1-2009

Assessment centres: A tale about dimensions, exercises, and dancing bears

Filip LIEVENS

Singapore Management University, filiplievens@smu.edu.sg

DOI: https://doi.org/10.1080/13594320802058997

Published in European Journal of Work and Organizational Psychology, 2009 January, Volume 18, Issue 1, Pages 102-121
http://doi.org/10.1080/13594320802058997

Assessment centres: A tale about dimensions, exercises, and dancing bears

Filip Lievens
Ghent University, Ghent, Belgium

This study reviews prior construct-related validity research in assessment centres. Special focus is placed on disentangling possible explanations for the construct-related validity findings. The conclusion is that we now have a much better picture of the reasons behind the construct-related validity findings. Careful assessment centre design and high interrater reliability among assessors seem necessary albeit insufficient conditions to establish assessment centre construct-related validity. The nature of candidate performances is another key factor. This study next discusses how these empirical findings have changed how assessment centres are conceptualized (theoretical advancements framed in the application of trait activation theory), analysed (methodological advancements), and designed (practical advancements).

Keywords: Assessment centres; Construct validity; Constructs; Multitrait-multimethod matrix; Trait activation theory.

More than 25 years ago, Robinson (1981) published a remarkable article in *Personnel Psychology*. This article was remarkable for two reasons. First, his article described the meticulous development of a content-oriented selection procedure for selecting a construction superintendent in a small business setting. The selection procedure consisted of various work samples and assessment centre exercises such as a blueprint reading task, a "scrambled subcontractor" task, a construction error recognition task, and a scheduling task. Second, the final sample size of this study was the smallest possible as

Correspondence should be addressed to Filip Lievens, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@ugent.be

This article is based on the invited presentation given as winner of the Distinguished Early Career Contributions Award at the Annual Conference of the Society for Industrial and Organizational Psychology (New York, April 2007). I would like to thank Liesbet De Koster for her help with the article.

$N = 1$. This was because one person was hired (of the 17 individuals who applied).

This article still serves as a prime example of the behavioural consistency approach to personnel selection (see also Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). In addition, it perfectly illustrates the features that continue to make assessment centres (ACs) an attractive procedure for selecting and developing managerial talent. As high-fidelity simulations, assessment centre exercises focus on actual candidate behaviour, which is observed and evaluated by trained assessors on various dimensions in multiple job-related situations (exercises). On the basis of these components, assessment centres enjoy a good psychometric record. On average, assessment centres have good criterion-related validity, ranging from .25 to .39, depending on the dimension measured (Arthur, Day, McNelly, & Edens, 2003). The interrater reliability is found to be moderate to high (.60–.90), depending on the level of experience and the training of assessors (Lievens, 2002a; Thornton & Rupp, 2005). Assessment centres further demonstrate good utility (Hoffman & Thornton, 1997) and little adverse impact (Terpstra, Mohamed, & Kethley, 1999). Finally, assessees react positively to the procedure (Hausknecht, Day, & Thomas, 2004).

Despite these positive features, a recurring issue in the assessment centre domain deals with the lack of evidence of assessment centres to measure the constructs (dimensions) they are purported to measure (see Bowler & Woehr, 2006; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens & Conway, 2001, for large-scale reviews). As so-called ''method-driven'' predictors, assessment centres are by definition measurement methods that might assess a plethora of constructs. Many research studies, however, show that there is weak evidence for the convergent and discriminant validity of these constructs.

This potential lack of construct-related validity has various implications. From a theoretical point of view, it is important to determine whether the two cornerstones (dimensions and exercises) of assessment centres are indeed represented by the AC ratings. If this is not the case, perhaps the assessment centre framework should be changed (Lowry, 1997). From a research perspective, it is important to establish evidence for the dimensions as building blocks of assessment centres because relationships between assessment centres and relevant job criteria are often based on final dimension ratings. For example, the most recent meta-analysis of assessment centres (Arthur et al., 2003) focused on the criterion-related validity of assessment centre dimensions. For these results to be meaningful, it is important that there is indeed evidence that these final dimension ratings measure the dimensions consistently across the various exercises. The same argument applies when final dimension ratings are placed in a nomological network with other predictors such as cognitive ability tests or personality

inventories. At a practical level, it is important to ascertain that dimensions are measured in assessment centres because the developmental feedback given to candidates is typically formulated around dimensions. So, if the dimensions are not valid indicants of the managerial abilities, the feedback and subsequent action plans could have detrimental effects. The following example by Kudish, Ladd, and Dobbins (1997) exemplifies these practical ramifications:

