1-2012

Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior?

Eveline SCHOLLAERT
Ghent University

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

DOI: https://doi.org/10.1080/08959285.2012.683907

Citation

SCHOLLAERT, Eveline and LIEVENS, Filip. Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior?. (2012). *Human Performance*. 25, (3), 255-271. Research Collection Lee Kong Chian School Of Business.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/5518

Published in Human Performance, 2012 January,
Volume 25, Issue 3, Pages 255-271
http://doi.org/10.1080/08959285.2012.683907

Building Situational Stimuli in Assessment Center Exercises: Do Specific Exercise Instructions and Role-Player Prompts Increase the Observability of Behavior?

Eveline Schollaert and Filip Lievens
Ghent University

Little is known about how assessment center exercises might be designed to better elicit job-relevant behavior. This study uses trait activation theory as a theoretical lens for increasing the number of behaviors that can be observed in assessment centers. Two standardized exercise stimuli (specific exercise instructions and role-player prompts) are proposed, and their effects on the observability of candidate behavior are examined. Results showed a significant effect of role-player prompts in increasing both the general number of behavioral observations and the number of behavioral observations related to three out of four dimensions. Specific exercise instructions did not have effects on observability. Implications for trait activation theory and assessment center practice are discussed.

> “Behaviors, not exercises, are the currency of assessment centers. Exercises are merely the stimuli to elicit behaviors”
>
> — Howard (2008, p. 101)

This key focus on behavior in assessment centers (ACs) is also well reflected in the most recent Guidelines of Assessment Center Operations, in which the observation of overt behavioral responses is described as a necessary and fundamental component of ACs (International Task Force on Assessment Center Guidelines, 2009). The Guidelines further state that AC designers should attempt to design exercises that evoke a large number of job-related behaviors because this gives assessors enough opportunities to observe this behavior (International Task Force on Assessment Center Guidelines, 2009). Observing a substantial amount of job-related behavior in ACs is crucial for rating and developmental purposes as behavioral observations serve as a basis for providing participants with detailed developmental feedback about their strengths and weaknesses. The more behavioral examples one can give, the more useful the feedback might be (Woo, Sims, Rupp, & Gibbons, 2008).

Although the AC guidelines and the AC literature emphasize the importance of exercises providing sufficient opportunities for observing job-related behavior, little is known about concrete and effective approaches one might implement for increasing the observability of behavior in AC

Correspondence should be sent to Eveline Schollaert, Ghent University, Department of Personnel Management, Work and Organizational Psychology, Henri Dunantlaan 2, B-9000 Ghent, Belgium. E-mail: schollaerteveline@hotmail.com

exercises. The traditional paradigm focuses on the exercise as a whole. In this molar approach, the entire exercise is seen as a vehicle for evoking behavior (Howard, 2008; Lievens, Tett, & Schleicher, 2009; McFarland, Yun, Harold, Viera, & Moore, 2005). However, research has shown that this molar exercise approach might be sometimes problematic in that an insufficient number of behaviors is elicited. Bycio, Alvares, and Hahn (1987) even suggested that "assessors within an exercise are sometimes, if not usually, forced to base *all* of their judgments on four or five behaviors" (p. 472). Along these lines, other authors mentioned the occurrence of so-called red hot items. This refers to the fact that assessors often need to rely on one particular behavioral reaction to score candidates on several dimensions (Brannick, Michaels, & Baker, 1989).

To address this problem, our study builds on trait activation theory (TAT) to propose two molecular and potentially practical approaches of increasing the observability of job-related behavior in AC exercises. We propose planting multiple stimuli *within* exercises as a structured means of eliciting candidate behavior. Specifically, we aim to examine the effects of (a) specific exercise instructions and (b) role-player prompts on the observability of behavior.

This research is of both conceptual and practical importance. Conceptually, this study represents the first application of TAT as a useful prescriptive framework in AC design. At a practical level, the present work introduces two tools that can help AC designers to develop AC exercises that elicit a large number of behavioral observations. It should also be noted that strategies for eliciting and observing behavior within AC exercises are beneficial for traditional dimension-based ACs (wherein assessors rate performance on dimensions within exercises) as well as for task-based ACs (wherein assessors rate general exercise performance; Jackson, Stillman, & Atkins, 2005) as the observation of sufficient behavior plays a pivotal role in both types of ACs.

BASIC TENETS OF TAT

It is generally acknowledged that candidate behavior in ACs is solely determined not by dispositional factors (i.e., stable personal characteristics of candidates) or by situational factors (i.e., AC exercises) but by the interaction of the person and the situation. Therefore, it is relevant to conceptualize the occurrence of candidate behavior in ACs in terms of a recent interactionist theory such as TAT (Lievens et al., 2009; Tett & Burnett, 2003). TAT focuses on the person–situation interaction to explain behavior based on responses to trait relevant cues found in situations (Tett & Guterman, 2000). These observable responses subsequently serve as the basis for ratings made by assessors in ACs (Tett & Burnett, 2003).

