the different conditions in naturally occurring settings. To this end, we worked together with the consultancy firm that was responsible for the ACs. Basically, the "no cue" condition (see below) represented how ACs were carried out all along, whereas the other condition was an upgrade of their exercise, rating, and training practices. Although in this quasi-experimental design candidates were not randomly assigned to the conditions, we did our utmost best to ensure that the AC was similar across these conditions. For instance, we tried to keep the exercises (role-plays), dimensions, and type of assessors (one professional assessor and one assessor of the organization) constant. In addition, the candidate profiles and jobs were similar across conditions.

Two conditions were distinguished. In one condition ($N = 314$), no formal attempts were implemented to evoke dimension-related behavior. Role-players followed their portrayal and exercise guidelines. In the other condition ($N = 184$), role-players were trained to use situational stimuli for evoking behavior *and* assessors were made aware of these stimuli. The number of cues per dimension was maximized to four, from relatively weak to relatively strong (e.g., nonverbal expression, verbal implicit expression, verbal explicit expression). As in the prior studies, role-players did their best to elicit the dimensions in a sequence that matched a "natural" conversation.

Results and Discussion

To test H7 we computed intraclass correlations (ICC 1.2) among the ratings of the assessor pair. The average ICC value was .51 in the low-behavior elicitation condition. In the high-behavior elicitation condition that familiarized assessors with the cues, the average ICC value equaled .75. These results support H7.

Given that Study 3 was conducted in the field (where a small number of assessors typically rate a large number of candidates), we could run various confirmatory factor analysis (CFA) models on the correlation matrix[6] of each condition to examine the other hypotheses. In line with prior research (e.g., Bowler & Woehr, 2006), we specified various models that represented different conceptualizations of the structure of AC ratings, namely the dimensions-only model, the exercises-only model, the exercises and one general dimension model, and the dimensions and exercises model. Evidence for the convergent and discriminant validity of ratings is obtained when there is support for the dimensions and exercises model and when dimension loadings are significant.

Table 6 shows that in the condition in which role-players did not use situational cues the best fit was obtained for the dimensions and exercises model: The RMSEA was below .06 and the RNI

[4] It is not unexpected that the number of incorrect classifications did not decrease for other dimensions following the Study 2 manipulation. The "control" condition in Study 2 (i.e., use of role player cues but assessors unaware of those cues) already builds on the best available evidence regarding assessor training (e.g., establishing a frame-of-reference).

[5] As the notes of assessors were not kept in this operational AC we could not reexamine the hypotheses related to behavioral observation and classification.

[6] The full multitrait–multimethod matrices are available from the first author.

Table 6
Results of Confirmatory Factor Analyses Broken Down by Condition in Study 3

Models	Chi^2	*df*	TLI	RNI	RMSEA	90% CI of RMSEA	Number of improper estimates
No formal behavior elicitation (N = 314)							
Dimensions-only	543.012	80	.382	.529	.138	[.127, .149]	1
Exercise-only	343.92	87	.685	.739	.099	[.088, .110]	0
Exercises and one dimension	220.81	72	.779	.849	.083	[.070, .095]	0
Dimensions and exercises	126.09	62	.890	.935	.059	[.044, .073]	1
Behavior elicitation and assessors familiar with cues (N = 184)							
Dimensions-only	307.24	80	.445	.577	.153	[.135, .171]	2
Exercise-only	175.24	87	.802	.836	.092	[.071, .111]	0
Exercises and one dimension	151.48	72	.784	.852	.096	[.074, .116]	1
Dimensions and exercises	66.65	62	.985	.991	.025	[.000, .061]	0

Note. TLI = Tucker Lewis Index; RNI = Relative Centrality Index; RMSEA = Root Mean Square Error of Approximation.

higher than 90. However, the TLI was below .90 and there was an improper parameter estimate (negative error variance). In the condition in which role-players used situational cues and assessors were familiar with them, the best fit was obtained for the dimensions and exercises model (RMSEA = .025; TLI = .985, and RNI = .991) and there were no improper parameter estimates. Inspection of these estimates showed that all dimension loadings were significant. Notably, dimensions explained on average more variance (32.5%) in assessor ratings than exercises (28.3%) did. Overall, these results support H6 (discriminant validity) and H8 (convergent validity).

