Journal of Applied Psychology
2006, Vol. 91, No. 2, 247–258

0021-9010/06/$12.00 DOI: 10.1037/0021-9010.91.2.247

Large-Scale Investigation of the Role of Trait Activation Theory for Understanding Assessment Center Convergent and Discriminant Validity

Filip Lievens
Ghent University

Christopher S. Chasteen and Eric Anthony Day
University of Oklahoma

Neil D. Christiansen
Central Michigan University

This study used trait activation theory as a theoretical framework to conduct a large-scale test of the interactionist explanation of the convergent and discriminant validity findings obtained in assessment centers. Trait activation theory specifies the conditions in which cross-situationally consistent and inconsistent candidate performances are likely to occur. Results obtained by aggregating correlations across 30 multitrait–multimethod matrices supported the propositions of trait activation theory, shedding a more positive light on the construct validity puzzle in assessment centers. Overall, convergence among assessment center ratings was better between exercises that provided an opportunity to observe behavior related to the same trait, and discrimination among ratings within exercises was generally better for dimensions that were not expressions of the same underlying traits. Implications for assessment center research and practice are discussed.

Keywords: assessment centers, trait activation, convergent validity, discriminant validity

The construct-related validity of assessment center ratings continues to be controversial among researchers and practitioners. A focal issue in the debate involves the common finding that ratings of the same dimension do not converge well across exercises (i.e., poor convergent validity) and that distinctions between dimensions appear to be blurred within exercises (i.e., poor discriminant validity). These findings are especially troublesome for developmental assessment centers in which dimensional ratings serve as a basis to provide participants with detailed feedback about their strengths and weaknesses. The crux is that this feedback and resulting action plans might be flawed without evidence that assessment centers provide a consistent and distinct measurement of the dimensions (Bycio, Alvares, & Hahn, 1987; Fleenor, 1996; Joyce, Thayer, & Pond, 1994; Lievens & Klimoski, 2001).

Traditionally, it has been argued that the poor construct-related validity is the result of problems with assessment center design, evoking assessors' biases and inaccuracy. Hence, the general strategy has been to modify the design of the assessment center to facilitate assessors' rating processes. Lievens and Conway (2001) conducted a large-scale review to test the viability of this traditional explanation. Generally, their statistical review of 34 multitrait–multimethod (MTMM) matrices confirmed the importance of careful assessment center design. Design features such as limiting the number of dimensions to be rated and using psychologist assessors significantly increased the quality of construct measurement in assessment centers. Another large-scale review of assessment center design issues reached similar conclusions (Woehr & Arthur, 2003).

Although many studies found improvements in construct validity by modifying the design of assessment centers, these procedural interventions were less successful in other studies. For instance, although Chan (1996), Schneider and Schmitt (1992), and Fleenor (1996) carefully implemented many design recommendations, evidence of construct validity was still lacking. Hence, some researchers have argued that careful assessment center design may install necessary but insufficient conditions for ensuring construct validity (Lance, Foster, Gentry, & Thoresen, 2004; Lance et al., 2000; Lievens, 2001, 2002).

Recently, actual differences in candidates' performance across situations have been put forward as an additional explanation of the construct validity findings reported in the assessment center literature (Haaland & Christiansen, 2002; Lance et al., 2000; Lievens, 2001, 2002). This explanation builds on interactionist models in social and personality psychology. The main argument is that assessment center exercises such as in-baskets, group discussions, or role-plays are dissimilar situations that place very different psychological demands on participants. Hence, they cannot be considered to be parallel measures of the same dimensions because they evoke inconsistent behavior from candidates across exercises, and as a result evidence of construct validity tends to be poor.

Although some primary studies have shown that differences in candidate performances across exercises affect the construct validity of assessment center ratings, no large-scale test of this

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; Christopher S. Chasteen and Eric Anthony Day, Department of Psychology, University of Oklahoma; Neil D. Christiansen, Department of Psychology, Central Michigan University.

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

interactionist explanation has been conducted. Even more important, a theoretical rationale as to why candidates often exhibit inconsistent performance across situations has yet to be offered and investigated. This is surprising because conclusive evidence in favor of this more recent explanation might shed a different and more positive light on the construct validity puzzle. In fact, whereas the traditional explanation puts the blame on assessors and on flawed assessment center design, this more recent explanation involves a focus on candidate performances.

The purpose of this study was to provide a large-scale and systematic investigation of this more recent interactionist explanation of the typical construct validity results of assessment centers. We did so by using trait activation theory (Tett & Burnett, 2003; Tett & Guterman, 2000) to predict when candidates' behavior would be expected to be consistent across exercises and when ratings on dimensions from the same exercise could be expected to be discriminable. Using the five-factor model (FFM) of personality dimensions as an organizing framework, we therefore aggregated correlations from both published and unpublished MTMM matrices derived from assessment center ratings to test hypotheses regarding convergent and discriminant validity.

Trait Activation Theory and Assessment Center Exercises

Trait activation theory is a recent theory that focuses on the person-situation interaction to explain behavior on the basis of responses to trait-relevant cues found in situations (Tett & Guterman, 2000). These observable responses serve as the basis for behavioral ratings on dimensions used in a variety of assessments, such as performance appraisal, interviews, or assessment centers (Tett & Burnett, 2003). The emphasis in trait activation theory is on the importance of situation trait relevance in order to understand in which situations a personality trait is likely to manifest in behavior. A situation is considered relevant to a trait if it provides cues for the expression of trait-relevant behavior (Tett & Guterman, 2000), an idea that has roots in Murray's (1938) notion of "situational press." For example, it would generally not be productive to assess individuals on the trait of aggression during a religious service because there are few cues likely to elicit aggressive behavior.

Also relevant from the trait activation perspective is the role of situation strength. Strong situations involve unambiguous behavioral demands where the outcomes of behavior are clearly understood and widely shared (Mischel, 1973). Relatively uniform expectations result in few differences in how individuals respond to the situation, obscuring individual differences on underlying personality traits even where relevant. Conversely, weak situations are characterized by more ambiguous expectations, enabling much more variability in behavioral responses to be observed. A related concept involves what has been referred to as the competency demand hypothesis (e.g., Mischel & Shoda, 1995), where research has shown that individual differences are obviated when situations have demanding behavioral requirements in terms of ability, skills, or personality traits.