According to TAT, two factors are important for understanding in which situations a trait is likely to manifest itself in behavior. First, TAT emphasizes the importance of *situation trait relevance*. A situation is considered relevant to a trait if it provides cues for the expression of trait relevant behavior (Tett & Guterman, 2000). Thus, situation trait relevance is a qualitative feature of situations that is essentially trait specific; it is informative with regard to which cues are present to elicit behavior for a given latent trait. *Situation strength* is the second relevant factor from the trait activation perspective. Situation strength is more of a continuum that refers to the clarity with regard to how the situation is perceived. Strong situations involve unambiguous behavioral demands and are therefore likely to negate almost all individual differences in behavior without regard to any specific trait (see also Bem & Allen, 1974; Meyer, Dalal, & Hermida, 2010;

Mischel, 1973). Conversely, weak situations are characterized by more ambiguous expectations, enabling more variability in behavioral responses to be observed.

So far, two studies (Haaland & Christiansen, 2002; Lievens, Chasteen, Day, & Christiansen, 2006) have used TAT in the context of ACs. Both studies were framed in the construct-related validity debate on AC ratings and showed that poor convergence of AC ratings across exercises was due to correlating ratings from exercises that differed in trait activation potential. In other words, the poor convergence was attributed to a substantial discrepancy in the degree to which exercises offered the possibility to observe differences in trait-relevant behavior. A drawback of these two prior studies is that they evaluated AC exercises in an existing operational AC. There was no manipulation of the trait activation potential in different exercises. Thus, so far, there have been no tests of the actual implementation of TAT in AC exercise design.

The key focus of the present study is on the observability of behavior. We examine the impact of the use of exercise stimuli on the opportunity for assessors to observe behavior and to note down a higher number of behavioral (also known as "good") observations. A behavioral or good observation is a behavioral statement that specifically describes what a person says or does (Gaugler & Thornton, 1989). The following section discusses the two exercise stimuli and their hypothesized effects on observability.

STRATEGIES FOR BUILDING STIMULI WITHIN AC EXERCISES

AC Exercises and Situational Stimuli

In current AC practice, exercises are developed with two goals in mind. One goal is to increase fidelity for purposes of maximizing criterion-related validity (Ahmed, Payne, & Whiddett, 1997; Thornton & Mueller-Hanson, 2004). That is, exercises are developed to represent the most important task, social, and organizational demands of the target job. Another goal of exercises consists of eliciting job-related behavior as indicative of specific dimensions (Howard, 2008; McFarland et al., 2005). For instance, as shown by dimension-exercise matrices of operational ACs, a cooperative leaderless group discussion is typically seen as a way of activating leadership emergence and interpersonal competencies, whereas a presentation exercise is expected to trigger dimensions relating to Emotional Stability and communication.

The problem inherent in such an approach to exercise development is that an AC exercise largely remains a black box (Howard, 2008). As noted by Brummel, Rupp, and Spain (2009), stimuli in AC exercises are complex, making it difficult to ascertain which specific aspects of the exercise map onto which dimensions. As exercises provide freedom and latitude to candidates to act on the situational stimuli included, the process and outcome of the exercise might also differ across candidates. To address this problem, we propose that the exercise–behavior linkage be examined at a more molecular level. Essentially, this implies building various stimuli in each exercise. With this respect, Brannick (2008) cogently argued to implement and rate multiple job-related items or problems within the exercise. Similarly, Howard (2008) posited that

> Designers should construct specific stimuli to elicit the kinds of behaviors to be measured and guide assessors to their placement and relative importance. . . . Designers must do more than create work samples; they must develop simulations that will best elicit the desired behaviors. We need to develop

> a much better understanding of the kinds of AC challenges that will bring out the behaviors associated with current and evolving positions and diverse business challenges. (p. 101)

Therefore, our study aims to include situational stimuli for evoking behavior within AC exercises. When interpersonal exercises (e.g., role-plays and oral presentations) are used, *role-player prompts* might serve as a first strategy for eliciting job-related behavior. In current AC practice, role-player training mainly focuses on providing role-players with information concerning their role. They learn to play their role objectively and consistently (see International Task Force on Assessment Center Guidelines, 2009). However, the guidelines do not contain information of *how* the role-player should be consistent. Moreover, empirical research in this area is also virtually nonexistent. We suggest that the use of prompts might serve as a practical means for role-players to evoke more job-related behavior from candidates. Role-player prompts are defined as predetermined verbal and nonverbal cues that a role-player consistently provides during the AC exercise across candidates to elicit job-related behavior (Schollaert & Lievens, 2011). An example is that the role-player asks what is top priority to evoke behavior related to planning and organizing.

A second way of eliciting job-related behavior is through *specific exercise instructions*. To date, we know little about how exercise instruction variations might affect the behavior demonstrated in ACs. However, prior research in the situational judgment test (SJT) domain has shown the importance of response instructions by suggesting that the use of different response instructions for identical SJT items might have important consequences on SJT performance (McDaniel, Hartman, Whetzel, & Grubb, 2007). Currently, exercise instructions in ACs provide background information about the fictitious organization, key persons, and problems. The instructions might also evoke some general expectations to candidates about what behavior (not) to show. With this respect, some authors posited that specific exercise instructions provided to candidates at the beginning of AC exercises might be deliberately constructed to serve as mechanisms for eliciting job-related behavior (Thornton & Mueller-Hanson, 2004). Such exercise instructions are defined as predetermined specific instructions given to candidates at the beginning of AC exercises to elicit job-related behavior. An example is the specific instruction to candidates to consider the opinion and the feelings of the role-player (to increase behavior related to interpersonal sensitivity).