Thus, Study 3 replicated our prior results in an operational AC because the measurement properties of assessor ratings in the condition in which both role-players and assessors were taught the situational cues were superior to the ones in the condition without use of such cues. Taken together, the three prior studies suggest that strategies for evoking behavior should be coupled with adjustments in assessor evaluation models for leveraging the detection, utilization, and reliable and distinct evaluation of behavior in ACs. Yet, we still do not have evidence as to whether this strategy also exerts positive effects on rating accuracy as key criterion in the RAM model, which is the focus of Study 4.

Study 4

Study 4 tests whether behavior elicitation via situational cues and familiarizing assessors with these cues leads also to more accurate ratings. The RAM posits that accurate ratings will result from assessors detecting and correctly utilizing a large quantity of relevant behavior. Social psychological RAM research typically dealt with "*operant*" behaviors. These are spontaneous behaviors that are shown when two unacquainted people are put together in a room to have a conversation with each other (without clear instructions on the topics to be discussed, Funder, 1999; Funder & Colvin, 1991; McClelland, 1984). In our integration of the RAM with TAT, we deliberately planted situational cues for evoking behavior in the exercises. Such behaviors are referred to as "*respondent*" behaviors because they are elicited in response to a specific identifiable and set up stimulus.

Essentially, the distinction between operant and respondent behavior is related to the concept of situational strength in TAT (Meyer et al., 2010; Tett & Burnett, 2003), with the situation assumed to be stronger for respondent than for operant behavior. As noted above, the logic behind the use of situational stimuli for evoking behavior is that they are relevant for evoking dimension-related behaviors, while at the same time not being too strong to obviate individual differences in exhibiting these behaviors. In other words, we argue that planting situational stimuli in exercises for evoking behavior should increase accuracy in the high-behavior elicitation condition when assessors are aware of these stimuli. Thus,

> *Hypothesis 9:* There will be an interaction between behavior elicitation and assessor familiarization. Accuracy will be highest in the high-behavior elicitation condition only for assessors familiar with the behavior evoking situational cues.

Method

Sample. The sample consisted of 111 industrial and organizational psychology students who participated in the study to receive credit for a human resource management course at the same large Belgian university of Study 1 and 2. The sample included 66% women and 34% men, with a mean age of 21.6 years (SD = 2.3 years). Participants had been in college for 3 years and had an undergraduate psychology degree. They had no prior experience as assessors.

Design. A 2 × 2 design was used by crossing two video recorded assessee performances (low vs. high-behavior elicitation via situational cues) with two assessor familiarization strategies (unfamiliar vs. familiar with situational cues). As explained below, behavior elicitation was a within-subjects factor and assessor familiarization strategy a between-subjects factor.

Video recorded assessee performances. We randomly selected one video recorded role-play performance from each of the two behavior elicitation conditions (low and high) of Study 1. So, in Study 4 all assessors rated two video recorded assessee performances as stimuli. The order of presenting the assessee performance was counterbalanced. We decided to use real candidate performances instead of scripted/constructed ones for external validity reasons. In each of these video recorded performances, a male assessee interacted with a female role-player. Inspection of the codings of Study 1 confirmed that one role-player was high on behavior elicitation (55% of her interventions contained situational cues), whereas the other one was low on behavior elicitation (10% of her interventions had situational cues).

Assessor familiarization. Participants were randomly assigned to one of two conditions: an assessor training without situational cues versus one with situational cues. The trainings (lecture, exercises, feedback, and trainer) were the same as in Study 2.

Measures. The observation and rating process was similar to Study 1 and 2. We computed the accuracy of assessors' ratings of each of the two assessees. To this end, Borman's differential accuracy index (BDA, Borman, 1977) was calculated per assessor. This index was obtained by computing within-assessor correlations between the assessor's ratings on the dimensions of the assessee and the corresponding "target" scores (with an *r*-to-Fisher's-*z* transformation). Higher scores on BDA indicate better accuracy. Cronbach's accuracy indices for the assessors in each condition could not be computed because they require multiple assessees.

To develop target scores of the two video recorded performances, three assessors with a graduate I/O psychology degree and mean assessor experience of 14 years viewed them under optimal conditions (e.g., possibility to pause and rewind, see Sulsky & Balzer, 1988) and gave dimensional ratings on a 5-point scale, with 1 (*poor*) and 5 (*excellent*). The interrater agreement among the experts equaled .89 (ICC 2.1, Shrout & Fleiss, 1979). We averaged the ratings per dimension across the experts to obtain target scores per dimension.