> Telling a candidate that he or she needs to improve his or her overall leadership skills may be inappropriate if the underlying construct being measured is dealing with a subordinate in a one-on-one situation (i.e., tapping individual leadership as opposed to group leadership). (p. 131)

Given the far-reaching importance of this construct-related validity issue, this study aims to provide an overview of prior construct-related validity research in assessment centres. Specific attention is paid to disentangling possible explanations for the findings. In addition, we aim to show how the thinking about the construct-related validity of assessment centres has substantially shifted over the years. In particular, we delineate the theoretical, methodological, and practical implications of this paradigm shift.

THE CONSTRUCT-RELATED VALIDITY FINDINGS: OVERVIEW OF PRIOR RESEARCH

The discrepancy between AC theory and empirical evidence was first reported by Sackett and Dreher (1982). Given that AC theory emphasizes the dimensions (besides the exercises) as key cornerstones of assessment centres, correlations between ratings on these dimensions across exercises are expected to be higher than correlations between ratings within exercises. Sackett and Dreher, however, reported exactly the opposite findings. They investigated assessment centre ratings in three organizations. In each of these organizations, they found low correlations among ratings of a single dimension across exercises (i.e., weak convergent validity) and high correlations among ratings of various dimensions within one exercise (i.e., weak discriminant validity). Furthermore, factor analyses indicated more evidence for exercise factors than for dimension factors. Although these findings seemed ''troublesome'', the authors underlined that this does not mean that assessment centres lack construct-related validity. Sackett and Tuzinski (2001) again cautioned for this misinterpretation of the basic findings, noting: ''Assessment centres do not lack 'construct validity,' but rather lack clear consensus as to the constructs they assess'' (pp. 117–188).

The findings described have proven to be very robust as they were found in both selection and developmental assessment centres. In addition, they have been found in assessment centres conducted all over the world. Apart from the USA (Bycio, Alvares, & Hahn, 1987; Harris, Becker, & Smith, 1993; Joyce, Thayer, & Pond, 1994; Kudisch et al., 1997; Reilly, Henry, & Smither, 1990; Sackett & Dreher, 1982; Schneider & Schmitt, 1992; Silverman, Dalessio, Woods, & Johnson, 1986), these results have been established in the United Kingdom (Anderson, Lievens, van Dam, & Born, 2006; Crawley, Pinder, & Herriot, 1990; Robertson, Gratton, & Sharpley, 1987), Germany (Kleinmann & Koller, 1997; Kleinmann, Kuptsch, & Koller, 1996), Belgium (Lievens & van Keer, 2001), France (Borteyrou, 2005; Rolland, 1999), Australia (Atkins & Wood, 2002), New Zealand (Jackson, Stillman, & Atkins, 2005), China (Wu & Zhang, 2001), and Singapore (Chan, 1996).

As a side note, it is remarkable that these findings of situation-specific variance being larger than construct variance are not unique to assessment centres. Similar results have also been obtained for other method-driven predictors such as structured interviews (Conway & Peneno, 1999; van Iddekinge, Raymark, Eidson, & Attenweiler, 2004) and situational judgement tests (Trippe & Foti, 2003). For example, in the interview field, convergence between the same construct measured by different types of structured interviews (behaviour description and situational interviews) has been low. Moreover, the findings seem to extend to all fields wherein different constructs are measured in multiple performance-based exercises. For example, predominance of situation-specific variance over construct variance has been found in studies about patient-management problems for physicians (e.g., Julian & Schumacher, 1988), military examinations (e.g., Shavelson, Mayberry, Li, & Webb, 1990), hands-on science tasks (e.g., Shavelson et al., 1991), bar examinations (e.g., Klein, 1992), and direct writing assessments (e.g., Dunbar, Koretz, & Hoover, 1991). One exception was a physiotherapy study wherein three physical examination skills (palpation, tension irritation, and touch irritation) were assessed on three anatomic sites (elbow, ankle, and shoulder) (Stratford et al., 1990). This study found that the ratings of the same skill across anatomic sites equalled .65, whereas ratings across skills for the same site equalled .41. So, in this particular study, the convergent validity coefficients are indeed higher than the discriminant validity coefficients.

