Human Performance, 22:375–390, 2009

ISSN: 0895-9285 print/1532-7043 online
DOI: 10.1080/08959280903248310

Routledge
Taylor & Francis Group

The Importance of Exercise and Dimension Factors in Assessment Centers: Simultaneous Examinations of Construct-Related and Criterion-Related Validity

Filip Lievens
Ghent University

Stephan Dilchert and Deniz S. Ones
University of Minnesota

This study presents a simultaneous examination of multiple evidential bases of the validity of assessment center (AC) ratings. In particular, we combine both construct-related and criterion-related validation strategies in the same sample to determine the relative importance of exercises and dimensions. We examine the underlying structure of ACs in terms of exercise and dimension factors while directly linking these factors to a work-related criterion (salary). Results from an AC ($N = 753$) showed that exercise factors not only explained more variance in AC ratings than dimension factors but also were more important in predicting salary. Dimension factors explained a smaller albeit significant portion of the variance in AC ratings and had lower validity for predicting salary. The implications of these findings for AC theory, practice, and research are discussed.

In assessment centers (ACs), trained assessors observe and evaluate candidates' behaviors in job-related exercises. Although the makeup of these high-fidelity simulations differs considerably across ACs, they can be brought back to five generic types: in-baskets, case analyses, role-plays, oral presentations, and group discussions (Bowler & Woehr, 2006; Lievens, Chasteen, Day, & Christiansen, 2006). Another hallmark of ACs is that candidates are rated on job-related dimensions. Recent taxonomic work has grouped this variety of dimensions into six broad psychological constructs: communication, consideration/awareness of others, drive, influencing others, organizing and planning, and problem solving (Arthur, Day, McNelly, & Edens, 2003).

Over the last years, these exercise and dimension taxonomies have served as useful frameworks for directing research on the validity of AC ratings. One stream of research has examined the ability of AC ratings to predict *external* work-related outcomes such as job performance, promotion, managerial potential, and salary. A meta-analysis of this strand of studies showed that ratings on the AC constructs of organizing and planning, problem solving, and influencing others emerged as the most valid predictors of job performance (Arthur et al., 2003). Another voluminous stream of studies has focused on the *internal* structure of AC ratings and investigated whether AC ratings indeed reflect candidates' standing on dimensions as measured by multiple exercises (Bowler & Woehr, 2006; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens &

Correspondence should be sent 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

Conway, 2001). According to the most recent meta-analysis of this research stream (Bowler & Woehr, 2006), exercises (especially the in-basket and oral presentation exercises) accounted for the largest portion of the variance in ratings, even though AC dimensions (especially communication and influencing others) also explained substantial amounts of variance.

So far, these two research streams have evolved separately from each other. That is, studies scrutinizing the underlying structure of AC ratings as well as linking these underlying components to prediction in the same study are nonexistent. Woehr and Arthur (2003) summarized this lack of integration among the two research streams as follows.

> A cursory examination of the literature suggests that studies examining assessment center construct-related validity and those examining criterion-related validity are largely independent. Thus, one important question with respect to the assessment center validity paradox is how many individual studies have demonstrated a lack of construct-related validity while also demonstrating criterion-related validity for a specific assessment center application? (p. 234)

In line with this, a common thread running through a recent series of AC review papers was that future research would profit from using broad validation designs. Specifically, these articles (Arthur, Day, & Woehr, 2008; Howard, 2008; Lance, 2008; Rupp, Thornton, & Gibbons, 2008) have criticized the reliance on the use of the multitrait–multimethod (MTMM) matrix approach as the *sole* validation strategy, as the assumptions underlying the MTMM approach are not realistic in the case of AC exercises and dimensions.

The current lack of integration of both construct-related and criterion-related validation designs leaves several key questions in the AC domain unanswered. For example, we do not know whether specific dimensions (e.g., problem solving) that account for variance in AC ratings are also good predictors when compared to other dimensions (e.g., consideration for others)? A similar question can be asked for AC exercises. Although we know that exercises represent true cross-situational performance variation (e.g., Lance et al., 2000), we do not know whether specific exercises (e.g., oral presentation) that explain most of the variance in AC ratings are also the best predictors of work-related criteria. Fundamentally, we do not know whether the dimension and exercise factors that explain a substantial amount of variance in AC ratings are also the best predictors of work-related criteria. Clearly, answering such questions would increase our understanding of which AC components (exercises and/or dimensions) enable ACs to predict important work-related criteria.

This study aims to advance AC research by combining both construct-related and criterion-related approaches in the same sample to determine the relative importance of exercises and dimensions. Specifically, we examine the underlying structure of ACs in terms of latent exercise and dimension factors while linking these factors to a work-related criterion (salary). The next sections review prior studies in each of these two independent streams of AC research. We also elaborate why an investigation of multiple evidential validity bases is of paramount conceptual and practical importance for ACs.