Trait relevance and strength therefore represent distinct characteristics of situations that figure into the concept of trait activation potential (TAP; Tett & Burnett, 2003). On the one hand, 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. The traits considered are typically cast in the FFM framework because FFM traits consist of clearly understood behavioral domains and represent the natural categories that individuals use to describe and evaluate social behavior (e.g., Costa & McCrae, 1992; Goldberg, 1992; Haaland & Christiansen, 2002; Lievens, De Fruyt, & Van Dam, 2001). Hence, they facilitate classification of exercises with similar situational demands. On the other hand, situation strength is more of a continuum that refers to how much clarity there is with regard to how the situation is perceived. Very strong situations are therefore likely to negate almost all individual differences in behavior without regard to any specific personality trait. The analogy used by Tett and Burnett (2003) to distinguish between the two concepts is that trait relevance is akin to which channel a radio is tuned to whereas situation strength is more similar to volume; relevance determines what is playing and strength (inversely) whether it will be heard.

These concepts are relevant to assessment centers because exercises are developed to allow a broad range of behavior to be observed across exercises and to be demanding enough that differences in candidates' performance can be observed. Because of this, they will necessarily differ in the cues present with regard to various FFM traits. For example, it would be expected that a leaderless group discussion would provide ample opportunity to observe differences in behavior relevant to the FFM trait of Extraversion; however, it might be difficult to observe such differences while a candidate completes an in-basket exercise. Assessment center exercises therefore represent situations that differ in terms of their TAP. The more likely it is that behavior can be observed within an exercise that is relevant to a particular FFM trait, the higher the activation potential would be for that FFM trait. The opportunity to observe differences in trait-relevant behavior within a situation depends upon both the relevance and strength of the situation and has relevance to both the convergent and discriminant validity of dimension ratings.

Trait Activation and the Convergent Validity of Assessment Center Ratings

One implication for the construct validity of assessment centers is that when exercises differ in the extent that behavior relevant to the same personality traits can be observed, it will be more difficult to observe consistent behavior across exercises. Because personality traits manifest in behavior as responses expressed to trait-relevant cues that vary from situation to situation, cross-situational behavior will appear consistent only when behaviors relevant to the same personality traits can be observed in each situation. Thus, convergence will be poor for ratings of a dimension related to a given personality trait when exercises differ in their activation potential for that trait because there is less consistency in the actual behavior of the candidates. On the other hand, stronger convergence would be expected when such ratings are based on exercises where there is significant opportunity to observe trait-relevant behavior in each exercise (i.e., two exercises high in TAP).

Consider ratings made on the dimension of "interpersonal influence," which has been defined as being based on behaviors that are known expressions of the FFM trait of Extraversion. Because a leaderless group discussion and a role-play exercise would be both expected to provide cues relevant to this FFM trait, conver-

gence between ratings would be expected. However, as mentioned above, because the in-basket exercise probably does not provide many cues for expression of trait-relevant behavior, ratings on the interpersonal influence dimension from this exercise would not be expected to correlate as strongly with those from the other exercises.

Empirical support for this implication of trait activation theory can be found in a recent study by Haaland and Christiansen (2002), who examined whether there would be poor convergence results from correlating ratings on exercises that differed in the extent to which that behavior was relevant to personality traits. The first step in their research was to have individuals familiar with the specific assessment center exercises make judgments on whether it would be possible to observe behavior relevant to the traits represented in the FFM per exercise. The subject matter experts (SMEs) were also asked to link the dimensions of the assessment center with the FFM traits because greater convergence would be expected only on dimensions that were conceptually relevant to a given trait. The correlations between ratings from exercises where there was more opportunity to observe trait-relevant behavior were compared with the correlations between ratings from exercises where there was less opportunity, with the results providing support for the implication that the TAP of the exercises plays an important role in determining the construct validity of the ratings.

The research by Haaland and Christiansen (2002) therefore expanded understanding of the conditions when convergence might be expected. Highhouse and Harris (1993) showed that convergence is better across exercises where the same behavior could be observed. It is noteworthy that the trait activation approach extends beyond this because the exact same behavior need not be observed for two exercises to be considered similar; behaviors can be involved that may on the surface appear different but are related to the same personality trait. The example used by the researchers to illustrate this point involved consideration of one exercise that required risk-taking behavior to successfully resolve the situation and one that involved persuading a group of people to adopt the candidate's position. Because these behaviors can be seen as falling within the construct domain of Extraversion, convergence on ratings from a dimension linked to that FFM trait would be expected across these exercises.

Another important issue is that research in the area of person perception has shown that cues relevant to some traits are more easily observed and tend to be used more effectively, resulting in more accurate judgments based on those behaviors (e.g., Funder & Colvin, 1988; Funder & Dobroth, 1987; Funder & Sneed, 1993). A large body of research has examined convergence of ratings made by observers of individuals interacting in limited social situations similar to those found in assessment centers, with the consensus being that convergence is best for judgments based on behavioral cues related to the FFM traits of Extraversion and Conscientiousness (Kenny, Horner, Kashy, & Chu, 1992; Kenny, Albright, Malloy, & Kashy, 1994). Reconciling this with trait activation theory, it suggests that the effect of TAP on the convergent validity of assessment center ratings may depend on how easy it is to observe and use the trait-relevant behavioral cues. Theoretical support for this can be found in Funder's realistic accuracy model, which posits that judgmental accuracy is a multiplicative function of the extent that trait-relevant cues are available for observation in the particular context and the extent that the cues are detected and used appropriately (Funder, 1999). Thus, the opportunity to observe differences in behavioral cues related to the same FFM trait (Extraversion and Conscientiousness) may have more impact for FFM traits where cue processing has been shown to be more effective. All of this leads to the following hypotheses:

> *Hypothesis 1a:* Convergence on ratings from a given dimension linked to a specific FFM trait will be stronger across exercises that activate this specific FFM trait (i.e., exercises high in TAP) than across exercises that do not activate this specific FFM trait (i.e., exercises low in TAP).
>
> *Hypothesis 1b:* The trait activation effect will be more pronounced for dimensions related to Extraversion and Conscientiousness.

Trait Activation and the Discriminant Validity of Assessment Center Ratings

The relationship between the TAP of exercises and the underlying personality traits of the candidates also has implications for the discriminant validity of assessment center ratings. Specifically, exercises may have cues for behaviors that are related to different dimensions but are expressions of the same trait. Because of this, dimensions that are based on behaviors that are expressions of the same trait will correlate more strongly because they share a common cause. Thus, discriminant validity will be worse in part because the MTMM approach assumes that constructs are totally discrete whereas assessment center dimensions may not be. Some support for this can be found in Lievens's (1998) summary of assessment center research where the use of conceptually distinct dimensions had positive effects on discriminant validity.