Returning to the assessment centre field, three studies have been conducted to quantitatively summarize the construct-related validity findings. First, Lievens and Conway (2001) reanalysed 34 multitrait–multimethod (MTMM) matrices of assessment centre ratings. Their main conclusion was that a model consisting of exercises (specified as correlated uniquenesses) and dimensions represented the best fit to the data. In this

model, exercises and dimensions explained the same amount of variance (34%). In addition, dimensions were found to correlate substantially (.71).

A second quantitative review came to different conclusions (Lance, Lambert, et al., 2004). According to Lance, Lambert, et al., Lievens and Conway's (2001) results of exercises and dimensions explaining about the same amount of variance were due to a statistical artefact (i.e., the use of the correlated uniqueness model that systematically overestimated dimension variance). In their reanalysis, a model with correlated exercises and one general dimension prevailed. In addition, exercise variance (52%) was clearly more important than dimension variance (14%).

Recently, Bowler and Woehr (2006) conducted a third quantitative review because a limitation inherent in the two prior quantitative reviews was that each MTMM matrix was individually reanalysed. Hence, estimates of dimension and exercise variance were based on CFA results from models with different sample sizes, dimensions, and exercises. Bowler and Woehr used meta-analytical methods to combine 35 MTMM matrices into one single matrix. Therefore, the CFA results from this meta-analytically derived MTMM matrix represent probably the best available estimates of exercise and dimension variance. The best fit was obtained for a CFA model with correlated dimensions and exercises. Exercises explained most of the variance (33%), although 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 of others, drive). Similar to the Lievens and Conway (2001) study, dimensions were found to correlate highly (.79).

This study differs from these prior studies in that it is not a quantitative review of assessment centre construct-related validity. Instead, we aim to delve deeper into the explanations given for the construct-related validity findings. To this end, the three explanations that were originally provided in the seminal study of Sackett and Dreher (1982) will serve as anchors. Specifically, we review recent empirical research that has directly or indirectly tested these three explanations. Sackett and Dreher presented the following three explanations: (a) poor assessment centre design might lead to assessor biases, (b) exercise variance confounds exercise and assessor variance, and (c) assessees might behave cross-situationally inconsistently across the exercises.

After scrutinizing the viability of these explanations by confronting them with recent empirical studies, we discuss how the findings of these recent studies have changed how we conceptualize (theoretical advancements), analyse (methodological advancements), and design (practical advancements) assessment centres.

EMPIRICAL ADVANCEMENTS IN UNRAVELLING THE EXPLANATIONS

Explanation 1: Poor assessment centre design

According to the first explanation, the construct-related validity findings result from poor assessment centre design, prompting assessor biases and inaccuracies. For example, according to this explanation, poorly designed assessment centres (e.g., inadequate training of assessors, asking assessors to rate a large number of dimensions) might result in assessors being prone to halo bias when rating the candidates, which in turn might lead to strong exercise factors. Woehr and Arthur (2003) excellently summarized this explanation by noting that "assessment centres as measurement tools are probably only as good as their development, design, and implementation" (p. 251). Many studies have taken this explanation as a point of departure for "fixing" potential flaws in assessment centre design. In fact, this explanation has been the dominant explanation in the assessment centre field for quite some time. Table 1 summarizes quantitative and qualitative review studies about the effects of assessment centre design changes on construct-related validity results. As can be seen in Table 1, there are two distinct perspectives within this first explanation. According to the first perspective, the main problem lies within assessors' limited information processing capacity. Therefore, interventions should be developed to support assessors in their observation and evaluation activities. For example, limiting the number of dimensions or using behavioural checklists can ease this process. The second perspective argues that assessors might use

TABLE 1
Summary of reviews examining the effects of design considerations on the quality of construct measurement in assessment centres

	Lievens (1998)/ Sackett and Tuzinski (2001)	*Lievens and Conway (2001)*	*Woehr and Arthur (2003)*
Limited cognitive capacity perspective			
Limit the number of dimensions	++	++	+
Use behavioural checklists	=	+	
Reduce assessor–assessee ratio			–
Make dimensions transparent	+	=	
Schema-based perspective			
Use psychologists as assessors	++	++	
Train assessors longer		–	++
Use frame-of-reference training	++		+
Use task-based "dimensions"	=		
Use within-dimension ratings	++		++

incorrect schemas for categorizing the information observed, which might lead to assessor biases. Consequently, assessors should be trained to apply more correct schemas. According to this view, using psychologist assessors and frame-of-reference training might lead to better results.