STUDY BACKGROUND

Modern conceptualizations of the validation process regard validation as a form of hypothesis testing (Binning & Barrett, 1989; Landy, 1986). Validation is viewed as a process of gathering var-

ious lines of evidence that should contribute to a better understanding of the meaning of a set of test scores and their inferences. According to these modern conceptualizations, construct validity is proposed as the central and unifying concept, with content-related validity, criterion-related validity, and construct-related validity being possible strategies for collecting construct validity evidence (Woehr & Arthur, 2003). Thus, this unitarian conceptualization highlights the importance of evaluating the validity of test scores on the basis of multiple evidential bases instead of on the basis of single coefficients.

As noted earlier, the different streams of validity research in the AC domain have followed largely independent paths. The first and earliest stream of research focused primarily on the criterion-related validity of AC ratings (see the Management Progress Study, Bray & Grant, 1966). This is understandable as ACs are more costly and labor intensive to develop and administer than most other selection tools and processes. The meta-analysis by Gaugler, Rosenthal, Thornton, and Bentson (1987) of early AC studies showed that overall assessment ratings were good predictors of several criteria such as job performance, potential, training performance, and career advancement (e.g., salary). A recent meta-analysis by Hermelin, Lievens, and Robertson (2007) confirmed that the overall assessment rating is a good predictor of supervisor ratings of job performance. Another meta-analysis (Arthur et al., 2003) examined the validity of final ratings on the six major AC dimension categories (see earlier) for predicting job performance. Ratings on the dimensions of organizing and planning, problem solving, and influencing others emerged as the most valid predictors.

A second stream of research has tried to answer the question whether the dimension ratings do measure the constructs they are purported to measure. This internal construct-related stream of research focused on the dimensional ratings that assessors make after *each* exercise (so-called within-exercise dimension ratings). These ratings were arranged in what resembles an MTMM matrix. Next, the amount of variance in AC ratings attributable to exercises and dimensions was estimated. Recently, Bowler and Woehr (2006) used meta-analytic methods to combine the existing matrices of correlations among post-exercise dimension ratings into one single matrix. The best fit was obtained for a model with correlated dimensions and exercises. Exercises explained most of the variance (33%). Especially, the in-basket and presentation exercises accounted for large parts of variance. Dimensions also explained a substantial amount of variance (22%). In addition, some dimensions (i.e., communication, influencing others, organizing and planning, and problem solving) explained significantly more variance than others (i.e., consideration/awareness of others, drive).

Recently, the focus on the internal construct-related validity of AC ratings and especially the use of the MTMM approach as the only validation strategy of AC ratings have been criticized in a series of AC review articles (Arthur et al., 2008; Howard, 2008; Lance, 2008; Rupp et al., 2008). These articles cogently argued that the criteria of the MTMM approach are overly stringent as AC exercises cannot be considered alternate measures and AC dimensions are often not stable traits. In other words, in this view exercises are not mere alternative measurement methods but rather represent different performance situations and behavior related to dimensions is not traitlike in the sense that it should be expected to be cross-situationally consistent (see Lance, Baranik, Lau, & Scharlau, 2009). A common thread running through these various AC articles is that future research would benefit from investing in broader validation strategies. That is, combining an investigation of the internal structure of AC ratings with criterion-related and external construct-related validation strategies (e.g., Arthur et al., 2008, p. 110).

The specific call to complement the construct-related validity approach (the so-called internal validation strategy; Schwab, 1980) with a criterion-related validity approach (external validation strategy) is not new. In fact, articles on the internal construct-related validity of ACs typically end by stating that determining the amount of exercise/dimension variance in AC ratings reflects only one side of the equation (Haaland & Christiansen, 2002; Kolk, Born, & van der Flier, 2002; Lievens & Conway, 2001; Robie, Osburn, Morris, Etchegaray, & Adams, 2000; Woehr & Arthur, 2003). Many of these researchers have called for an investigation of how much variance exercises and dimensions explain in an *external* criterion such as job performance or extrinsic career success (i.e., salary, career progression).

So far, only some studies have attempted to combine these validation designs in the same sample. In one group of studies, AC ratings were linked to external criteria in a correlational design (Chan, 1996; Fleenor, 1996; Henderson, Anderson, & Rick, 1995; Jansen & Stoop, 2001). For example, Chan validated an AC for police officers in Singapore ($N = 46$). He found little evidence that the dimensions were actually measured as exercise variance was predominant. In addition, the AC was not predictive of job performance. However, it predicted promotions received. Jansen and Stoop investigated the validity of a Dutch AC ($N = 679$). Results showed that dimensions were poorly measured and only a few had substantial correlations with salary progression. Henderson et al. validated an AC for graduates in the United Kingdom ($N = 311$). The structure underlying the AC ratings could again be primarily explained by the exercise component of ACs. In terms of prediction, only final ratings on one dimension (adaptability) had a significant positive relationship with job performance. Similar conclusions were drawn by Fleenor in a developmental AC for public sector managers ($N = 102$). Thus, in three of these four studies, weak evidence of dimensions being measured was paired with weak prediction.