Results and Discussion

Hypothesis 9 (H9) posited that accuracy would be highest when assessors who are familiar with the behavior evoking situational cues rate candidates in the high-behavior elicitation condition. Descriptive statistics are presented in Table 7. To test H9 we conducted an ANOVA with accuracy as dependent variable and behavior elicitation (video recorded assessee) as a within-subjects factor and assessor familiarization strategy as between-subjects factor. There was a multivariate interaction effect between behavior elicitation and assessor familiarization strategy, $F(1, 109) = 9.062$, $p < .01$ (partial $\eta^2 = .08$).[7] Follow-up univariate analyses showed that the accuracy of assessors who were taught the situational cues and who rated the candidate in the high-behavior elicitation condition was significantly higher than assessors' accuracy in the three other conditions. Note that there were no significant differences in accuracy across these other conditions. So, there was no reduction in the accuracy of assessors who observed behavior that was explicitly elicited via situational cues (instead of "naturally" occurring behavior).

In sum, Study 4 expands the earlier reported beneficial effects[8] of the interplay between behavior elicitation and evaluation to rating accuracy because accuracy was highest in the condition when assessors were made aware of the cues. Conceptually, these results show that assessors can interpret the meaning of detected behaviors more accurately when they are made aware of the situational cues that evoke these behaviors, thereby underscoring the importance of contextualization for accurately drawing inferences about people's characteristics (Christiansen et al., 2005; Gilbert, 1989; Trope, 1986).

General Discussion

This study examined the interplay between behavior elicitation and evaluation. The results obtained across different exercises, settings, and studies add to our conceptual understanding of behavioral assessment and have several implications for improving current AC practice. We summarize these key contributions below.

Interplay Between Behavior Elicitation and Evaluation

We started to integrate the RAM with TAT by proposing to plant multiple job-related stimuli in interpersonal exercises. We hypothesized that such situational stimuli (e.g., in the form of role-player prompts) would enhance the observability, interrater reliability, measurement, and accuracy of dimension ratings. A key finding was that behavior elicitation via situational stimuli affected only the observation process, with assessors detecting more relevant behaviors for three out of four dimensions. Yet, solely increased behavior elicitation (i.e., without making assessors familiar with the cues) did neither affect the correct utilization of behavior nor the reliability or validity of the ratings.

As a second key finding, we discovered that it is of paramount importance to familiarize assessors with the situational stimuli that elicit behavior in order for them to correctly utilize the larger amount of behavioral information and subsequently provide more reliable, distinct, and accurate ratings. Apparently, when assessors are made familiar with the cues that activate specific dimension-related behaviors, these cues structure the rating process and serve as "hints," alerting them to classify behaviors in the correct dimension.

These findings have several theoretical contributions. One conceptual implication is that this study constitutes an important step to integrate the RAM and TAT. In particular, evidence for the interplay between behavior elicitation and evaluation shows that increasing assessees' behavioral manifestations of underlying dimensions and the subsequent assessor processes of evaluating these dimensions represent two sides of the same coin. This bridges the RAM and TAT as it shows that situational strategies for evoking (non)verbal behavior (as highlighted in TAT) should be aligned with the processes by which assessors perceive and interpret these behavioral manifestations (as stressed in the RAM).

Furthermore, the strategies recommended in this study illustrate how careful design might increase the probability to provide "good information" to assessors for rating dimensions. Although the RAM model posits good information to be a moderator of rating accuracy, prior research typically compared different information sources, interactions, and media in terms of the quality of information presented to raters (Funder, 1999). Importantly, this study extends this line of research by specifying concrete approaches for increasing the quality of information in ongoing interactions (Letzring, 2008; Letzring & Human, 2014).

[7] There was no effect of the sequence in which the videotaped performances were shown.

[8] We also ran within-assessee generalizability analyses to conduct additional tests of H6 (discriminant validity) and H7 (inter-rater reliability). Results showed the discriminant validity and inter-rater reliability of assessor ratings to be highest in the condition when behavior evoking cues were used and when assessors were familiar with these cues, further supporting H6 and H7.

Table 7
Accuracy Results Broken Down by Condition in Study 4

	Low behavior elicitation		High behavior elicitation	
	Assessors unfamiliar with cues ($N = 56$)	Assessors familiar with cues ($N = 55$)	Assessors unfamiliar with cues ($N = 56$)	Assessors familiar with cues ($N = 55$)
M	$.26_a$	$.14_a$	$.27_a$	$.61_b$
SD	.51	.44	.50	.73

Note. Means differ from each other (at $p < .01$) when different subscripts are used.