However, by implicating links to underlying traits as a causal explanation for strong dimension correlations within exercises, trait activation theory again goes beyond the simpler conceptualization that dimensions may overlap because they require the same behaviors. For example, ratings on the dimensions of oral communication and impact may be based on very different behaviors but all may be expressions of the FFM trait of Extraversion. Similarly, planning and organizing and initiative dimensions may be based on behaviors that are expressions of the FFM trait of Conscientiousness. It is therefore noteworthy that the trait activation perspective provides psychological depth here as well.

This rationale is also consistent with research that has shown that different dimensions commonly represented in assessment center ratings can be both conceptually and empirically linked to personality traits such as those in the FFM (e.g., Furnham, Crump, & Whelan, 1997; Goffin, Rothstein, & Johnston, 1996; Haaland & Christiansen, 2002). The trait activation perspective suggests that part of the reason dimension ratings correlate so highly is that they may be based on behavioral cues related to a common personality trait. For example, strong evidence of discriminant validity might not be expected between ratings of presentation skills and persuasion dimensions within an exercise because both of these dimensions may be expressions of the FFM trait of Extraversion. Alternatively, better discrimination might be observed when correlating ratings of problem solving and interpersonal skills because neither of these dimensions is an expression of the same underlying trait(s). Traditional approaches for examining the discriminant

validity of assessment center ratings have not taken this into account—they typically involve analyzing all correlations among all dimension ratings without regard to possible personality traits. Accordingly, we tested the following hypothesis.

> *Hypothesis 2:* Discrimination among ratings within exercises will be better for dimensions that are not expressions of the same underlying traits than for dimensions that are.

Method

Database

To find MTMM matrices of assessment center ratings, we conducted a search using a number of computerized databases (i.e., PsycLit, Dissertations Abstracts, and Current Contents). Keywords included "assessment center" in combination with "constructs," "validity," or "multitrait–multimethod." In addition, we contacted assessment center researchers and scrutinized reference lists from obtained studies to find other published and unpublished studies.

We used several criteria for inclusion. First, only so-called within-exercise dimension ratings were used. These are ratings made upon completion of each exercise. Second, these ratings had to be cast in an MTMM correlation matrix in which at least two dimensions served as traits and at least two exercises served as methods. When methods did not represent assessment center exercises, they were removed from the MTMM matrix. For example, in some matrices methods represented different assessors, different rating sources, structured interviews, or (video-based) situational judgment tests. We also removed from the MTMM matrix dimensions that were rated in only one exercise because the convergent validity of these dimensions could not be determined. Third, MTMM matrices had to be based on data from operational assessment centers. Fourth, consistent with assessment center practices, we included only matrices in which assessors rotated across the assessment center exercises.

The resulting database included 30 matrices, dating from 1976 to 2004. Sources for these matrices are marked with an asterisk in the Reference List. Nineteen matrices came from published articles, eight came from unpublished dissertations, two came from an unpublished manuscript, and one was a conference presentation. The mean sample size was 225.63 (*Mdn* = 176, total N = 6,769), the mean number of dimensions per study was 7.20 (*Mdn* = 7, total N = 216, total number of unique dimension names = 99), and the mean number of exercises per study was 4.00 (*Mdn* = 4, total N = 119, total number of unique exercise names = 88).

Classification of Dimensions and Exercises

Given the large number of dimensions and exercises gathered across the MTMM matrices, we decided to group the assessment center dimensions and exercises in a manageable set of generic dimensions and exercises. Using the taxonomy of assessment center dimensions developed by Arthur, Day, McNelly, and Edens (2003), we collapsed the 99 dimensions into seven categories: communication, consideration and awareness of others, drive, influencing others, organizing and planning, problem solving, and tolerance of stress and uncertainty. On the basis of Thornton's (1992) taxonomy of assessment center exercises, we collapsed the 69 exercises into six types: in-basket, cooperative leaderless group discussion, competitive leaderless group discussion, case study, role-play, and oral presentation. Appendixes A and B provide descriptions of the dimension and exercise categories, respectively.

Filip Lievens and Eric A. Day classified both dimensions and exercises independently into these taxonomies. Kappa coefficients for both dimensions and exercises were above .90. If consensus could not be reached in 10 min of discussion for instances when there was not independent agreement, then the data points for these dimensions and exercises were excluded from our analyses. Table 1 provides a summary of the number of separate study names that were classified into the dimension and exercise categories we used.

Identifying Exercises High in TAP

To identify which exercises were high in TAP, six SMEs, all of whom had at least a master's degree in industrial and organizational psychology (three had doctoral degrees), were surveyed using the trait activation rating instrument developed by Haaland and Christiansen (2002). The individuals surveyed qualified as SMEs because they typically had served as assessors in at least two assessment centers. Most of our SMEs also published at least one manuscript regarding either assessment centers or the FFM; some have published manuscripts on both assessment centers and the FFM. The trait activation rating instrument first required SMEs to make judgments about how the traits related to each exercise. Second, SMEs made judgments regarding the extent to which the exercises provided opportunities for trait-relevant behaviors to be displayed. Total scores across the items pertaining to each trait-exercise link could range from 3 to 30. Kendall's coefficient of concordance among SME ratings was .58 (the overall intraclass correlation was .90).

The upper part of Table 2 summarizes the trait-exercise TAP linkages. To decide which exercises would be considered high in TAP for a given trait, we used the same a priori cutoff of 18 used by Haaland and Christiansen (2002). However, using this approach, Emotional Stability was not linked to any exercise category. To permit analyses for Emotional Stability, we linked Emotional Stability to the two exercises that had the highest TAP ratings for Emotional Stability: competitive leadership group discussion (LGD; 16.5) and oral presentation (17.17). The remaining exercises had TAP ratings below 15.20 for Emotional Stability.[1]

Linking Dimensions to the FFM

To determine which dimensions were conceptually related to the FFM traits, we surveyed five additional SMEs (similar in composition to the aforementioned SME sample), again using the instruments developed by Haaland and Christiansen (2002). Kendall's coefficient of concordance was .70 (the overall intraclass correlation was .95). A summary of the trait-dimension linkages is shown in the lower part of Table 2. The format for making trait-dimension linkages was similar to that used for exercises. So, again, we used the same a priori cutoff of 18 used by Haaland and Christiansen to decide which dimensions would be linked to FFM traits. Table 2 shows that each dimension was related to only one FFM trait. Note that we did not artificially restrict the dimensions to be linked to only one FFM trait. This was simply the result of using the a priori cutoff of 18 used by Haaland and Christiansen

Results

Convergent Validity

To test the hypothesis that convergence would be better between exercises that provide an opportunity to observe behavior related to the same trait, we compared the average monotrait–heteromethod (MTHM) correlation across all matrices for exercises that were both high in activation potential for a given trait to the average MTHM correlation across all matrices for exercises in which one or both exercises were low in activation potential. This approach is analogous to the approach used by Haaland and

[1] The overall results for both the MTHM correlations (i.e., convergent validities) and HTMM (i.e., discriminant validities) correlations were essentially no different when the cutoff of 18 was applied to Emotional Stability.