Another set of studies has examined the correlations between AC ratings and external criteria via structural equation modeling. In particular, Lance and colleagues (Lance et al., 2000; Lance et al., 2004; Lance et al., 2007) have reported correlations between latent exercise factors and a general latent performance factor on one hand and external criteria such as cognitive ability, personality, job knowledge, and job performance on the other. Results showed that exercise factors were differentially related to these external criteria (job knowledge, cognitive ability, and job performance but not personality), whereas the general performance factor was consistently related to all of these criteria. Thus, these studies have advanced our understanding of the model with one general performance factor and exercise factors. However, they have not explicitly compared the criterion-related and construct-related validity evidence of AC dimension factors versus exercise factors, which is the focus of this study.

Taken together, our review of these prior streams of AC research attests to the relative lack of integration in AC validity research. First, previous criterion-related validity research has focused on final dimension ratings, typically ignoring the importance of the exercise component of ACs. Second, previous construct-related validity research has shown that dimension factors are actually less important contributors than exercise factors. However, these studies have neglected to investigate the relationship of latent exercise *and* dimension factors with external work-related criteria (job performance, promotion, salary, etc.). Third, in studies that examined both the construct-related and the criterion-related validity, these validity investigations were run independently. Indicators of internal construct-related validity evidence (dimension and exercise factors) were not linked to external criteria.

The objective of this study is to complement an internal validation strategy with an external validation strategy. Hence, we investigate the relative contribution of dimensions versus exercises both with an internal and external criterion. Accordingly, we are able to *directly* link internal construct-related validity indicators (i.e., the amount of exercise and dimension variance) to an external criterion (i.e., extrinsic career success). In this way, we are able to examine whether those latent dimension and exercise factors that explain a substantial amount of variance in AC ratings are also the most valid ones in terms of criterion-related validity.

Such a joint investigation of construct-related and criterion-related validity is of conceptual importance. Conceptually, this endeavor might provide broader evidence that speaks to the issue of whether exercises, dimensions, or both deserve their place in ACs. If exercises (and not dimensions) explain most of the variance in AC ratings *and* if this exercise variance (instead of dimension variance) is also strongly correlated with external work-related criteria, this would provide a much broader evidential basis to view exercises rather than dimensions as the cornerstones of ACs. At a practical level, our study therefore might provide a stronger evidential basis for making future modifications to AC design (e.g., dimension-based vs. task-based job analysis, dimension-based vs. exercise-based feedback).

METHOD

Sample

A sample of 765 individuals applying for middle management positions attended the AC. Of these candidates, 78.9% were male; the majority of the sample (89.8%) was of Caucasian descent. The mean age was 42.8 years ($SD = 7.1$); the majority of candidates (> 88.7%) held university degrees. Candidates applied for jobs in various industries, the largest proportions being diverse manufacturing industries (24%; e.g., electrical, food, heavy manufacturing), retail (13%), and professional and health care jobs (10% and 9%, respectively). More than 75% of candidates were seeking jobs at relatively large organizations (> 1,000 organizational members). Apart from the AC, the selection procedure usually consisted of cognitive ability tests, personality inventories, and interviews. A large consultancy firm provided access to the AC data.

Assessment Center Dimensions and Exercises

Four exercises were used in the AC. Across these exercises, participants were evaluated across four AC dimensions. We included only those dimensions in the analyses that were measured in multiple exercises (see dimension-exercise matrix in Table 1). The four AC dimensions assessed across exercises corresponded directly to four of the six overarching AC dimensions in the taxonomy of Arthur et al. (2003). These dimensions were problem solving, organizing and planning, influencing others, and consideration/awareness of others. Definitions of these dimensions are given in the appendix.

The exercises used to assess dimensions were typical for many ACs. First, an in-basket presented a number of items (e-mails, memos, voice messages) to candidates who were asked to respond to each of them and to create plans for addressing the top three issues over the next 6

TABLE 1
Dimension by Exercise Matrix

	In-Basket	*Direct Report Meeting*	*Task Force*	*Presentation*
Problem solving	X	X	X	X
Organizing and planning	X	X	X	
Influencing others	X	X	X	X
Consideration/Awareness of others	X	X	X	

months. Second, a role-play was conducted in the form of a direct report meeting with a regional sales manager. Candidates took the role of the sales manager's superior. The objectives were to address the subordinate's performance and to convince the subordinate to adopt an alternate sales strategy. Third, candidates were asked to give a strategy presentation to their superior. Finally, another role-play simulated a task force meeting in which candidates assumed the role of group leader. The team was charged with resolving a problem. As both peers of the candidates (played by assessors) displayed varying degrees of commitment to resolving this issue, candidates needed to ensure that consensus was reached regarding the steps required for short-term and long-term resolution of the problem.

Rating Process

Generally, the rating process adhered to standard AC practices and guidelines. Experienced consultants, most of whom were psychologists, served as assessors. All assessors had previously undergone the AC themselves and had attended a comprehensive training seminar that lasted several days. Training content included general information on psychological assessment as well as detailed explanations of the specific dimensions assessed in the ACs. Assessors familiarized themselves with the scenarios in all exercises, including the purpose of simulations, background information available, information about their role, and the scripts to which they have to adhere. Subsequently, assessors in training observed seasoned assessors in playing the respective roles in each simulation, and practiced their roles with other assessors. Additional practice was then coupled with feedback from a primary trainer. The training also provided practice in observing, recording, and scoring assessee behavior. Assessors practiced scoring until they reached a predetermined level of agreement with an experienced assessor in rating behavior in the simulations and exercises. In addition, calibration sessions were held at regular intervals.