Another conceptual implication is that the use of situational stimuli for evoking behavior implies that the whole exercise is no longer seen as the sole vehicle for evoking behavior. Instead, the exercise situation is broken down and structured in several mini situations, which might facilitate the observation and rating process. So, AC exercises should no longer be conceptualized solely at a holistic (molar) level but also at an elementalistic (molecular) level. By configuring AC exercises as the sum of a series of mini interpersonal situations provided by role-players the exercise is less of a black box (Brummel et al., 2009; Lievens, 2008) because this might facilitate determining which exercise aspects evoke particular dimension-related behaviors.

We believe these implications go beyond role-plays in an AC context. As noted in the beginning of the article, interpersonal exercises are the mainstay of many forms of behavioral assessment outside the employment context (e.g., in the health professions). Our results might also be insightful for predictor instruments such as work samples and employment interviews. For instance, recent research on employment interviews has started to examine the effectiveness of probing (Levashina, Hartwell, Morgeson, & Campion, in press).

Shedding Light into the Black Box of AC Exercises

The lack of research on behavior elicitation via situational stimuli in AC exercises can be broadened to a lack of AC research on the impact of exercise factors in general (Lievens et al., 2009; Schneider & Schmitt, 1992). We know little about how variations in exercise design influence AC behavior and the psychometric properties of the ratings made. This is surprising because decades of research on ACs reveal that exercises explain important portions of variance in assessor ratings (Bowler & Woehr, 2006; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Jackson, Stillman, & Atkins, 2005; Kuncel & Sackett, 2014; Lance, 2008; Monahan, Hoffman, Lance, Jackson, & Foster, in press; Putka & Hoffman, 2013). In addition, there is relative consensus that this substantial exercise variance does not represent measurement bias but true cross-situational variance of assessees across exercises (Gibbons & Rupp, 2009; Lance, 2008; Lievens, 2002; Putka & Hoffman, 2013). In line with the interactionist mixed-model AC approach (Borman, 2012; Lievens & Christiansen, 2012; Melchers et al., 2012), AC exercises are then thought to present different situational demands to candidates, thereby producing variability performance across exercises (see also Gibbons & Rupp, 2009; Howard, 2008; Putka & Hoffman, 2013; Speer et al., 2014).

This study adds new insights to this large literature of exercise effects in ACs. We discovered that making assessors aware that multiple situational stimuli per dimension are built into the exercises increases both the convergent and discriminant validity of their ratings. The use of similar cues by significant others (i.e., role-players) across exercises might have enhanced the behavioral consistency of assessees across exercises (Gibbons & Rupp, 2009; Lievens et al., 2009). This explanation is consistent with broader social psychological research on cross-situational consistency showing that social demands are key psychological contextual features invoking cross-situational (in)consistency in behavior (Mischel & Shoda, 1995, see also Oliver et al., in press). In turn, alerting assessors to these situational cues might make them more aware of these contextual features affecting behavior ("If . . ., then . . ." templates).

Aligning Stimulus Development With Rating Systems

At a practical level, this study provides evidence-based guidelines for improving ACs. In particular, our findings translate to actionable advice for adjusting exercise design, assessor training, and rating instruments. Prior to presenting these practical recommendations, a disclaimer is on target. We do not propose that current exercise design, assessor training, and scoring practices should be abandoned. Rather, we posit that the following recommendations should play a more prominent role in AC development, with the goal of making a good tool even better.

Exercise design. A first practical implication deals with exercise design. Current AC exercise design approaches focus on the whole exercise as a vehicle for evoking candidate behavior. This study suggests that AC developers should plant multiple dimension-related stimuli (e.g., situational cues via role-players) in interpersonal AC exercises as a systematic and efficient tool for increasing the frequency of behavior relevant to focal constructs. Similar to our study, the situational cues can be added to already existing exercises. An even stronger approach consists of building these cues in as part of the exercise development process itself. Note further that the cues should represent stimuli that people might experience in their work tasks and context. Planting such situational cues in existing and/or new exercises should decrease the probability that assessors have to rely on one or two behavioral reactions to rate several dimensions (Brannick et al., 1989). Hereby we emphasize that both role-players and assessors should be familiarized with the situational cues.

To date, the current guidelines on AC operations are silent about the use of situational stimuli such as role-player prompts (International Task Force on Assessment Center Guidelines, 2009). Hence, we suggest adding this exercise design consideration to the next version of the guidelines. Regardless of how evoked behavior

is subsequently evaluated by assessors (dimension-based, task-based, or mixed-model ACs), eliciting and observing behavior is key to the provision of rich feedback.