Table 1
Classification Summary of Study Dimensions and Exercises

Classification category	No. of unique study names[a]	% of studies with at least one[b]	Examples
		Dimensions	
Communication	5	63	Communication Oral communication Presentation skills
Consideration and awareness of others	17	80	Interpersonal skills Managing interaction Sensitivity
Drive	6	33	Achievement motivation Energy Initiative
Influencing others	16	47	Impact Leadership Persuasiveness
Organizing and planning	17	67	Controlling Delegation Time sensitivity
Problem solving	25	83	Analyzing Decision making Judgment
Tolerance for stress and uncertainty	7	33	Adaptability Flexibility Stress tolerance
Unable to classify	6	47	Internal contacts Oral comprehension Recruitment and selection
Total	99		
		Exercises	
Case analysis	4	20	Policy development plan Problem analysis Written problem analysis
Competitive leaderless group discussion	8	27	Competitive allocation exercise Competitive group exercise Grant allocation competitive group discussion
Cooperative leaderless group discussion	12	43	Leaderless group discussion Noncompetitive management problem exercise Team manufacturing cooperative group discussion
In-basket	5	50	In-basket In-basket (operational issues) In-tray
Oral presentation	8	33	Analytical problem and presentation Presentation Problem analysis presentation
Role-play	15	60	Customer meeting Performance counseling Meeting with a subordinate
Unable to classify	17	47	Interview Patterned/situational interview Written situational test
Total	69		

[a] Number of nonredundant dimension/exercise labels classified in the dimension/exercise category. [b] Percentage of studies that included at least one representative from the dimension/exercise category.

Christiansen (2002). We performed this comparison for each trait of the FFM, using those dimensions that had been judged by our SMEs as related to that trait. Although we did not conduct a meta-analysis (we were not interested in validity generalization), we did calculate the sample-weighted means, the *SD*s, and the medians for these effects to have an estimate of the variability of our results. The results of these analyses are shown in Table 3.

Although the difference in effects was small, the results were consistent with Hypothesis 1: Organizing the MTHM correlations by activation potential resulted in a larger overall average correlation using the traits of the FFM. That is, high TAP correlations were larger (mean $r = .33$) than low TAP correlations (mean $r = .27$). To further test Hypothesis 1, we conducted a 2 (TAP: high vs. low) $\times$ 5 (FFM trait) between-subjects analysis of variance

Table 2
Summary of Dimension and Exercise Linkages to Traits

Exercise/dimension	FFM trait				
	E	A	C	ES	O
Exercise					
Case analysis	5.17	6.50	**23.83**	10.00	17.67
Competitive LGD	**24.17**	**21.67**	**20.00**	**16.50**	**19.33**
Cooperative LGD	**24.00**	**22.33**	**19.50**	15.17	**21.00**
In-basket	6.33	12.17	**26.50**	13.33	15.50
Oral presentation	16.33	11.50	**19.67**	**17.17**	14.33
Role-play	**20.00**	**20.17**	17.00	15.00	16.67
Dimension					
Communication	**19.20**	15.80	8.00	8.60	9.00
Consideration and awareness of others	13.00	**21.60**	5.80	6.80	9.20
Drive	15.60	3.80	**18.20**	9.20	8.80
Influencing others	**19.00**	10.40	6.60	7.00	8.60
Organizing and planning	6.00	4.40	**24.20**	9.40	6.20
Problem solving	5.60	4.20	14.60	7.80	**22.60**
Tolerance for stress and uncertainty	7.00	10.60	8.40	**24.20**	11.00

Note. Boldface font indicates a high trait linkage. FFM = Five factor model; LGD = leadership group discussion; E = Extraversion; A = Agreeableness; C = Conscientiousness; ES = Emotional Stability; O = Openness to Experience.

(ANOVA) with the observed correlations from all the matrices as the data points. The results indicated that the high TAP versus low TAP distinction was statistically significant, $F(1, 339) = 7.68$, $p < .01$.

The ANOVA results also indicated a significant main effect for trait, $F(4, 339) = 6.79$, $p < .01$. Correlations associated with dimensions linked to Extraversion and Emotional Stability tended to be stronger than the correlations associated with dimensions linked to Conscientiousness, Openness, and Agreeableness. We conducted planned comparisons to test our hypothesis that TAP effects would be more pronounced for dimensions linked to Extraversion and Conscientiousness, as compared with the effects for Openness and Agreeableness.[2] The difference between high and low TAP correlations for dimensions linked to Extraversion and Conscientiousness (high mean $r = .32$; low mean $r = .26$) was not stronger than the difference between high and low TAP correlations for dimension linked to Openness and Agreeableness (high mean $r = .33$; low mean $r = .27$). However, it should be noted that the difference for Openness and Agreeableness did not reach conventional levels of statistical significance, $t(172) = 1.70$, $p = .09$, whereas the difference for Extraversion and Conscientiousness was statistically significant, $t(146) = 2.18$, $p < .05$. This difference in statistical significance is due to the disproportionately low number of correlations (26 for Openness and Agreeableness) involving two high TAP exercises. To further test this difference in TAP effects, we calculated 95% confidence intervals around the sample-weighted mean correlations (Whitener, 1990). Consistent with our hypothesis, there was no overlap in confidence intervals for Extraversion and Conscientiousness (low TAP = .24, .31; high TAP = .32, .38), but the intervals largely overlapped for Openness and Agreeableness (low TAP = .27, .31; high TAP = .27, .35).