Hypotheses

Conceptually, there is a clear link between TAT and the inclusion of situational stimuli in AC exercises. As previously noted, two main factors determine the activation of behavior related to traits (AC dimensions). The first factor underlying trait activation relates to situational trait relevance, which might be increased by enhancing the cues presented to elicit behavior for a given latent trait. We expect that this is accomplished by both role-player prompts and exercise instructions. In fact, the use of role-player prompts creates multiple mini situations for evoking job-related behavior within exercises. So, basically role-player prompts serve as vehicles for increasing the opportunity for candidates to demonstrate behavior. In turn, this might have beneficial effects on the opportunity for assessors to observe behavior and to note down a higher number of behavioral observations. The same is true for exercise instructions. As previously noted, specific exercise instructions at the beginning of AC exercises might reveal to candidates expectations about what behavior to show in the exercises. As such instructions trigger relevant candidate behavior we expect that the amount of observed candidate behavior will be enhanced and that assessors will have the opportunity to note down more behavioral observations. We anticipate these positive effects of increasing the situational trait relevance via role-player prompts and exercise instructions both at the overall level (i.e., the overall number of behavioral observations noted down by assessors) and at the level of dimensions (i.e., the number of behavioral observations per dimension noted down by assessors). Taken together, this leads to the following hypotheses:

H1a: The frequency of observed candidate behavior (i.e., the number of behavioral observations) will be higher in AC exercises designed to evoke behavior by implementing role-player prompts than in AC exercises without role-player prompts.

H1b: The frequency of observed dimension-related behavior (number of behavioral observations per dimension) will be higher in AC exercises designed to evoke behavior by implementing role-player prompts than in AC exercises without role-player prompts.

H2a: The frequency of observed candidate behavior (the number of behavioral observations) will be higher in AC exercises designed to evoke behavior by implementing specific exercise instructions than in AC exercises without specific exercise instructions.

H2b: The frequency of observed dimension-related behavior (the number of behavioral observations per dimension) will be higher in AC exercises designed to evoke behavior by implementing specific exercise instructions than in AC exercises without specific exercise instructions.

As stated previously, both role-player prompts and exercise instructions might be effective ways to increase the situational trait relevance of AC exercises. However, situational trait relevance is only one side of the equation of activation trait-related behavior. According to TAT, situational strength represents the second factor determining trait activation potential. As previously noted, situational strength represents a continuum. Across the two situational stimuli (i.e., prompts and instructions), the provision of stimuli needs to be explicit enough to activate candidates' propensities while subtle enough to avoid presenting candidates with too strong of a situation (in terms of behavioral demands). There are conceptual reasons to expect that role-player prompts might create somewhat stronger situational demands than exercise instructions. In this study, role-players provide cues to candidates to elicit behaviors related to specific dimensions (e.g., To evoke interpersonal sensitivity mentioning "Our conversation makes me feel uncomfortable"). Role-player prompts are interactive in that the candidate is expected to respond to them. As the prompts used by role-players constitute a person-based means of eliciting behavior (Lievens et al., 2009; Schollaert & Lievens, 2011), role-players might also try repeated times during the exercise to evoke behavior related to the dimensions targeted. Conversely, exercise instructions represent a task-based means of eliciting behavior. Another difference with prompts is that exercise instructions (e.g., "Create a concrete solution and finish the conversation with some clear agreements," to evoke planning and organizing) are provided before the start of the exercise (together with other exercise materials such as the exercise description, context, etc.). For these reasons we expect that in this study role-player prompts might represent stronger situations than exercise instructions. As the activation of trait-related behavior is determined by both the situational trait relevance and the situational strength, this leads to the following hypotheses:

- H3a: The frequency of observed candidate behavior (the number of behavioral observations) will be higher in AC exercises with role-player prompts than in AC exercises with specific exercise instructions.
- H3b: The frequency of observed dimension-related behavior (the number of behavioral observations per dimension) will be higher in AC exercises with role-player prompts than in AC exercises with specific exercise instructions.

METHOD

Sample and Procedure

Data were collected from 103 AC candidates. They were final-year students of a large university (57.3% female, *M* age = 23.0, *SD* = 2.0). Most of the participants had majors in Law and Sciences. Individuals were recruited by an invitation e-mail to participate in a simulated selection process. This simulated selection setting lasted approximately 1 day. Candidates were tested in groups of 3 or 4. We tried to simulate an actual selection situation. First, each candidate was individually welcomed and guided during the day by a certified test administrator. At the start of the session, participants learned that they could increase their experience with a selection process. Subsequently, each candidate had to complete a variety of questionnaires and took part in a real AC exercise. After scrutinizing the catalogues of several consultancy firms, we purchased a specific role-play that was targeted to applicants who pursued an entry-level managerial job. The dimensions measured were interpersonal sensitivity, planning and organizing, problem solving, and tolerance for stress as these were identified as relevant for entry-level managerial jobs. Some weeks after the selection process, candidates received a detailed feedback report.