Assessor training. Second, we believe that the incorporation of situational stimuli in ACs also adds a new angle to assessor training. In current assessor training, the focus is placed on imposing a consistent frame-of-reference on assessors (Lievens, 2001; Schleicher et al., 2002). In such training programs, the dimensions and the accompanying behaviors play a crucial role. Although this focus has proven its merits, this study shows that it is also important to make assessors familiar with the specific context (situational stimuli) in which specific behavior is activated.

Rating instruments. A third implication pertains to adjustments to the rating tools. Although behavioral checklists and BARS are routinely used in AC practice, this study suggests it might be worthwhile to include the situational cues in these rating instruments. This easy to implement and cost-effective approach might provide guidance to assessors (especially inexperienced ones) in that it helps them to better detect, utilize, and rate the evoked behaviors in the stream of candidate behavior. Some consultancy firms are already adopting this approach[9] of making ratings at the behavior level in the rating instruments (via probes introduced as cues in the exercises). So, investigating whether including the cues in the rating tools further increases reliability and validity represents a next logical step in this research. When empirical evidence becomes available for its effectiveness this practice should become widespread. Note that incorporating the situational cues directly into the rating instruments is not only applicable to AC exercises but also to work samples, employment interviews, and so forth.

In short, as a common thread running through these practical suggestions, we suggest there should be a better alignment between stimulus development (behavior elicitation) and the rating system (behavioral evaluation) in ACs (see Brannick, 2008). Metaphorically, one can compare the interplay between behavior elicitation and evaluation with a tango. Role-players take the first step by using a cue for evoking a specific dimension. Next, assessors follow by concentrating on the candidate reaction. This process between role-players and assessors is repeated multiple times per dimension in an exercise.

Limitations

In this study, situational stimuli for activating behavior were developed only for role-plays and oral presentations. Hence, all stimuli reflect a person-based approach for eliciting behavior (i.e., stimuli given by role-players such as Buster and Kippy in the quote above). Although it should also be possible to build task-based stimuli in individual exercises (e.g., exercise instructions, video and audio stimuli in computerized simulations), so far little is known about the effects of such approaches on behavior elicitation and scoring.

This study also found only evidence for the effectiveness of situational cues for (inter)personal dimensions. Including situational stimuli for activating behavior for cognitively oriented dimensions (i.e., problem solving) did not produce the expected effects in terms of observability. As noted above, future research should examine this issue more closely by framing it into the difference between different types of dimensions (social, etc.).

Directions for Future Research

Apart from the ideas for future research already mentioned throughout the Discussion, we propose the following avenues for future research. First, situational cues (stimuli, prompts, probes) are central in the interplay between behavior elicitation and evaluation. One area of research should focus on making finer distinctions between the type of situational cues. For instance, using the concept of situational strength one might divide the exercise stimuli into activation free (e.g., filler information), activation-light ("weak" cues), and activation-heavy ("strong" cues) information. The strength of the cues is among others important in light of the transparency of the AC dimensions and the ability in which candidates might identify what is being measured (Jansen et al., 2013; Kleinmann et al., 2011). Future studies should test the effects of the provision of situational cues on the transparency of ACs. An equally important issue deals with the number of situational cues to be provided. In this study, we typically used three to four situational cues for eliciting behavior related to a given dimension. However, we agree that this is a rule-of-thumb. Therefore, future research should examine which number of situational cues is ideal in terms of being sufficient to reliably elicit behavior while at the same time not overwhelming assessees and assessors.

The level of generalizability of the cues across AC exercises represents another useful criterion for distinguishing among the cues. As shown in the tables in Appendix A, some situational cues can be used across exercises, whereas others are more specific to the exercise at hand. As noted before, these differences might have implications for the situational breadth of the AC exercises (the extent to which AC exercises present different situational demands to candidates, see Speer et al., 2014), the variability in candidate performance across these exercises, and the convergent validity of the AC ratings. Along these lines, a key principle of TAT is that cues that might be superficially different still activate behavior related to the same trait (Haaland & Christiansen, 2002). In any case, this constitutes another important area for future research.