Because of the nonnormal distribution and the nonindependence of the observed correlations, we also used a nonparametric test to compare the median correlations. This analysis is analogous to the analysis conducted by Haaland and Christiansen (2002). The result of this test was a chi-square with one degree of freedom that indicated 62% of the pairwise comparisons (23,859) showed stronger correlations for high TAP exercises (*Mdn* $r = .33$) compared with low TAP exercises (*Mdn* $r = .26$), $\chi^2 = 1{,}328.04$, $p < .01$. Similar to the results above using ANOVA, these results support our prediction that the convergence among ratings from exercises high in TAP would be stronger than the convergence observed for exercises low in TAP.

Discriminant Validity

To examine the hypothesis that discrimination among ratings within exercises would be better for dimensions that are not expressions of the same underlying traits than for dimensions that are expressions of the same underlying traits, we compared the average heterotrait–monomethod (HTMM) correlation across all matrices for dimensions that were both linked to a given FFM trait to the average HTMM correlation across all matrices for dimensions that did not share a common link to any of the FFM traits. We broke these results down for each of the exercise categories and also analyzed the overall results. These results are shown in Table 4.

Although the difference in effects was small, the results were consistent with Hypothesis 2: HTMM correlations involving two dimensions that did not share a link to any of the traits in the FFM were significantly lower (mean $r = .53$) than the HTMM correlations involving two dimensions that had a similar link to traits (mean $r = .57$), $t(1{,}010) = 1.97$, $p < .05$. However, these differences were statistically significant only for the competitive leaderless group discussion, role-play, and oral presentation exercises.

Similar to the analyses conducted with the MTHM correlations, we also used a nonparametric test to compare the median HTMM correlations. The result of this test indicated 55% of the pairwise

[2] Emotional Stability was excluded from these analyses because of the small number of data points available for Emotional Stability.

Table 3
Average Monotrait–Heteromethod Correlations by Trait Activation Potential

Personality trait/ (AC dimensions)	Trait activation potential of exercises									
	Low					High				
	k	*M*	*SD*	*Mdn*	SWM	*k*	*M*	*SD*	*Mdn*	SWM
Extraversion										
Communication/influencing others	43	.32	.19	.33	.33	31	.39	.13	.40	.40
Conscientiousness										
Drive/organizing and planning	35	.20	.12	.18	.22	39	.26	.14	.29	.31
Openness										
Problem solving	107	.28	.15	.29	.29	6	.33	.14	.38	.33
Agreeableness										
Consideration and awareness of others	41	.25	.18	.22	.27	20	.33	.14	.27	.30
Emotional Stability										
Tolerance of stress and uncertainty	15	.34	.27	.34	.34	3	.45	.05	.48	.46
Overall	241	.27	.17	.26	.29	99	.33	.14	.33	.34

Note. Results for high trait activation potential were derived from ratings between two exercises both high in activation potential for the same trait, whereas results for low involved at least one exercise that was not high in activation potential for that trait. The "overall" row indicates the average computed across all the traits. AC = assessment center; *k* = number of correlation coefficients; SWM = sample-weighted mean correlation.

comparisons (145,638) showed lower correlations when two dimensions did not share a link to any trait (*Mdn r* = .54) compared with when two dimensions shared a link to the same trait within an exercise (*Mdn r* = .59), $\chi^2 = 1{,}512.96$, $p < .01$. Coupled with the results above, these results support Hypothesis 2. Discrimination among ratings within exercises was better for dimensions that were not expressions of the same underlying traits than for dimensions that were expressions of the same underlying traits. However, this was not the case for all exercises. Specifically, median HTMM correlations were higher when two dimensions shared a link to a trait for competitive leaderless group discussions, role-plays, and oral presentations.

Discussion

A common thread running through this study is that the vast majority of past research on assessment center construct validity has neglected to examine the nature of performance in assessment centers. In fact, most prior studies have focused on assessment center design. Although some of these design modifications have

Table 4
Average Heterotrait–Monomethod Correlations by Linkages to Personality Traits

AC exercise	Link to personality trait									
	Dissimilar					Similar				
	k	*M*	*SD*	*Mdn*	SWM	*k*	*M*	*SD*	*Mdn*	SWM
Case analysis	35	.64	.18	.60	.73	8	.60	.25	.59	.69
Competitive LGD	64	.44	.22	.46	.44	20	.59	.17	.57	.58
Cooperative LGD	192	.44	.23	.39	.58	34	.48	.20	.49	.58
In-basket	178	.56	.20	.55	.63	39	.55	.20	.53	.62
Oral presentation	107	.67	.15	.70	.68	22	.72	.12	.73	.69
Role-play	259	.51	.22	.52	.48	51	.57	.19	.57	.53
Overall	837	.53	.22	.54	.57	174	.57	.20	.59	.59

Note. Results for similar links were derived from ratings between two dimensions that shared a link to the same personality trait, whereas results for dissimilar links involved two dimensions that did not share a link to any personality trait. The "overall" row indicates the average computed across all the exercises. AC = assessment center; *k* = number of correlation coefficients; SWM = sample-weighted mean correlation; LGD = leaderless group discussion.

been successful, a more fundamental question has remained virtually unexplored: Namely, how can the typical construct validity findings be explained? It is in this context that trait activation provides a much-needed theoretical framework for better understanding and explaining within-person behavioral variations and consistencies in assessment centers. This general value of trait activation theory is also reflected in this study's main contributions. In the following sections, we discuss these contributions and the key implications for assessment center research and practice.

Main Contributions

A first contribution of this large-scale study is that we showed that trait activation provides a deeper and more sophisticated approach for looking at the convergence of ratings of the same dimensions across assessment center exercises. An advantage of using trait activation theory is that convergence should not be expected among all dimension ratings. In fact, trait activation posits that convergence should be expected only between exercises that provide an opportunity to observe behavior related to the same trait. The greater psychological depth of trait activation is also exemplified by the fact that convergence is also expected across exercises that look different on the surface but activate the same traits on a deeper trait level. On the basis of this more sophisticated approach for examining convergent validity, we found support for trait activation as a theoretical framework for understanding convergent validity in assessment centers. That is, we found support for the proposition that convergence would be better between exercises that provided an opportunity to observe behavior related to the same trait.