Candidates were scored on each dimension using behaviorally anchored rating scales. For each exercise and dimension, different behaviorally anchored rating scales were used. As only one assessor was rating candidate's behavior at any given time, it was not possible to compute interrater reliability. Consistent with current AC practices, candidates were rated by different assessors across all exercises. After completion of all exercises, assessors met to discuss their observations and ratings with one another; however, data were integrated using a mechanical procedure allowing for only minor adjustments after these discussions.

External Criterion Measure: Salary

This study employed managerial salary level as an external criterion for evaluating the validity of AC exercises and dimensions. Our use of managerial salary as the criterion has ample precedent in primary and secondary validity studies of ACs (Bray, Campbell, & Grant, 1974; Hinrichs, 1978; Jansen & Stoop, 2001; Jansen & Vinkenburg, 2006; Lievens & Van Keer, 2005; Mitchel, 1975; Tziner, Ronen, & Hacohen, 1993). Some advantages of using salary as a criterion include data collection ease and freedom from self-enhancement bias (Heslin, 2003). We gathered salary data using self-reports, as is most often the case in research conducted in across-organization samples. Prior research has shown that self-reports and organizational records of earnings are virtually identical when gathered for research purposes. For example, Judge, Cable, Boudreau, and Bretz (1995) reported an average deviation of only 1% between the two types in a sample of 1,338 executives.

Although job-related, salary level should not be regarded as a direct indicator of job performance (cf. Hilton & Dill, 1962), primarily because many factors influencing salary are outside of the direct control of individual employees. This renders salary a contaminated indicator of job performance. Salary is also a deficient measure of performance as it does not capture all relevant performance dimensions (Campbell, Dunnette, Lawler, & Weick, 1970). Yet we should note that the strength of the salary–job performance link can be expected to increase where performance-based compensation systems are instituted. For managers, though, the correspondence is less than perfect as performance evaluations tend to be related to salary as well as salary increases (Dyer, Schwab, & Theriault, 1976; Lawler, 1966). On the basis of commonly hypothesized moderators of predictor-salary relationships in the psychological and economics literature, this study controlled for tenure, gender, job type, and industry type in computing concurrent criterion-related validities for managerial salary (see next).

Given the nature of earnings data (see earlier), we prefer to conceptualize salary as an objective indicator of extrinsic career success, and not job performance per se. Extrinsic career success "refers to outcomes that are both instrumental rewards from the job or occupation" (Seibert & Kraimer, 2001, p. 2) and includes easily observable outcomes such as salary and promotions (Greenhaus, Parasuraman, & Wormley, 1990; Judge et al., 1995). Indeed, these two outcomes are the two most frequently used indicators of extrinsic career success (e.g., Judge et al., 1995; Ng, Eby, Sorensen, & Feldman, 2005; Seibert & Kraimer, 2001). Using and interpreting salary as an indicator of extrinsic career success also makes sense from a conceptual viewpoint as individuals rely on salary to evaluate their objective career success and career planning decisions (Harrell, Harrell, McIntyre, & Weinberg, 1977; Weinstein & Srinivasan, 1974).

ANALYSES AND RESULTS

Descriptive Statistics

Descriptive statistics and correlations among the within-exercise dimension ratings are presented in Table 2. The mean same dimension–different exercise correlation was .17, whereas the mean different dimension–same exercise correlation equaled .41. The mean different dimension–differ-

TABLE 2
Descriptive Statistics and Correlations Among Assessment Center Ratings

	M	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
In-basket															
1. Problem solving	2.76	.57													
2. Organizing and planning	2.56	.61	.53												
3. Influencing others	2.68	.66	.47	.60											
4. Consideration/Awareness of others	3.06	.63	.34	.34	.37										
Direct report meeting															
5. Problem solving	3.07	.62	.19	.14	.10	.06									
6. Organizing and planning	2.75	.64	.12	.11	.08	.05	.53								
7. Influencing others	3.01	.66	.13	.11	.13	.04	.47	.57							
8. Consideration/Awareness of others	3.00	.78	.10	.04	–.06	.09	.36	.17	.01						
Task force															
9. Problem solving	3.14	.71	.20	.14	.10	.10	.16	.12	.07	.13					
10. Organizing and planning	2.49	.62	.11	.06	.06	.06	.14	.17	.12	.09	.52				
11. Influencing others	2.96	.69	.05	.05	.10	.01	.10	.09	.18	–.02	.47	.55			
12. Consideration/Awareness of others	2.96	.74	.15	.07	.04	.19	.14	.11	.03	.26	.44	.32	.11		
Presentation															
13. Problem solving	2.98	.71	.21	.12	.13	.05	.22	.17	.14	.05	.24	.19	.15	.11	
14. Influencing others	3.10	.60	.14	.14	.10	.00	.18	.16	.20	.04	.15	.21	.25	.08	.56

Note. $N = 765$.

ent exercise correlation was .10. *T* tests showed that the different dimension–same exercise correlations were significantly higher than the same dimension–different exercise correlations, $t(17) = 5.79$, $p = .00$, which were in turn significantly higher than the different dimension–different exercise correlations, $t(17) = 3.06$, $p = .01$.