Second, we posit it is important to examine whether the provision of situational stimuli changes the type of performance being assessed. That is, does the performance shown change from reflecting typical to maximum performance? The provision of cues might make the situation stronger, and suggest to assessees what they should do, rather than allow them to choose what to do (see Smith-Jentsch, 2007). Interestingly, the effects might differ depending on the type of dimension (personality-like dimensions vs. ability-like dimensions). For example, consider a role-play with a problem subordinate. Without cues, the candidate might (or might not) engage in coaching behaviors. Conversely, with cues from the role-player (e.g., "Well, what can you do to help me with my problem?"), the candidate might 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, in this case 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, &

[9] We thank an anonymous reviewer for mentioning this.

Grubb, 2007). In the same role-play, problem analysis/solving might also be measured. Without the inclusion of cues, the candidate might engage in little systematic problem analysis. However, if the role-player were to provide cues such as "What led you to say you would do?" or "What other solutions did you consider?", it is possible that the candidate gives answers that (s)he did not initially considered. Thus, one could evaluate the ability to do problem solving because more of those behaviors would be evoked. In summary, to measure "personality-like" dimension, providing cues might make the situation stronger and reduce individual differences in the behavior one wants to observe. In case of "ability-like" dimensions, providing cues may ensure that relevant behaviors are displayed, and thus enhance measurement accuracy.

Third, an important implication of these issues it that in case spontaneously bringing up information represents criterion-relevant variance, planting situational cues for eliciting that information might reduce the criterion-related validity of AC ratings. However, there are also arguments that suggest that the inclusion of situational cues might exert beneficial effects on validity. The inclusion of specific exercise stimuli that trigger job relevant behavior might increase the point-to-point correspondence with the criterion dimensions. In addition, in this study accuracy was higher when cues were used and assessors were familiar with them. To disentangle these rival assumptions behind the effects of planting situational cues in AC exercises on validity, future studies might include the level of ambiguity in the job and the purpose of the AC as important factors.

Conclusion

This study presents an integrative framework that simultaneously considered behavior elicitation and assessor rating issues. Both experimental and field studies demonstrate the importance of the interplay between behavior elicitation and evaluation via situational cues in order to improve the quality of AC ratings. At a practical level, this study's recommendation for better aligning stimulus development with rating systems has implications for the design of exercises, assessor training, and rating instruments. These theoretical and practical implications should inspire both researchers and practitioners to work together in developing theory-driven strategies that further improve the domain of behavioral assessment.

References

Adamo, G. (2003). Simulated and standardized patients in OSCEs: Achievements and challenges 1992–2003. *Medical Teacher, 25,* 262–270. http://dx.doi.org/10.1080/0142159031000100300

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716. http://dx.doi.org/10.1111/j.1744-6570.1990.tb00679.x

Assessment Staff, O. S. S. (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services*. New York, NY: Rinehart & Co.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238–252. http://dx.doi.org/10.1016/0030-5073(77)90004-6

Borman, W. C. (2012). Dimensions, task and mixed models: An analysis of the three diverse perspectives on assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 309–320). New York, NY: Routledge.

Boulet, J. R., Smee, S. M., Dillon, G. F., & Gimpel, J. R. (2009). The use of standardized patient assessments for certification and licensure decisions. *Simulation in Healthcare, 4,* 35–42. http://dx.doi.org/10.1097/SIH.0b013e318182fc6c

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91,* 1114–1124. http://dx.doi.org/10.1037/0021-9010.91.5.1114

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 131–133. http://dx.doi.org/10.1111/j.1754-9434.2007.00025.x

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74,* 957–963. http://dx.doi.org/10.1037/0021-9010.74.6.957

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-3456-0

Brummel, B. J., Rupp, D. E., & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62,* 137–170. http://dx.doi.org/10.1111/j.1744-6570.2008.01132.x

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72,* 463–474. http://dx.doi.org/10.1037/0021-9010.72.3.463

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). New York, NY: Routledge.

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance, 18,* 123–149. http://dx.doi.org/10.1207/s15327043hup1802_2

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. http://dx.doi.org/10.1177/001316446002000104

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

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579. http://dx.doi.org/10.1037/0021-9010.80.5.565

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin, 83,* 956–974. http://dx.doi.org/10.1037/0033-2909.83.5.956

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. http://dx.doi.org/10.1037/0022-3514.37.7.1097

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670. http://dx.doi.org/10.1037/0033-295X.102.4.652

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego, CA: Academic Press.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21,* 177–182. http://dx.doi.org/10.1177/0963721412445309

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Per-*

sonality and Social Psychology, 60, 773–794. http://dx.doi.org/10.1037/0022-3514.60.5.773

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74,* 611–618. http://dx.doi.org/10.1037/0021-9010.74.4.611

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management, 35,* 1154–1180. http://dx.doi.org/10.1177/0149206308328504

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). New York, NY: Guilford Press.