This study further showed that trait activation seemed to work best for two traits, namely Extraversion and Conscientiousness. Why did trait activation work for these traits and not for others? The most likely explanation deals with the observability of behavior. As previously mentioned, research about the ease of trait judgment in social psychology (Funder, 1999; Funder & Colvin, 1988; Funder & Dobroth, 1987; Funder & Sneed, 1993; John & Robbins, 1993; Trope, 1986) has consistently shown that not all personality traits are equally observable and detectable in short social interactions (which are comparable to assessment center exercises). In particular, prior meta-analytic work (Connolly & Viswesvaran, 1998) revealed that there was much higher convergence among stranger and self-ratings on traits such as Extraversion and Conscientiousness as opposed to traits such as Openness to Experience, Emotional Stability, or Agreeableness. This indicates that even strangers (i.e., assessors) who had limited opportunity to observe a target person were able to make relatively accurate judgments about that person's level of Extraversion and Conscientiousness. Generally, these meta-analytic results from the social psychology literature fit well with our findings. In our study, high TAP correlations for dimensions linked to Extraversion and Conscientiousness were significantly stronger than the low TAP correlations, whereas high TAP correlations for dimensions linked to Openness and Agreeableness were not significantly stronger than the low TAP correlations.

As another contribution, this large-scale study provides a novel look at the discriminant validity issue in assessment centers. To our knowledge, all prior assessment center studies correlated all dimensions within an exercise to obtain an index of discriminant validity. Such a broad approach focuses only on the surface dimensions and ignores that these dimensions are conceptually related to underlying traits. In addition, specific traits might be expressed in performance dimensions that appear to be distinct on the surface. Therefore, it is important to examine discriminant validity when taking into account how the dimensions relate to underlying traits. This study revealed that discrimination among ratings within exercises was better for dimensions that were not expressions of the same underlying traits than for dimensions that were expressions of the same underlying traits. This was especially the case for exercises that involved challenging interactions with others (i.e., oral presentations, role-plays, competitive LGDs) as compared with individual exercises (i.e., in-baskets or case-analyses). We believe that the finding that better discrimination among ratings is found in exercises that involve "challenging" interactions with others relates to the fact that such competitive situations with others (e.g., fellow candidates, role player, a panel of assessors) elicit a wider variety of behaviors (see the trait relevance concept). When assessors observe a wider variety of behaviors per candidate, it is logical that they can make more fine-grained distinctions among dimensions. Future research is needed to examine more closely this differential effect for some exercises.

Even when one takes account of the underlying principles of trait activation theory, it should be noted that the effects found are rather small. To provide a yardstick of the size of our effects, we compared them to the results of a large-scale study on the effects of assessment center design features (e.g., type of assessor, number of dimensions measured) on construct validity evidence (Lievens & Conway, 2001). The results of both studies were converted into effect sizes based on Cohen's *d* comparing levels of the design features. The effect of design features is sometimes larger and sometimes smaller than the effects of TAP reported in our study. The average effect size for the design features (not including the anomalous result for length of training for dimension variance) in Lievens and Conway's (2001) study was .27, whereas the average effect size in our study was .32, indicating that our effect sizes are in the same range as the ones found in prior research.

It is important to keep in mind that trait activation is only one part, albeit a crucial one, in understanding the assessment center construct validity puzzle. Trait activation theory deals with the variability of behavior across exercises and does not deal with variability among assessors. In addition, we were unable to integrate the trait activation approach with features of the design of assessment centers that have been shown to facilitate evidence of construct validity (see Lievens & Conway [2001] and Woehr & Arthur [2003] for large-scale reviews). It might be that trait activation interacts with design features and stronger effects might be realized when considered in tandem. We were not able to examine whether design factors moderate trait activation effects because there were not enough studies to afford a reasonable breakdown. As more MTMM matrices become available in the future, it should be possible to provide a test of both trait activation and design characteristics. In addition, future research might find larger differences in construct validity based on TAP using more narrow traits than the FFM traits used in this study. Our strategy of aggregating results across a large number of studies constrained us to look at broad categories of exercises and dimensions. As we determined TAP on the basis of generic descriptions of relatively

broad dimensions and exercises, we probably lost some more fine-grained information.

Methodological Implications

Trait activation has key implications for interpreting the construct validity evidence gleaned from an MTMM matrix. In the past, unrealistic assumptions have been placed on the interpretation of MTMM matrices in assessment centers. In fact, high within-exercise correlations have often been interpreted as indicative of low discriminant validity, even though this might have resulted from correlating dimensions that are behavioral manifestations of the same underlying trait. Similarly, high convergent validity coefficients have been unrealistically expected for dimensions, considering that some exercises (situations) frequently do not activate the same underlying traits. All of this illustrates that prior expectations when looking at convergent and discriminant validity in assessment centers may have been overly stringent (also Lance et al., 2000; Lievens, 2002).

Another methodological implication involves the measurement models used to understand assessment center ratings. Very often covariance models adopt an MTMM framework that includes effects of both dimensions and exercises that combine in an additive fashion. However, the interactionist perspective suggests that some exercises may be better at assessing some dimensions than others as a function of both the cues present in the exercises and the relationship of the underlying traits to the behaviors that serve as the basis for ratings. The implication is that the effects of dimension factors on ratings will depend on exercises and that they combine in a more multiplicative fashion. Such approaches have been modeled successfully in the area of multitrait–multirater measurement models using the direct products model (Conway, 1996; Goffin & Jackson, 1992) and could be extended to assessment center research as well as how method (situation) effects are modeled in other areas of assessment.

Implications for Future Research

This study leads to several new directions for assessment center research. First, in this study, a trait activation model served as theoretical framework to shed light on assessment center construct validity. This model focuses on candidate behavior and outlines that cross-situationally consistent behavior on the part of candidates can be expected only when exercises contain similar trait-relevant cues. However, the flip side of this trait activation model is a trait perception model. This model focuses on assessor judgmental processes and specifies that cross-exercise consistency in trait-expressive behavior might be washed out by judgments of assessors (Tett & Burnett, 2003). Indeed, it is possible that even though behavior related to the same trait is consistently expressed in different exercises, assessors judge the appropriateness of this behavior differently across exercises. In this case, assessor ratings will not show consistency, although there was behavioral consistency per se. We know very little about the schemas assessors use to interpret candidate behavior and to make trait judgments (Lievens & Klimoski, 2001; Zedeck, 1986). One notable exception is a recent study by Lance et al. (2004). They compared different models of assessor cognitive processes and discovered that assessors mainly used an exercise-specific model for judging candidate behavior. If assessors use exercise-specific schemata when making trait judgments about candidates, these global schemata might well override candidate behaviors that were consistent across exercises. However, all of this remains speculation. To shed light onto these assessment center issues, future research should combine trait activation and trait perception models.