Underlying Structure of Assessment Center Ratings

We tested several models that represented different conceptualizations of ACs (see Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001, for reviews). First, we tested a *dimensions-only* model. In this model, ACs were conceived as measuring stable individual differences constructs, reflecting the traditional "personalist" perspective in personality psychology. Second, we tested an *exercises-only* model. In this "situationist" model, exercises are the building blocks of ACs that are then conceptualized as a series of miniaturized work samples designed to elicit job-relevant behavior (Jackson, Stillman, & Atkins, 2005; Lowry, 1997; Robertson, Gratton, & Sharpley, 1987; Sackett & Dreher, 1982). Such a model is also consistent with empirical research showing that exercise effects in ACs represent true cross-situational variability of candidates across exercises, and not simply unwanted method variance (Lance, Foster, Gentry, & Thoresen, 2004; Lance et al., 2000). Third, a model with *one general dimension and correlated exercises* was specified (Lance et al., 2000). In this model, it is posited that assessors are not able to distinguish among the various dimensions. Conceptually, this model builds on general impression models prevalent in performance appraisal (Lance, Foster, et al., 2004). The fourth model was a *combination model containing both latent exercise and dimension factors* (e.g., Donahue, Truxillo,

Cornwell, & Gerrity, 1997; Kudisch, Ladd, & Dobbins, 1997). The underlying rationale for this "interactionist" model is that ACs aim to measure multiple job-related dimensions in multiple job-related exercises.

To test the relative fit of these models through confirmatory factor analysis, we employed EQS (Bentler, 1995). Maximum likelihood estimation was used. We relied upon several fit indices to assess how each model represented the data. In particular, the comparative fit index (CFI), Tucker Lewis Index (TLI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA) were used. These goodness-of-fit measures were suggested by Hu and Bentler (1999). Their extensive simulation study evaluated the adequacy of cut-off values based on the criterion that the adequate cut-off values should result in minimum type I and type II errors. On the basis of this study, Hu and Bentler proposed the following cutoff values: .95 (minimum values for CFI and TLI), .08 (maximum value for SRMR), and .06 (maximum value for RMSEA).

Results of the confirmatory factor analyses (CFAs) are presented in Table 3. The best fit was obtained for the *correlated exercises and correlated dimensions model*, with a CFI of .973 and a RMSEA of .044. This model was also not plagued by estimation problems (e.g., improper estimates). So this model serves as basis for our study. Table 4 presents summary parameter estimates of this model, showing that exercise loadings were higher than dimension loadings.

Internal Structure of Assessment Center Ratings: Latent Exercise and Dimension Factors

Table 5 presents the squared parameter loadings associated with dimensions and exercises for the *correlated exercises and correlated dimensions* model. Latent exercise factors on average accounted for 41% of the variance, whereas latent dimension factors on average accounted for 15% of the variance. Among the latent exercise factors, the *oral presentation* factor explained most of the variance (55%). Among the latent dimension factors, the *consideration/awareness of others* factor accounted for most of the variance (21%). So although omnibus fit results indicated that assessor ratings are best represented by a combination of dimensions and exercises, parameter fit results showed that exercise variance is larger than dimension variance. These findings are in line with the most recent meta-analysis of AC construct-related validity (Bowler & Woehr, 2006).

TABLE 3
Summary Results of Confirmatory Factor Analysis Models

Model	χ^2	*df*	*CFI*	*TLI*	*SRMSR*	*RMSEA and 90% CI*	*Proper Solution*
Correlated dimensions only	1746.100	71	.415	.250	.131	.176 [.169–.183]	No
Correlated exercises only	432.430	71	.874	.838	.056	.082 [.074–.089]	Yes
One dimension and correlated exercises	171.518	57	.960	.936	.030	.051 [.042–.060]	Yes
Correlated dimensions and correlated exercises	127.624	51	.973	.952	.029	.044 [.035–.054]	Yes

Note. $N = 765$. CFI = comparative fit index; TLI = Tucker Lewis Index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation; CI = confidence interval.