Although this explanation has generated a large strand of studies, it is not without problems. First, even well-designed assessment centres have exhibited weak evidence of construct-related validity. For example, the analyses of Schneider and Schmitt (1992) revealed that most of the variance in assessment centre ratings was explained by exercises instead of dimensions, even though they carefully implemented various recommendations for improving construct-related validity (e.g., limiting the number of dimensions, using behavioural checklist, and providing thorough assessor training). Similarly, Chan's (1996) rigorous assessment centre design did not improve construct-related validity. These results can be explained by the fact that many of these design considerations have only small effects. That is, they do not change the basic pattern that discriminant validity coefficients are higher than convergent validity coefficients.

Second, one might question whether all of these design interventions have beneficial effects on the criterion-related validity. In fact, most of these design interventions have been tried out in assessment centres conducted in laboratory settings where candidates are typically rated on the basis of a limited number of exercises and dimensions. Yet, limiting the number of dimensions might detract from the criterion-related validity of actual assessment centres in the field. A similar argument might be made with respect to making the dimensions transparent to candidates in selection assessment centres (Kleinmann et al., 1996).

Explanation 2: Exercise variance confounds exercise and assessor variance

As a second explanation, it has been posited that exercise variance is so large because it represents not only variability across exercises but also variability across assessors. In other words, exercise variance is basically a confounding of exercise and assessor variance. This confounding is due to the common practice of assessors rotating through the various exercises. Indeed, to save costs, a given assessor does not evaluate each candidate in each exercise.

Two research studies have directly tested this explanation (Kolk, Born, & van der Flier, 2002; Robie, Osburn, Morris, Etchegaray, & Adams, 2000). In both studies exercise variance was separated from assessor variance by asking one assessor to rate only one dimension per exercise. In line with this second explanation, this rating method led in both studies to more evidence for dimension factors. However, it should be noted that the discriminant validity coefficients were still higher than the convergent validity coefficients.

This is conceivable in light of the fact that interrater reliability among assessors has typically been satisfactory (Thornton & Rupp, 2005). Hence, controlling for assessor variance has only marginal effects. Thus, this explanation is at best a partial explanation for the construct-related validity results typically established. Another problem with this explanation is that it is very costly to use a large number of assessors in practice.

Explanation 3: Cross-situationally inconsistent assessee performances

According to a third explanation, the construct-related validity findings are due to candidates who behave cross-situationally inconsistently across structurally different exercises. Thus, exercises are not conceptualized as parallel measures. That is, an assessment centre is composed of several exercises which are carefully selected to cover specific job-related competences. Consequently, they place different psychological demands on the assessees. For instance, one might expect an assessee to behave differently—even inconsistently—in a one-to-one role-play as compared to in a group discussion.

Research that has tested the viability of this third explanation is relatively scarce. In three studies, Lance and colleagues (Lance, Foster, Gentry, & Thoresen, 2004; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Lance et al., 2000) correlated exercise factors with external variables such as job performance and cognitive ability. They hypothesized that if exercise factors constituted unwanted method bias, exercise factors and performance criteria should be unrelated. Conversely, when exercise effects turned out to reflect true cross-situational specificity of performance, positive relations between exercise factors and performance criteria should emerge. Their results confirm that exercise factors do not represent unwanted method bias. Instead, they reflect true performance differences. In another study, Hoeft and Schuler (2001) tried to estimate the amount of variability in assessment centre performance. In line with the third explanation, their study revealed that assessment centre performance was more situation-specific (57%) than situation-consistent (43%). They also found that candidates performed more consistently on some dimensions than on others. In particular, Activity (53%) and Oral Communication (55%) were the most consistently rated dimensions across exercises.