There was anecdotal evidence that participants were motivated and perceived the selection simulation in a similar way as real selection settings. For instance, they decided themselves to take part in the simulated selection setting. In addition, they all reported to be nervous and anxious to take the tests. They also wore business attire. To measure their motivation, candidates were asked to complete a test motivation scale (of the Test Attitude Survey; Arvey, Strickland, Drauden, & Martin, 1990). This scale was administered after the AC exercise and consisted of five items (e.g., “Doing well on this exercise was important to me”). Candidates were asked to indicate how accurately each statement described them, using a Likert-type scale ranging from 1 (*very inaccurate*) to 5 (*very accurate*). The internal consistency reliability alpha was .82 and the mean score was 3.8 (*SD* = .53).

Experimental Design

A 2 (specific exercise instruction) × 2 (prompt training) between-subjects design was used. The first factor had two levels, either general exercise instructions or specific exercise instructions. The second factor had also two levels. Half of the candidates were confronted with role-players who attended a role-player training without prompts, whereas the other half was confronted with role-players who attended a prompt training. Candidates were randomly assigned to one of the four conditions.

Role-Player Prompts

Generation of Role-Player Prompts

To generate relevant role-player prompts, we conducted a prestudy to develop role-player prompts linked to the four relevant dimensions. First, to ensure a collection of prompts that were actually used in AC practice seven experienced assessors (*M* age = 38.6, *SD* = 7.87, 57% male, *M* experience in selection = 13.3, *SD* = 8.80) were asked to report possible prompts that role-players could use to evoke job-related behavior in the role-play. During this phase, 198 unique prompts were reported. Second, we refined this list by reducing prompts that were (a) inappropriate, (b) too vague (c), too concrete, and (d) redundant. After this procedure, only 84 prompts were left. Third, these 84 prompts were presented to two other groups of assessors: eight graduate students in I-O Psychology (62% male, *M* age = 27.1, *SD* = 2.17) and 12 actual experienced assessors (42% male, *M* selection experience = 5.71, *SD* = 7.35). They were asked to retranslate the prompts to the dimensions. If there was an agreement of at least 70% we considered the prompt to be a good cue to evoke the respective dimension. This further reduced the number of prompts to 21 (Appendix A).

Role-Player Training

Nineteen role-players (58% female; *M* age = 22.9, *SD* = 1.5) were randomly assigned to one of two conditions (role-player training without prompts and role-player training with prompts). The trainer was a consultant with a degree in I-O Psychology and disposed of 15 years of personnel selection experience. Both trainings took half a day and had an identical format. The first 1.5 hr consisted of a lecture wherein participants learned the AC exercise content and their specific role. Next, a videotape of role-player models was presented (1.5 hr). In the role-player training without prompts, the trainer introduced the videotape (of a role-player without prompts) by explaining that role-players have to play their role objectively and consistently, following the Guidelines and Ethical Considerations for AC operations (see International Task Force on Assessment Center Guidelines, 2009). In the prompt training, the trainer also introduced the video (of a role-player using prompts) by explaining that role-players play their role objectively and consistently, but added to this a demonstration of using standardized role-player prompts to evoke behavior. Role-players were taught to use a script, knowing that they had to use at least two cues per dimension. In both conditions, the last part of the training (2 hr) was composed of practical exercises, feedback, and discussion.

Manipulation Check

To check whether role-players did indeed use prompts after a prompt training, four master students I-O Psychology (100% female; *M* age = 21.8, *SD* = .96) coded the role-player behavior.

To this end, the coders received a half day of training. They independently wrote down the verbatim behavior of the role-players. Next, they counted the number of times a role-player used prompts to evoke dimension-related behavior. They also counted the number of interventions that could not be considered to be evoking dimension-related behavior. Interrater agreement (κs > .70) was satisfactory for all dimensions. Discrepancies were resolved through discussion. The role-players that had followed a prompt training used significantly more prompts. In the role-play, the average proportion of prompts was .07 (1.54 prompts) in the condition without prompts and .48 (13.12) in the condition with prompts. Effect sizes were large, varying from 1.43 to 1.69. So we concluded that the role-players were able to use the prompts taught in the training. The frequency with which prompts were displayed can be found in Appendix A.

Specific Exercise Instructions

Generation of Exercise Instructions

In the condition without specific exercise instructions, we used the AC exercise exactly as we purchased it. The general exercise instructions were as follows. First, the background of the role of the candidate was given (information about family, job, interests). Next, the problem was explained (the performance of one of her or his subordinates declined and the colleagues and customers complained about the subordinate). Finally, the candidate received the general instruction to have a conversation with the problem subordinate.

In the condition with specific exercise instructions, the candidates received the same general exercise instructions added with four specific exercise instructions. To this end, we developed one exercise instruction per dimension (Appendix B). For the generation of the four specific exercise instructions we relied on the definitions of the dimensions used by Arthur, Day, McNelly, and Edens (2003).

Manipulation Check

To check whether the specific exercise instructions were relevant instructions for the respective dimensions, 15 graduate students in I-O Psychology (53% female; *M* age = 25.4, *SD* = 3.7) were asked to independently retranslate the specific exercise instructions by assigning exercise instructions to one of the four dimensions. The graduate students had been in college for 5 years and were unaware of the study's purpose. They received the definitions of the dimensions. Besides the specific exercise instructions developed for this study, we also added four exercise instructions related to other dimensions as distracters. For each exercise instruction we found an agreement of at least 90%. So we considered the specific exercise instructions to be relevant instructions to represent the four dimensions.