TABLE 4
Summary of Parameter Estimates for Correlated Dimensions and Correlated Exercises Model

Source of the Rating	Dimension Factors				Exercise Factors				
	PS	*OP*	*IO*	*CA*	*IB*	*DRM*	*TF*	*P*	*U*
IB									
PS	.19				.66				.73
OP		.03			.79				.62
IO			.10		.75				.65
CA				.16	.46				.87
DRM									
PS	.51					.63			.58
OP		.19				.77			.61
IO			.24			.72			.66
CA				.65		.12			.75
TF									
PS	.25						.89		.39
OP		.65					.53		.55
IO			.71				.48		.52
CA				.42			.42		.81
P									
PS	.13							.82	.57
IO			.24					.66	.71
	Dimension Intercorrelations				*Exercise Intercorrelations*				
PS	1.00				1.00				
OP	.39	1.00			.20	1.00			
IO	.26	.65	1.00		.18	.10	1.00		
CA	.73	.28	–.12	1.00	.22	.26	.27	1.00	

Note. In EQS uniquenesses are not variances. Standardized solution is presented here. PS = problem solving; OP = organizing and planning; IO = influencing others; CA = consideration/awareness of others; IB = in-basket; DRM = direct report meeting; TF = task force; P = presentation; U = Uniqueness.

Criterion-Related Validity: Latent Exercise and Dimension Factors

Next, we linked the dimension and exercise factors to an external work-related criterion (extrinsic career success – salary). To this end, we specified an additional latent factor for the criterion and added covariances between this latent factor and the latent exercise and dimension factors. In this analysis, we controlled for tenure, gender, job type, and industry type by partialling out these variables from the input matrix (Fletcher, Selgrade, & Germano, 2006; Kammeyer-Mueller & Wanberg, 2003). As this information was missing for some candidates, the sample size was reduced to 520 for these analyses. The main question was whether latent exercise factors would also be more valid predictors than latent dimension factors. Table 6 presents the results broken down by exercises and dimensions. The mean correlation between the latent exercise factors and salary level was .24, whereas the mean correlation between the latent dimension factors and salary was .16. All four latent exercise factors were consistently related to salary (as indicated by the fact that

TABLE 5
Squared Parameter Loadings of Dimensions and Exercises for the Correlated Exercises and Correlated Dimensions Model

Exercises	
In-basket	.46
Direct report meeting	.38
Task force	.37
Oral presentation	.55
M	.41
SD	.22
Dimensions	
Problem solving	.09
Organizing and planning	.15
Influencing others	.16
Consideration/Awareness of others	.21
M	.15
SD	.18

Note. $N = 765$.

TABLE 6
Confirmatory Factor Analyses Derived Criterion-Related Validity Estimates of the Exercise and Dimension Factors for Predicting Salary

	r	*CI*
Exercises		
In-basket	.20	.13–.27
Direct report meeting	.19	.12–.26
Task force	.11	.03–.18
Oral presentation	.48	.42–.53
Adj. *R*	.49	
Adj. R^2	.24	
Dimensions		
Problem solving	.18	.11–.25
Organizing and planning	.15	.08–.22
Influencing others	.08	.01–.15
Consideration/Awareness of others	.21	.14–.28
Adj. *R*	.22	
Adj. R^2	.05	

Note. $N = 520$. *r* = correlation between salary and dimension/exercise factors as estimated by the CFA model; CI = 95%, two-tailed confidence interval; Adj. *R* = adjusted multiple correlation for exercises and dimensions predicting salary, respectively (based on unit-weights).

the respective 95% confidence intervals did not include zero), with the *oral presentation* being the most valid one (.48). All four dimensions were also valid predictors of salary. *Consideration/awareness of others* (.21) and *problem solving* (.18) were the most valid dimensions. We also computed multiple correlations and squared multiple *R* s to compare the predictive power of the four latent exercise factors to that of the four latent dimension factors as a set. We unit-weighted corre-

lations across exercises and dimensions (by using their respective average intercorrelations in the multiple regressions) to obtain stable estimates that do not capitalize on sample specific weighting schemes (Schmidt, 1971). The four exercise factors together explained more variance in salary than the four latent dimension factors (adjusted *R*s = .49 and .22, respectively).

Incremental Validity: Latent Exercise and Dimension Factors

To examine the incremental validity of latent exercise and dimension factors over and above one another in a structural equation modeling framework, we conducted nested comparison tests. The full model including covariances between the latent salary factor and all latent exercise and dimension factors was compared to a model with only covariances between salary and latent exercise factors and to a model with only covariances between salary and latent dimension factors. The full model obtained $\chi^2 = (57) = 90.81$, CFI = .983. Removing the exercise factor covariances with salary from the full model significantly decreased fit, difference in $\chi^2(4) = 67.66$, $p < .01$, ΔCFI = .03. The same significant fit decrease was obtained when removing the dimension factor covariances with salary, although the decrease was smaller, $\chi^2(4) = 16.78$, $p < .01$, ΔCFI = .01. So, the incremental validity of exercises over dimensions was much higher than vice versa.

DISCUSSION

This study aimed to bridge two major AC research streams that have evolved apart from one other, namely, research on construct-related and criterion-related validity of AC ratings. To this end, we directly linked indicators of construct-related validity (i.e., the amount of exercise and dimension variance) to an external criterion indicating extrinsic career success (salary). Our simultaneous examination of multiple bases of validity evidence in an AC provides several new key insights to the AC field.