Second, in this study we examined trait activation within assessment centers. Therefore, one could describe our effort as an internal validation oriented approach. However, as argued by Tett and Burnett (2003), trait activation is a framework that applies to many assessment methods. Essentially, as long as assessment methods create the opportunity to observe similar trait-relevant behavior as assessment center exercises, one should expect these methods to obtain convergent results. Conversely, when various assessment methods do not lend themselves to observe similar trait-relevant behavior, divergent results should be expected. Therefore, an intriguing avenue for future studies consists of incorporating trait activation ideas when externally validating assessment center ratings with those from nonassessment center situations with similar activation potential. For example, one could correlate ratings of managers' interpersonal sensitivity from exercises that are high in TAP for Agreeableness with subordinates' ratings of managers' consideration or sensitivity (e.g., from multisource feedback). Evidence of construct validity could then be obtained if that correlation is higher than those found from other assessment center dimensions not relevant to Agreeableness. A similar theory-driven strategy could be followed when correlating assessment center exercises with construct-oriented situational judgment tests. To date, researchers have externally validated assessment center ratings without paying attention to trait activation.

Implications for Practice

Trait activation theory does not need to be reserved as a theoretical framework for understanding what is happening in assessment centers. If desired, researchers and practitioners should go even further and use trait activation theory as a useful prescriptive framework in assessment center design. Before presenting some examples, we want to emphasize that trait activation theory does not mean that assessors should directly rate traits and that dimensions should be removed from assessment centers. Organizations choose dimensions for a variety of reasons, only one of which is their representation of traits. An important advantage of dimensions is that they are often formulated in the language of work behavior, increasing their apparent relevance to management. In fact, dimensions capture acquired work skills (e.g., negotiation and organization skills) and are closely linked to job activities and organizations' competency models.

One way to use the logic of trait activation in practice concerns the development of exercises. In current assessment center practices, exercises are primarily developed to increase fidelity and criterion-related validity. Similarly, dimensions are based on job analysis. We are not proposing that these practices should be abandoned. However, trait activation theory should also play a role. For example, once job analysis has identified the dimensions to be measured, trait activation theory might be used to eliminate or combine dimensions within an exercise that seem to capture the

same underlying trait (e.g., "innovation" and "adaptability" are based on behaviors that might be expressions of Openness).

Another concrete example is that assessment center users might fruitfully build on trait activation theory when constructing role-player instructions. In current assessment center practice, role-players are typically given a specific list of things to do and to avoid. Role-players are also trained to perform realistically albeit consistently across candidates. Although these best practices have proven their usefulness over the years, a key function of trained role-players consists of evoking dimension-related behavior from candidates (Thornton & Mueller-Hanson, 2004). Trait activation might help identify which specific traits can be evoked by specific role-player stimuli (i.e., specific statements or actions).

Apart from implications on dimension selection, exercise design, and role-player instructions, trait activation theory has also implications regarding assessment center feedback. There has been some debate about whether assessment center feedback reports should be built around dimensions versus exercises (Thornton, 1999). When feedback is built around dimensions (e.g., "You score weak on resilience"), the advantage is that such dimension-specific feedback is relevant across a wide variety of situations. Yet this feedback assumes that these dimensions are indeed measured across many situations (exercises). Research shows this is often not the case. Conversely, feedback might also be built around exercises (e.g., "You score weak in the oral presentation"). This is in line with most of the research evidence showing that exercises capture much of the variance in assessment center ratings. Yet this feedback lacks depth as it generalizes to only one specific situation (one exercise). The interesting point is that trait activation theory takes a middle-of-the-road position between these two extremes. Specifically, trait activation theory suggests building feedback reports around the situations that activate the traits (e.g., "You score weak in situations where you are put under pressure"). Such a feedback approach is supported by construct-related evidence (see our results), while at the same time avoiding the specificity of the exercise-based approach as it refers to all exercises that activate Emotional Stability (namely oral presentations and competitive group discussions, see Table 2).

Conclusion

In sum, a large-scale investigation of the role of trait activation in assessment centers was conducted by aggregating results across numerous studies. Generally, support was found for the hypotheses, confirming that trait activation is a useful theoretical framework for understanding the within-person behavioral variations and consistencies that might partly explain the typical construct validity results obtained in assessment centers. This sheds a slightly more positive light on the construct validity puzzle in assessment centers and focuses attention on the meaning of dimensions and exercises as related to well-understood individual difference variables in order to understand the validity evidence. Although this was a large-scale investigation, it can be seen as just the beginning of research on the role of trait activation in assessment centers. Specifically, future studies should examine the role of trait activation in the mental models that assessors use to interpret candidate behavior, across various assessment methods, and in a prescriptive way to shape assessment center characteristics.

References

References marked with an asterisk indicate studies included in the meta-analysis.

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

*Arthur, W. A., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26,* 813–835.

*Atkins, P. W. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55,* 871–904.

*Becker, A. S. (1990). *The effects of a reduction in assessor roles on the convergent and discriminant validity of assessment center ratings.* Unpublished doctoral dissertation, University of Missouri, St. Louis.

*Bobrow, W., & Leonards, J. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality, 12,* 217–236.

*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.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organisational Psychology, 69,* 167–181.

*Chorvat, V. P. (1994). *Toward the construct validity of assessment center leadership dimensions: A multitrait–multimethod investigation using confirmatory factor analysis.* Unpublished doctoral dissertation, University of South Florida, Tampa.

Connolly, J. J., & Viswesvaran, C. (1998, April). *The convergent validity between self- and observer ratings of personality.* Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Conway, J. M. (1996). Analysis and design of multitrait–multirater performance appraisal studies. *Journal of Management, 22,* 139–162.

Costa, P. T., Jr., & McCrae, R. R. (1992). *The Revised NEO-PI/NEO-FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

*Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10,* 319–333.

*Fredricks, A. J. (1989). *Assessment center ratings: Models and process.* Unpublished doctoral dissertation, University of Nebraska.

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

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55,* 149–158.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with inter-judge agreement. *Journal of Personality and Social Psychology, 51,* 409–418.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology, 64,* 479–490.

Furnham, A., Crump, J., & Whelan, J. (1997). Validating the NEO Personality Inventory using assessor's ratings. *Personality and Individual Differences, 22,* 669–675.