Greenberg, J., & Tomlinson, E. C. (2004). Situated experiments in organizations: Transplanting the lab to the field. *Journal of Management, 30,* 703–724. http://dx.doi.org/10.1016/j.jm.2003.11.001

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55,* 137–163. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00106.x

Harvey, P. D., Velligan, D. I., & Bellack, A. S. (2007). Performance-based measures of functional skills: Usefulness in clinical treatment studies. *Schizophrenia Bulletin, 33,* 1138–1148. http://dx.doi.org/10.1093/schbul/sbm040

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64,* 351–395. http://dx.doi.org/10.1111/j.1744-6570.2011.01213.x

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 98–104. http://dx.doi.org/10.1111/j.1754-9434.2007.00018.x

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21,* 264–276. http://dx.doi.org/10.1111/ijsa.12036

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17,* 243–253. http://dx.doi.org/10.1111/j.1468-2389.2009.00467.x

Jackson, D. J. R., Ahmad, M. H., Grace, G., & Yoon, J. (2011). An alternative take on AC research and practice: Task-based assessment centers. In N. Povah & G. C. Thornton III (Eds.), *Assessment centres and global talent management* (pp. 33–46). Surrey, UK: Gower.

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18,* 213–241. http://dx.doi.org/10.1207/s15327043hup1803_2

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology, 98,* 326–341. http://dx.doi.org/10.1037/a0031257

Kauffman, J. R., Jex, S. M., Love, K. G., & Libkuman, T. M. (1993). The construct validity of assessment centre performance dimensions. *International Journal of Selection and Assessment, 1,* 213–223. http://dx.doi.org/10.1111/j.1468-2389.1993.tb00115.x

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1,* 128–146. http://dx.doi.org/10.1177/2041386610387000

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology, 99,* 38–47. http://dx.doi.org/10.1037/a0034147

Lance, C. E. (2008). Why assessment centers (ACs) don't work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 84–97. http://dx.doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89,* 22–35. http://dx.doi.org/10.1037/0021-9010.89.1.22

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89,* 377–385. http://dx.doi.org/10.1037/0021-9010.89.2.377

Lane, S., & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Ed.), *Educational measurement* (pp. 387–424). Westport, CT: Praeger.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42,* 914–932. http://dx.doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D., & Human, L. J. (2014). An examination of information quality as a moderator of accurate personality judgment. *Journal of Personality, 82,* 440–451. http://dx.doi.org/10.1111/jopy.12075

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (in press). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152. http://dx.doi.org/10.1111/1468-2389.00085

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86,* 255–264. http://dx.doi.org/10.1037/0021-9010.86.2.255

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87,* 675–686. http://dx.doi.org/10.1037/0021-9010.87.4.675

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 112–115. http://dx.doi.org/10.1111/j.1754-9434.2007.00020.x

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247–258. http://dx.doi.org/10.1037/0021-9010.91.2.247

Lievens, F., & Christiansen, N. D. (2012). Core debates in assessment center research: Dimensions versus exercises. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 68–91). New York, NY: Routledge.

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

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Bingley, UK: JAI Press.

McClelland, D. C. (1984). *Motives, personality, and society: Selected papers*. New York, NY: Praeger.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91. http://dx.doi.org/10.1111/j.1744-6570.2007.00065.x

McFarland, L. A., Yun, G. J., Harold, C. M., Viera, L., & Moore, L. G. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology, 58,* 949–980. http://dx.doi.org/10.1111/j.1744-6570.2005.00374.x

Melchers, K. G., Wirz, A., & Kleinmann, M. (2012). Dimensions AND exercises: Theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 237–254). New York, NY: Routledge.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36,* 121–140. http://dx.doi.org/10.1177/0149206309349309

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252–283. http://dx.doi.org/10.1037/h0035002

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246–268. http://dx.doi.org/10.1037/0033-295X.102.2.246

Monahan, E., Hoffman, B. J., Lance, C. E., Jackson, D. J. R., & Foster, M. R. (in press). Now you see them, now you don't: The influence of indicator-factor ratio on support for assessment center dimensions. *Personnel Psychology*.