Observation and Rating Process

Assessors and Assessor Training

I-O Psychology students served as assessors. To this end, they received a standard half day of training. The trainer was a certified assessor with a graduate degree in I-O Psychology. The training program was composed of three main parts: (a) an introduction about the basics of ACs;

(b) a portrayal of the exercise content and the four dimensions; and (c) a workshop on the observation and rating process which included a lecture, practice, and feedback. In the introductory lecture (2.5 hr), ACs were defined and framed in the context of personnel selection. The trainer also discussed the components, purposes, history, and current usage of ACs. The second part (0.5 hr) provided the assessors with knowledge and understanding of the exercise and the four dimensions. To this end, they received background information concerning the exercise and the definitions of the dimensions. We told the assessors that their task consisted of observing candidates and rating them on four dimensions in a role-play. In the last part of the training (1 hr), assessors were taught the process of observing candidate behavior. The trainer instructed them to make behavioral descriptions of the participants' behavior instead of nonbehavioral interpretations. So assessors learned to write down behavioral observations (Gaugler & Thornton, 1989). The assessors completed practical exercises in observing real performances using a videotape with assessees participating in a role-play. They also learned to rate candidates on the dimensions on the basis of frame-of-reference principles. After watching the videotapes the trainer elicited among the assessors a discussion about the observations and ratings. Discrepancies were clarified and the trainer provided the assessors with feedback.

Those trained assessors were randomly assigned to candidates within one of the conditions to keep them blind to the manipulations. While observing a candidate, assessors wrote down the behaviors they noticed. The observation process took about 15 min. After noting down behavior, each assessor was asked to rate the candidate on the four dimensions on a 5-point scale, ranging from *poor* (1) to *excellent* (5).

Measures

The number[1] of behavioral observations written down by assessors served as dependent variable to test our hypotheses. Two independent and trained coders (100% female; *M* age = 22.5, *SD* = .71) with a graduate degree in I-O Psychology examined the individual assessor notes. In a preliminary phase, they independently coded the notes of 20 assessors randomly selected from the assessor pool. They counted the number of behavioral observations. Cohen's (1960) kappa was computed and equaled .92. As this level of interrater agreement among the coders was relatively high and because the observation forms yielded a total of 1,507 notes, it was decided to divide the observation forms in two parts. Each coder was randomly assigned one part of observation forms and coded the notes of these forms.

RESULTS

Means, standard deviations, and effect sizes of the number of behavioral observations broken down by role-player prompts are presented in Table 1 (the correlation matrices are available from

[1] We also conducted the analyses with the proportion of behavioral observations (i.e., the ratio of the number of behavioral observations of an assessor to the total number of observations of the same assessor) as it is possible that assessors who are more prolific produce a higher number of "raw" behavioral observations. The ratio of the number of behavioral observations of an assessor to the total number of observations can be considered as an index of prolificacy (Gaugler & Thornton, 1989). Conducting the analyses with the proportion of behavioral observations showed virtually the same results.

TABLE 1
Means, Standard Deviations, and Effect Sizes of the Number of Behavioral Observations Broken Down by Role-Player Prompts

	With Prompts[a]		*Without Prompts*[b]				
	M	*SD*	*M*	*SD*	*d*	*F*	*p*
Planning and organizing	2.89_b	1.48	1.76_b	1.13	.79	18.567	<.001
Interpersonal sensitivity	4.57_a	2.16	2.86_a	1.54	.83	20.738	<.001
Problem solving	2.85_b	1.45	2.73_a	1.91	.07	.107	.745
Tolerance for stress	2.41_c	1.62	1.45_b	1.12	.65	12.348	.001
Total	12.72	4.46	8.80	3.86	.85	22.562	<.001

Note. $N = 103$. The means and standard deviations refer to the number of behavioral observations. The *d* values are effect sizes calculated by the formula ($M_{\text{with prompts}} - M_{\text{without prompts}}$) / Total *SD*. Positive effect sizes (*d* values) mean that the number of behavioral observations was higher in the condition with prompts than in the condition without prompts. Means with different subscripts (vertical) are significantly different at .05 level.
[a]$n = 54$. [b]$n = 49$.

the first author). The first hypothesis concerned the influence of the use of role-player prompts (H1a) on the observed frequency of behavior during the role-play. We hypothesized that the number of behavioral observations will be higher in AC exercises with role-players who evoked behavior by using prompts than in AC exercises without role-player prompts. We conducted an analysis of variance (ANOVA) with role-player prompts as factor and with the total number of behavioral observations as a dependent variable. This ANOVA showed a significant effect of role-player prompts, $F(1, 101) = 22.562$, $p < .001$ (partial $\eta^2 = .18$). The mean number of behavioral observations was higher in the condition with prompts ($M = 12.72$, $SD = 4.46$) than in the condition without prompts ($M = 8.80$, $SD = 3.86$), supporting H1a. In terms of percentages, the mean number of behavioral observations in the condition with prompts was 44.55% higher than in the condition without prompts.