As a first contribution, we found that the criterion-related validity of exercise factors was higher than that of dimension factors for predicting salary. This result extends previous research showing that exercises explain more variance than dimensions in AC ratings. By showing that exercises not only explain the largest share of variance in AC ratings but also are most predictive of an external criterion, this study provides a broader evidential basis to establish exercises as the main cornerstones of ACs. However, this is not to say that the dimensions as underlying constructs of AC ratings should be ignored. Dimensions explain a significant albeit smaller portion of the variance in AC ratings. In addition, three latent dimension factors were significant predictors of salary.

A second contribution of our simultaneous examination of multiple evidential bases of validity is the conclusion that evidence of internal construct-related validity appears to be coupled with evidence of criterion-related validity. This convergence of validity evidence is in line with the unitarian framework of validity (Binning & Barrett, 1989; Landy, 1986). Generally, exercises explained more variance and were better predictors compared to dimensions. On the exercise level, the *exercise* that explained most of the variance in AC ratings (oral presentation) was also the most valid one in terms of criterion-related validity. On the dimension level, the *dimension* that explained most of the variance (consideration/awareness of others) was also the most valid one in terms of criterion-related validity. This correspondence between the loadings and the validity coefficients

might be explained by the fact that scales that are measured with less measurement error (as shown by higher loadings) will typically have higher validities.

However, there was also some discrepancy between the different forms of validity evidence. For example, two exercises with very similar loadings (.38 for direct report meeting and .37 for task force) had notably different criterion-related validities, namely, .19 and .11, respectively. Another example is that a dimension such as problem solving had a low loading (.09) but was still a significant predictor (.15). Therefore, measurement error cannot be the only explanation for validity differences. On a more general level, these results fit well with current explanations for the lack of construct validity in ACs. In a recent review, Woehr and Arthur (2003) postulated two explanations. First, they argued that "assessment center design, implementation, and other methodological factors may add measurement error" that leads to poor construct measurement (p. 234). As a second explanation, Woehr and Arthur proposed a "construct misspecification hypothesis," stating that some of the unexplained variance in dimensions might be due to unspecified constructs operating at a deeper level (e.g., self-monitoring, Extraversion, communication apprehension). Research should be conducted to test this construct misspecification hypothesis.

A third important contribution of our study is methodological in its nature: We used CFA to obtain unconfounded estimates of exercise and dimension variance in AC ratings. Hence, we were able to disentangle exercise and dimension variance because CFA partitions the variance into dimension, exercise, and error variance. Clearly, it would have been much easier to compute the zero-order correlation between the external criterion (salary) and final dimension and exercise ratings, respectively. Yet observed final dimension ratings are summary ratings of dimensions *across exercises* (e.g., Goffin, Rothstein, & Johnston, 1996). Likewise, observed final exercise ratings are summary ratings of exercises *across dimensions*. Accordingly, in the specific case of ACs, each of these observed final dimension ratings always confounds exercise variance with dimension variance. Therefore, observed final exercise/dimension ratings are not the right yardstick to obtain estimates of the contribution of exercises and dimensions in predicting external criteria.

In terms of future research, we urge other researchers to also gather multiple lines of evidence examining the validity of ACs. As construct-related and criterion-related validity research in the AC domain have largely gone separate ways, it is time to invest in validation designs that yield both types of validity evidence simultaneously and from the same sample. In this study, we combined an internal construct-related validity design with a criterion-related validity design. As suggested by several scholars (Arthur et al., 2008; Lance, 2008; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007), future studies might also combine internal and external construct-related validity designs. This means that the internal structure of AC ratings (latent dimension and exercise factors) is linked to constructs measured with other predictor measures (personality inventories, cognitive ability tests, etc.) in a nomological network.

Broad validation designs are also needed to evaluate the impact of AC design interventions. On the basis of prior research, several design considerations have been suggested for increasing the construct-related validity of ACs (see Lievens, 1998; Lievens & Conway, 2001; Woehr & Arthur, 2003). Consider, for instance, the suggestion to limit the number of exercises. Whereas a more diverse set of exercises seems to reduce the convergence of dimension ratings across exercises (i.e., lower convergent validity; Lievens, 1998; Schneider & Schmitt, 1992), the opposite might be true for criterion-related validity (i.e., a more diverse set of job-related exercises might increase criterion-related validity). We need to test these predictions by integrating different validity designs in the future.

Future research should also aim to broaden the criterion measures employed in validating ACs. In this study, we relied on salary as the external work-related criterion. Although salary as a measure of extrinsic career success has a long tradition in general validation research and AC research, it also has its limitations. Therefore, we controlled for various extraneous sources affecting salary (age, gender, type of job, type of industry). Yet, it should be acknowledged we did not control for organization. Future research should replicate our results using other criteria (job performance and its facets) and with other dimensions and exercises.

In conclusion, Landy and Conte (2004) recently characterized the research endeavors related to AC construct validity as follows: "Decomposing the assessment center into its constituent elements and asking which part makes the greatest contribution is like decomposing a bouillabaisse and asking which ingredient made it taste so good" (p. 146). We concur with this culinary simile when only internal construct-related validity evidence is available, as was the case in most prior AC research. However, when an external criterion is available, it is possible to determine which AC component is related to its predictive success. This study showed that both the internal construct-related validity and criterion-related validity of latent exercise factors is higher than that of latent dimension factors. To further advance our understanding of ACs and their validity our study should spur future research to investigate multiple bases of AC validity.