Oliver, T., Hausdorf, P., Lievens, F., & Conlon, P. (in press). Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management*.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education - The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57,* 22–36. http://dx.doi.org/10.1177/0022487105284045

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology, 98,* 114–133. http://dx.doi.org/10.1037/a0030887

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct-validity of assessment center dimensions. *Personnel Psychology, 43,* 71–84. http://dx.doi.org/10.1111/j.1744-6570.1990.tb02006.x

Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review, 12,* 311–329. http://dx.doi.org/10.1177/1088868308321721

Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. Hakel (Ed.) *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 113–129). Hillsdale, NJ: Lawrence Erlbaum Inc.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87,* 735–746. http://dx.doi.org/10.1037/0021-9010.87.4.735

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment-center dimension and exercise constructs. *Journal of Applied Psychology, 77,* 32–41. http://dx.doi.org/10.1037/0021-9010.77.1.32

Schollaert, E., & Lievens, F. (2011). The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment, 19,* 190–197. http://dx.doi.org/10.1111/j.1468-2389.2011.00546.x

Schollaert, E., & Lievens, F. (2012). Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance, 25,* 255–271. http://dx.doi.org/10.1080/08959285.2012.683907

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components* (2nd ed.). New York, NY: Wiley.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions transparent on relations with typical performance predictors. *Human Performance, 20,* 187–203. http://dx.doi.org/10.1080/08959280701332992

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology, 99,* 282–295. http://dx.doi.org/10.1037/a0035213

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497–506. http://dx.doi.org/10.1037/0021-9010.73.3.497

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500–517. http://dx.doi.org/10.1037/0021-9010.88.3.500

Thornton, G. C., III, & Cleveland, J. N. (1990). Developing managerial talent through simulation. *American Psychologist, 45,* 190–199.

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

Tippins, N., & Adler, S. (2011). *Technology-enhanced assessment of talent*. San Francisco, CA: Jossey-Bass. http://dx.doi.org/10.1002/9781118256022

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93,* 239–257. http://dx.doi.org/10.1037/0033-295X.93.3.239

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29,* 231–258. http://dx.doi.org/10.1177/014920630302900206

Woo, S. E., Sims, C. S., Rupp, D. E., & Gibbons, A. M. (2008). Development engagement within and following developmental assessment centers: Considering feedback favorability and self-assessor agreement. *Personnel Psychology, 61,* 727–759. http://dx.doi.org/10.1111/j.1744-6570.2008.00129.x

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

(*Appendix follows*)

Appendix

Situational Cues Used in Assessment Center Exercises

Table A1
Situational Cues Used in Role-Play Exercise

Problem solving
"Where did you get this information?"
"How did you find out?"
Answering questions in a vague way
"Why do you think I feel bad?"
"What is the main problem/solution?"
Interpersonal sensitivity
"I do not want to make myself look ridiculous to the clients."
"I feel offended by the fact that I had to come here."
"Do you still trust me?"
"Actually, I prefer not to do this."
"Our conversation makes me feel uncomfortable as I get the impression that the colleagues gossip about me."
"You are partially right."
Planning and organizing
"I do not have plenty of time, what is the agenda of the meeting? What do you want to discuss?"
"Is your proposition realistic in terms of time? I have a very busy schedule the next weeks."
"What do you expect from me in the next period?"
"What is the top priority?"
"How do we organize this?"
"Can you explain it in more details? Can you give the facts and figures of the plan?"
Tolerance for stress
"I also heard some complaints about you from other colleagues."
"You are not perfect either."
"No, I refuse to do that."
Shaking one's head (repeatedly)

Table A2
Situational Cues Used in Oral Presentation Exercise

Problem solving
"What is the main problem/solution?"
"What is the essence of all this?"
"Do you find some indications in the file for your ideas?"
"I heard some contradictions; can you explain the following, for instance, . . .?"
"What is your most important recommendation?"
Interpersonal sensitivity
"Excuse me to interrupt you, but your words make me feel uncomfortable."
"I thought that this company was doing a great job."
"This issue is delicate around here."
"Your idea touches on a sensitive issue; we put a lot of effort in this."
"You are partially right."
Planning and organizing
"Your presentation is herky-jerky, can you help me out?"
"How will we allocate the budget?"
"What is the timing for this?"
"It will be difficult to convince the employees of this timing because we are then in the holiday period."
"What are your concrete actions; how will you realize the implementation?"
Tolerance for stress
"I also heard complaints about your company."
"This idea is awful."
"We have new information that you have not received yet."

Received March 6, 2014
Revision received September 17, 2014
Accepted September 17, 2014 ■