H1b stated that the use of role-player prompts will have an effect on the observed frequency of dimension-related behavior. A multivariate analysis of variance (MANOVA) with role-player prompts and specific exercise instructions as factors and the number of behavioral observations for problem solving, interpersonal sensitivity, planning and organizing, and tolerance for stress as a set of four dependent variables showed a multivariate main effect of role-player prompts, $F(4, 96) = 8.720$, $p < .001$ (partial $\eta^2 = .27$). Follow-up univariate analyses revealed that the main effect of prompts was significant and in the expected direction for the dimensions planning and organizing ($M_{\text{with prompts}} = 2.89$, $M_{\text{without prompts}} = 1.76$, $d = .79$), interpersonal sensitivity ($M_{\text{with prompts}} = 4.57$, $M_{\text{without prompts}} = 2.86$, $d = .83$), and tolerance for stress ($M_{\text{with prompts}} = 2.41$, $M_{\text{without prompts}} = 1.45$, $d = .65$), but not for problem solving ($M_{\text{with prompts}} = 2.85$, $M_{\text{without prompts}} = 2.73$, $d = .07$). Overall, these results support H1b. We also conducted paired sample *T* tests to examine whether there were differences between the mean numbers of behavioral observations across the dimensions in the role-player prompt condition. Assessors noted significantly more behavioral observations for interpersonal sensitivity than for planning and organizing and problem solving. The lowest number of behavioral observations was found for tolerance for stress.

TABLE 2
Means, Standard Deviations, and Effect Sizes of the Number of Behavioral Observations Broken Down by Exercise Instructions

	With Specific Exercise Instructions[a]		*Without Specific Exercise Instructions*[b]				
	M	*SD*	*M*	*SD*	*D*	*F*	*p*
Planning and organizing	2.34_c	1.49	2.36_b	1.38	−.01	.011	.916
Interpersonal sensitivity	3.72_a	2.02	3.80_a	2.14	−.04	.097	.756
Problem solving	3.06_b	1.98	2.52_b	1.27	.32	2.531	.115
Tolerance for stress	1.81_d	1.24	2.10_b	1.69	−.20	1.029	.313
Total	10.93	4.66	10.78	4.63	.03	.025	.875

Note. $N = 103$. The means and standard deviations refer to the number of behavioral observations. The *d* values are effect sizes calculated by the formula ($M_{\text{with prompts}} - M_{\text{without prompts}}$) / Total *SD*. Positive effect sizes (*d* values) mean that the number of behavioral observations was higher in the condition with prompts than in the condition without prompts. Means with different subscripts (vertical) are significantly different at .05 level.

[a] $n = 53$. [b] $n = 50$.

Means, standard deviations, and effect sizes of the number of behavioral observations broken down by exercise instructions are presented in Table 2. H2a dealt with the effect of the use of specific exercise instructions on the observed frequency of behavior during the role-play. An ANOVA analysis with specific exercise instructions as factor and with the total number of behavioral observations as a dependent variable revealed no significant effect of specific exercise instructions, $F(1, 101) = .025$, $p = .875$ (partial $\eta^2 = .00$). So, H2a was not supported.

H2b posited the use of specific exercise instructions to have an effect on the observed frequency of dimension-related behavior in the role-play. We conducted a MANOVA with specific exercise instructions and role-player prompts as factors and the number of behavioral observations for problem solving, interpersonal sensitivity, planning and organizing, and tolerance for stress as a set of four dependent variables. Results showed no significant multivariate main effect for specific exercise instructions, $F(4, 96) = 1.147$, $p = .339$ (partial $\eta^2 = .05$). Thus, H2b, which stated that the use of specific exercise instructions will enhance the frequency of observed dimension-related candidate behavior during the role-play, was not supported. Note that we did also not find an interaction effect between prompts and instructions.

H3a posited that the frequency of observed candidate behavior (the number of behavioral observations) will be higher in AC exercises with role-player prompts than in AC exercises with specific exercise instructions. We tested H3a using planned comparisons (the total number of behavioral observations served as dependent variable). The planned comparison test was significant, $t(99) = 3.221$, $p < .01$, with the role-player prompts condition showing a higher number of behavioral observations than the condition with specific exercise instructions. The effect size was $d = .83$. Thus, H3a was supported.

H3b stated that the frequency of observed dimension-related behavior (the number of behavioral observations *per* dimension) will be higher in AC exercises with role-player prompts than in AC exercises with specific exercise instructions. Again, we used planned comparisons (with the number of behavioral observations per dimension as dependent variables). For the dimensions planning and organizing, interpersonal sensitivity, and tolerance for stress the planned comparison tests were significant, $t(99) = 3.111$, $p < .01$, $d = .81$, $t(99) = 3.429$, $p < .01$, $d = .88$, and $t(99) = 3.191$, $p < .01$, $d = .85$, respectively. For these dimensions the number of behavioral observations was higher in the condition with role-player prompts than in the condition with exercise instructions. For the dimension problem solving the planned comparison test was not significant, $t(99) = –.891$, $p = .375$, $d = –.15$. Thus, H3b gained support for three out of the four dimensions.

DISCUSSION

The present study used TAT as a theoretical lens for proposing two exercise stimuli which might activate behavior in AC exercises. Specifically, we examined the effects of the implementation of role-player prompts and specific exercise instructions on the observability of behavior. A first key result was that the use of role-player prompts led to greater observability of behavior. Thus, role-player prompts ensure that the situational elements of the target job might be more represented, thereby giving assessees more opportunities for showing relevant behavior. This effect was also found at the level of dimensions which is in line with Howard (2008) suggesting that AC developers should indicate the target behaviors that fit to the relevant dimensions and create simulations that evoke these behaviors. More specifically, when role-players used prompts assessors wrote down more behavioral observations for the dimensions planning and organizing, interpersonal sensitivity, and tolerance for stress. The use of prompts might be an answer to previous findings that assessors often need to rely on one particular behavioral reaction to score candidates on several dimensions (Brannick et al., 1989).