REFERENCES

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.

Arthur, W., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1*, 105–111.

Bentler, P. M. (1995). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

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

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

Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives*. New York: Wiley.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monograph, 80,* 1–27.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill.

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

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior & Personality, 12*, 85–108.

Dyer, L., Schwab, D. P., & Theriault, R. D. (1976). Managerial perceptions regarding salary increase criteria. *Personnel Psychology, 29*, 233–242.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319–335.

Fletcher, T. D., Selgrade, K. A., & Germano, L. M. (2006, May). *On the use of partial covariances in structural modeling*. Poster session presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Gaugler, B. B., Rosenthal, D. R., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validities. *Journal of Applied Psychology Monograph, 72,* 493–511.

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.

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal, 33*, 64–86.

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.

Harrell, M. S., Harrell, T. W., McIntyre, S. H., & Weinberg, C. B. (1977). Predicting compensation among MBA graduates five and ten years after graduation. *Journal of Applied Psychology, 62*, 636–640.

Henderson, F., Anderson, N., & Rick, S. (1995). Future competency profiling. *Personnel Review, 24*, 19–31.

Hermelin, E., Lievens, F. & Robertson, I. T. (2007). The validity assessment centers for the prediction of supervisory ratings: A meta-analysis. *International Journal of Selection and Assessment, 15,* 405–411.

Heslin, P. A. (2003). Self- and other-referent criteria of career success. *Journal of Career Assessment, 11*, 262–286.

Hilton, T. L., & Dill, W. R. (1962). Salary growth as a criterion of career progress. *Journal of Applied Psychology, 46*, 153–158.

Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63*, 596–601.

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

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: an Interdisciplinary Journal, 6*, 1–55.

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

Jansen, P. G. W., & Stoop, B. A. M. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology, 86*, 741–753.

Jansen, P. G. W., & Vinkenburg, C. J. (2006). Predicting management career success from assessment center data: A longitudinal study. *Journal of Vocational Behavior, 68*, 253–266.

Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz, R. D. (1995). An empirical investigation of the predictors of executive career success. *Personnel Psychology, 48*, 485–519.

Kammeyer-Mueller, J. D., & Wanberg, C. R. (2003). Unwrapping the organizational entry process: Disentangling multiple antecedents and their pathways to adjustment. *Journal of Applied Psychology, 88*, 779–794.

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325–338.

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may no be so troubling after all. *Journal of Social Behavior & Personality, 12*, 129–144.

Lance, C. E. (2008). Where have we been, how did we get there, and where shall we go? *Industrial and Organizational Psychology, 1*, 151–157.

Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis)application of the multitrait-multimethod methodology in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in organizational and social research* (pp. 339–362). New York: Routledge.

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., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345–362.

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

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

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Landy, F. J., & Conte, J. M. (2004). *Work in the 21st century.* New York: McGraw-Hill.

Lawler, E. E. (1966). Managers' attitudes toward how their pay is and should be determined. *Journal of Applied Psychology, 50*, 273–279.

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., 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., & 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., & Van Keer, E. (2005). Assessment centers in Belgium: The results of a study on their validity and fairness. *Psychologie du Travail et des Organisations, 11*, 25–33.
Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior & Personality, 12*, 53–62.
Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60*, 573–579.
Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success. A meta-analysis. *Personnel Psychology, 58*, 367–408.
Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13*, 355–370.
Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60*, 187–195.
Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology, 1*, 116–120.
Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401–410.
Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*, 699–714.
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.
Schwab, D. P. (1980). Construct validity in organizational behavior. In L. L. Cummings & B. M. Staw (Eds.). *Research in Organizational Behavior (Vol. 2)*: 3–43. Greenwich, CT.
Seibert, S. E., & Kraimer, M. L. (2001). The Five-Factor Model of personality and career success. *Journal of Vocational Behavior, 58*, 1–21.
Tziner, A., Ronen, S., & Hacohen, D. (1993). A four-year validation study of an assessment center in a financial corporation. *Journal of Organizational Behavior, 14*, 225–237.
Weinstein, A. G., & Srinivasan, V. (1974). Predicting managerial success of Master of Business Administration (MBA) graduates. *Journal of Applied Psychology, 59*, 207–212.
Woehr, D. J., & Arthur, W. (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.

APPENDIX

TABLE A1
Definitions of Assessment Center Dimensions

Dimension	*Definition*
Problem solving	Approaches issues from a broad perspective, considering a wide range of information and factors; grasps complexities and perceives relationships among problems or issues.
Organizing and planning	Organizes and prioritizes work activities; delegates responsibility; monitors progress.
Influencing others	Steps forward to address difficult issues; stands firm on behalf of the organization and key stakeholders.
Consideration/Awareness of others	Initiates and develops relationships with a wide variety of people based on trust; shows interest in and understanding of others' needs and concerns.