Goffin, R. D., & Jackson, D. N. (1992). Analysis of multitrait–multirater performance appraisal data: Composite direct product method versus confirmatory factor analysis. *Multivariate Behavioral Research, 27,* 363–385.

Goffin, R. D., Rothstein, M. G., & Johnston, N. G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81,* 746–756.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26–42.

*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.

*Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78,* 675–678.

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

John, O. P., & Robbins, R. W. (1993). Determinants of inter-judge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521–551.

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

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception. *Psychological Bulletin, 116,* 245–258.

Kenny, D. A., Horner, C., Kashy, D. A., & Chu, L. (1992). Consensus at zero acquaintance: Replication, behavioral cues, and stability. *Journal of Personality and Social Psychology, 62,* 88–97.

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

*Lance, C. E., 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.

*Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13,* 323–353.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152.

Lievens, F. (2001). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 65,* 1–19.

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.

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

Lievens, F., De Fruyt, F., & Van Dam, K. (2001). Assessors' use of personality traits in descriptions of assessment center candidates: A 5-factor model perspective. *Journal of Occupational and Organizational Psychology, 74,* 623–636.

*Lievens, F., & Harris, M. M., Van Keer, E., & Bisqueret, C. (2003). Predicting cross-cultural training performance: The validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *Journal of Applied Psychology, 88,* 476–489.

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

*Lievens, F., & Van Keer, E. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organisational Psychology, 74,* 373–378.

*Lovler, R. L., Rose, M., & Wesley, S. (2002, April). *Finding assessment center construct validity: Try behaviors instead of dimensions.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252–253.

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.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

*Parker, M. W. (1992). *A construct validation of the Florida Principal Competencies Assessment Center using confirmatory factor analysis.* Unpublished doctoral dissertation, University of South Florida, Tampa.

*Pittman, S. (1998). *An examination of construct validity within an assessment center.* Unpublished doctoral dissertation, George Mason University, Fairfax, VA.

*Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organisational Psychology, 70,* 103–108.

*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.

*Sweeney, D. C. (1976). *The development and analysis of rating scales for the Chicago Police Recruit Assessment Center.* Unpublished manuscript, Bowling Green State University, OH.

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.

Thornton, G. C., III. (1992). *Assessment centers and human resource management.* Reading, MA: Addison Wesley.

Thornton, G. C., III. (1999, April). *Reactions to attribute- versus exercise-based feedback in developmental assessment centers.* Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93,* 239–257.

*Van der Velde, E. G., Born, M. P., & Hofkes, K. (1994). Begripsvalidering van een assessment center met behulp van confirmatorische factoranalyse [Construct validity of an assessment center using confirmatory factor analysis]. *Gedrag en Organisatie, 7,* 18–26.

*Veldman, W. M. (1994). *Assessment centers and candidates' personal qualities: A study on the correlations between assessment center ratings.* Unpublished doctoral dissertation, Vrije Universiteit Amsterdam.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315–321.

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.

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

(*Appendixes follow*)

Appendix A

Classification Scheme for Dimensions

1. Communication: The extent to which an individual conveys oral and written information and responds to questions and challenges. (p. 133)

2. Consideration and awareness of others: The extent to which an individual's actions reflect a consideration for the feelings and needs of others as well as an awareness of the impact and implications of decisions relevant to other components both inside and outside the organization. (p. 133)

3. Drive: The extent to which an individual originates and maintains a high activity level, sets high performance standards and persists in their achievement, and expresses the desire to advance to higher job levels. (p. 134)

4. Influencing others: The extent to which an individual persuades others to do something or adopt a point of view in order to produce desired results and takes action in which the dominant influence is one's own convictions rather than the influence of others' opinions. (p. 134)

5. Organizing and planning: The extent to which an individual systematically arranges his or her work and resources as well as that of others for efficient task accomplishment and the extent to which an individual anticipates and prepares for the future. (p. 135)

6. Problem solving: The extent to which an individual gathers information; understands relevant technical and professional information; effectively analyzes data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; uses available resources in new ways; and generates and recognizes imaginative solutions. (p. 135)

7. Tolerance for stress and uncertainty: The extent to which an individual maintains effectiveness in diverse situations under varying degrees of pressure, opposition, and disappointment. (p. 136)

Note. These definitions come from Arthur et al.'s (2003) article. Page citations are shown in parentheses.

Appendix B

Classification Scheme for Exercises

1. In case analysis, the participant is given material to read that describes an organizational problem and is then asked to prepare a written set of recommendations or an action plan for higher management. The problem may require financial, system, or process analysis.

2. In a competitive leaderless group discussion, four to eight participants are given one or several problems to resolve in a fixed period of time, usually one hour. They are given the task to discuss the problems and arrive at a solution that provides them with the best solution for them individually. For example, each participant in a group of four individually prepared and presented an idea for allocating a grant. Then, all four participants discussed the ideas and decided how to distribute the money among them. Participants were given the individual goal of earning the most possible money for their project.

3. In a cooperative leaderless group discussion, four to eight participants are given one or several problems to resolve in a fixed period of time, usually one hour. They are given the task to discuss the problems and arrive at a group solution. So participants have to work as a team to develop the best solution by pooling and sharing information. For example, participants had to develop a proposal for solving a joint task. They were given some identical and some nonidentical information that they had to assess together and integrate in order to solve the problem.

4. An in-basket is a simulation of the paperwork that arrives in the mailbox or on the desk of the typical manager. It might include a large volume of memos, letters, reports, announcements, requests, and irrelevant information that present personnel, financial, accounting, or procedural problems for the manager. The participant is given a calendar, background information, general instructions, and paper and pencil for response. The participant must write out instructions, draft letters, make decisions, and set up meetings, all within a relatively short time period. The time pressures force the participant to set priorities and make decisions.

5. In an oral presentation exercise, participants are asked to deliver a presentation on a specific topic, case study, and so forth. Candidates are given a short time period (e.g., 30 min) to study the topic or case study and to prepare the presentation. The presentation is usually given to an assessor or group of assessors, who then ask questions intended to challenge the participant.

6. In role-plays, participants typically talk with someone playing the role of a subordinate, colleague, or customer. The participant must talk with the problem subordinate and find a solution to the problem. The role-player might ask questions, make suggestions, answer questions, and even act upset depending on what the situation calls for.

Note. The items above are adapted from Thornton's (1992) article.

Received June 8, 2004
Revision received February 9, 2005
Accepted February 18, 2005 ■