The results support our expectations that the use of role-player prompts during the exercise serve to increase the relevance of certain dimensions, which in turn might have beneficial effects on the opportunity for assessors to observe a higher number of behavioral observations per dimension. Given the beneficial effects of the use of role-player prompts we suggest that in AC practice role-players are trained to use multiple prompts in a consistent fashion across candidates. Prompts are useful for evoking behavior both for dimension-based and task-based ACs. To date, current AC guidelines do not include the notion of prompts (International Task Force on Assessment Center Guidelines, 2009).

In contrast with the beneficial effects of prompts for the dimensions planning and organizing, interpersonal sensitivity, and tolerance for stress, role-player prompts did not have an effect on the number of behavioral observations for the dimension problem solving. The g loading versus personality loading of AC dimensions (Whetzel, McDaniel, & Nguyen, 2008) might provide an explanation for these findings. Hereby cognitive ability provides the “can do” and personality the intrinsic “will do” of performance in ACs (Lievens et al., 2009). It might be that AC prompts developed for role-plays are better able to evoke dimensions with a higher personality loading (i.e., interpersonal sensitivity, planning and organizing, tolerance for stress vs. problem solving).

Despite the positive evidence of using prompts, our results showed no significant effects for exercise instructions. This finding might imply that exercise instructions are not appropriate to serve as AC stimuli to evoke specific behavior as they are more general. As such, exercise instructions might be too weak cues to elicit specific behaviors. Further, exercise instructions are provided before the start of the exercise. Hence, candidates might attend to them or might not attend to them, whereas role-player prompts are interactive in that the candidate is expected to respond to them. Role-players might also try repeated times during the exercise to evoke behavior related to the dimensions targeted, which is not possible for exercise instructions. Moreover, in today's AC best practices, exercise instructions (such as the AC exercise that we purchased in this study) might already be as specific and clear as possible.

This plausible explanation for the null findings related to exercise instructions invokes an interesting avenue for future research. For future studies, it might be relevant to evaluate the effects of behavior-related cues that are embedded in exercise materials (e.g., the memos, reports, background documents, etc.). For example, good problem solvers might put together Cue 1, Cue 2, and Cue 3 to come to a reasonable conclusion. Embedding situational stimuli in exercise instructions in such a way gets at the quality of problem solving, not just whether participants understand that they are supposed to solve a problem (as triggered by the exercise instruction manipulation in our study). Future studies that use such an embedded cue approach in exercise materials might find effects of exercise instructions on observability as planting behavior-related cues in the exercise materials themselves instead of mentioning at the beginning might make the situational strength (and therefore also the activation potential) of exercise instructions similar to that one of role-player prompts. So we believe it is still a bit premature to discount the value of exercise instructions as a strategy for increasing candidate behavior.

Third, from a theoretical perspective this study provides a test of the relative importance of situational trait relevance versus situational strength in activating behavior by examining the impact of role-player prompts versus exercise instructions. Although both interventions increase situational trait relevance by providing cues for activating behavior, they differ from each other in that there are conceptual reasons why role-player prompts might be higher on situational strength than the exercise instructions manipulated in this study. Our results confirmed that prompts elicited significantly more behavior for the total number of behavioral observations and for three out of four dimensions than the exercise instructions manipulated in this study. As the intervention (prompts) that was situational stronger was the most effective, the results also are indicative that situational trait relevance might be a necessary albeit insufficient condition to activate behavior. Only in combination with situational strength, sufficient conditions to activate behavior seem to be present.

Given the positive effects of the use of role-player prompts, future research should examine whether the use of role-player prompts has beneficial effects on the criterion-related validity as the inclusion of exercise stimuli that trigger relevant behavior might increase the point-to-point correspondence with the criterion dimensions (Tett & Schleicher, 2001). Along these lines, it would also be interesting to scrutinize whether the use of role-player prompts increases the interrater reliability among assessors (Schollaert & Lievens, 2011). Given that prior research has shown that the interrater reliability of structured interviews is higher than that of unstructured interviews, we expect the same effect when standardized role-player cues are used (Conway, Jako, & Goodman, 1995).

Another direction for future research consists of enlarging the assessor training with an explanation of the prompts used by role-players. In this study, assessors were not familiar with the prompts used. Assessors who are familiarized with the role-player prompts might be better able to focus on the behavior elicited. This may lead to a more structured rating process because assessors know when a particular dimension is being evoked by a role-player prompt. In that case, effects of prompts on psychometric criteria that deal with assessor ratings (i.e., reliability and validity) might be expected. Essentially, this leads to a double standardization as both role-players and assessors are familiar with the prompts used (viz. in eliciting behavior as well as in scoring it), which bears resemblance to the two main dimensions of interview standardization (i.e., interview question standardization and response scoring standardization; Huffcutt & Arthur, 1994).

Some limitations should be acknowledged. A first limitation is related to the potential lack of generalizability of our results to real hiring contexts because this was a simulated selection context. However, we tried as best as we could to ensure the external validity and realism of our study. A second limitation refers to the AC exercise used. Prompts were developed only for a role-play exercise. Future studies should examine their viability for other exercises. Third, this study increased the situational relevance of all four dimensions at the same time. A more detailed test of TAT consists of increasing the relevance of one or two dimensions relative to others by manipulating situational stimuli relative to those dimensions only and then in another condition reversing the manipulation so that the previously relevant dimensions are no longer the relevant ones and the previously nonrelevant ones are now relevant.

In conclusion, traditional AC design approaches mainly focus on the whole exercise as a vehicle for evoking behavior. This study sought to expand on the traditional view by planting multiple job-related stimuli in AC exercises to elicit behavior. Regardless of how candidate behavior is evaluated by assessors (task-based vs. dimension-based ACs), eliciting and observing behavior is key to effective assessment and development centers. From a theoretical viewpoint, this study showed how TAT can be used in a prescriptive way for improving the quality of AC exercises. At a practical level, we developed a useful tool (role-player prompts) for increasing the observability of candidate behavior in ACs.

ACKNOWLEDGMENTS

Funding for this study came from the SIOP Foundation via the Douglas W. Bray and Ann Howard Research Grant.

REFERENCES

Ahmed, Y., Payne, T., & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment*, *5*, 62–68.

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

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695–716.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 506–520.

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

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

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.

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

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.

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

Haaland, S., & Christiansen, N. D (2002). Implications of trait-activation theory for e evaluating the construct validity of assessment center ratings. *Personnel Psychology*, *55*, 137–163.

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

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184–190.

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.

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.

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.

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.

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.

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.

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.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*, 252–283.

Schollaert, E., & Lievens, F. (2011). The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment*, *19*, 190–197.

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

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

Tett, R. P., & Schleicher, D. J. (2001, April). Assessment center dimensions as ‘‘traits’’: New concepts in AC design. In M. Born (Chair), *Assessment center dimension validation: Are we asking the wrong questions?* Symposium conducted at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291–309.

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.

APPENDIX A

TABLE A1
Overview of the 21 Prompts and the Frequency With Which They Were Displayed

		Frequency			
		Without Prompts[a]		*With Prompts*[b]	
Prompts	*Dimension*	*0*	*≥1*	*0*	*≥1*
1. Asking “Where did you get this information?”	PS	77.6	22.4	17.3	82.7
2. Asking “How did you find out?”	PS	100.0	0.0	65.4	34.6
3. Answering questions in a vague way.	PS	100.0	0.0	92.3	7.7
4. Asking “Why do you think I feel bad?”	PS	98.0	2.0	36.5	63.5
5. Asking “What is the main problem/solution?”	PS	98.0	2.0	100.0	0.0
6. Mentioning “I do not have plenty of time, what is the agenda of the meeting? What do you want to discuss?”	PO	16.3	83.7	3.8	96.2
7. Asking “Is your proposition realistic in terms of time? I have a very busy schedule the next weeks.”	PO	98.0	2.0	25.0	75.0
8. Asking “What do you expect from me the next period?”	PO	89.8	10.2	23.1	76.9
9. Asking “What is the top priority?”	PO	100.0	0.0	76.9	23.1
10. Asking “How do we organize this?”	PO	100.0	0.0	88.5	11.5
11. Asking “Can you explain it in more details/more concrete? Can you give the facts and figures of the plan?”	PO	100.0	0.0	100.0	0.0
12. Mentioning “Actually, I prefer not to do this.”	IS	100.0	0.0	67.3	32.7
13. Mentioning “I do not aim to make myself look ridiculous to the clients.”	IS	98.0	2.0	36.5	63.5
14. Mentioning “I feel offended by the fact that I had to come here.”	IS	98.0	2.0	59.6	40.4
15. Asking “Do you still trust me?”	IS	100.0	0.0	57.7	42.3
16. Mentioning “Our conversation makes me feel uncomfortable as I get the impression that the colleagues gossip about me.”	IS	83.7	16.3	44.2	55.8
17. Mentioning “You are partially right.”	IS	95.9	4.1	51.9	48.1
18. Mentioning “I also heard some complaints about you from other colleagues.”	TS	100.0	0.0	36.5	63.5
19. Mentioning “You are not perfect either.”	TS	100.0	0.0	73.1	26.9
20. Saying “No, I refuse to do that.”	TS	100.0	0.0	63.5	36.5
21. Shaking one’s head (repeatedly).	TS	100.0	0.0	88.5	11.5

Note. PS = problem solving; PO = planning and organization; IS = interpersonal sensitivity; TS = tolerance for stress.

[a]$n = 49$. [b]$n = 54$.

APPENDIX B

TABLE B1
Overview of the Four Specific Exercise Instructions

Specific Exercise Instructions	*Dimensions*
1. X does not perform well in the last weeks. Try to find the real cause of the problem and look for a solution. You are not sure whether you will sign up Dominique for the course “Project management”. You have to correct and motivate Dominique to do her job better again.	Problem solving
2. Be concrete in your approach and finish the conversation with some clear agreements.	Planning and organizing
3. Do not ignore the perspective and the feelings of the role-player. Try not to offend the role-player.	Interpersonal sensitivity
4. The role-player will probably not immediately agree with your solution. Try to stay calm even when the role-player puts pressure on you.	Tolerance for stress