Note that this third explanation is radically different from the other two. Whereas the first two explanations put the "blame" on assessment centre design or lack of interrater reliability among assessors, this third explanation focuses on candidate performances. Assessors are viewed as relatively accurate. All of this highlights the key importance of disentangling these rival explanations. Unfortunately, few studies have put these three

explanations to the test. One exception is Lievens (2002b). In this study, both the effects of type of assessee performances and type of assessor were examined. In particular, three types of assessors (I/O psychologists, managers, and students) were asked to rate assessees whose performances varied along two continua, namely cross-exercise consistency (i.e., relatively inconsistent vs. relatively consistent) and dimension differentiation (i.e., relatively undifferentiated vs. relatively differentiated). Their ratings were analysed for convergent validity, discriminant validity, and interrater reliability evidence. Results showed large differences in evidence for convergent and discriminant validity across assessee performances. In fact, convergent validity was established only for consistent performances across exercises, whereas discriminant validity was established only for differentiated performances across dimensions. Evidence for convergent and discriminant validity varied across type of assessor too, although these differences were smaller. In particular, evidence for discriminant and convergent validity was more clearly established with I/O psychologists and managers than with students. In addition, interrater reliability also played a minor role in establishing convergent and discriminant validity evidence.

This study has key implications for the construct-related validity puzzle. First, this study shows that careful assessment centre design and assessor reliability (high interrater reliability among assessors) are necessary but insufficient conditions for establishing evidence for convergent and discriminant validity. This is because the nature of assessee performance may be a limiting factor for obtaining construct-related validity evidence. Second, this study provides no support for the method (exercise) bias explanation for the assessment centre construct-related validity findings. Finally, assessors appear to be relatively accurate. Hence, we might compare the seemingly impossible task of assessors providing relatively accurate ratings to the seemingly impossible task of bears dancing. Or to cite Funder (1989):

> Somebody once said that what makes a dancing bear so impressive is not that it dances well, but that it dances at all. I am impressed by human judgments of personality for roughly the same reason—not because the judgments are perfect, but because in the face of the enormous difficulties it seems remarkable they manage to have an accuracy at all. (p. 212)

THEORETICAL IMPLICATIONS

As noted earlier, the knowledge that candidate performances were one of the main reasons behind the construct-related validity findings was a crucial empirical advancement. However, the next step became to better understand candidate performances. Indeed, it is important to know why candidates

perform inconsistently across exercises. In the past, this question was answered by referring to the person-situation debate in personality and social psychology (e.g., Highhouse & Harris, 1993; Lievens, 2002b).

Recently, trait activation theory has provided a more comprehensive theoretical explanation for the variability in candidate performances across different assessment centre exercises. Trait activation theory is an interactionist theory to explain behaviour based on responses to trait-relevant cues found in situations (Tett & Guterman, 2000). A key characteristic of trait activation theory is that situational similarity is described in a trait-like manner, namely through the notion of *trait activation potential* (i.e., the capacity to observe differences in trait-relevant behaviour within a given situation; Tett & Guterman, 2000). The trait activation potential of a given situation is primarily determined by the *relevance* and *strength* of that situation. A situation is considered relevant to a trait if it provides cues for the expression of trait-relevant behaviour. Apart from situation relevance, situational strength also impacts on the variability and consistency of behaviour (Mischel, 1973, 1977). In particular, strong situations involve unambiguous behavioural demands, resulting in few differences in reactions to the situation, whereas weak situations are characterized by more ambiguous expectations, enabling more variability in behavioural responses.

So, if organizations want to assess candidates on a dimension such as resistance to stress that is related to the trait of emotional stability, they must use exercises that put people in a situation that might activate behaviour relevant to the trait of interest (without rating this trait). An oral presentation with challenging questions might be a good example as this kind of situation is likely to evoke dimension-relevant behaviour. Other examples might be the inclusion of stringent time limits, sudden obstacles, or information overload in exercises. Trait activation theory also suggests that exercises should not represent too strong situations. If organizations design exercises with clearly defined tasks, there might be few options left open for the assessees. Therefore, organizations typically design exercises with a certain amount of ambiguity so that differences in how assessees tackle the situation can be elicited and observed.

Haaland and Christiansen (2002) were the first to examine the convergent validity of assessment centre ratings in light of trait activation theory. They conducted a small-scale investigation of a promotional assessment centre ($N = 79$). Their findings pointed out that convergence between dimension ratings in exercises that were judged to be high in trait activation potential was stronger than convergence between dimension ratings in exercises low in trait activation. So, these results provided support for the relevance of trait activation theory for understanding assessment centres. A recent reanalysis of 30 existing AC studies also confirmed the propositions of trait activation theory (Lievens, Chasteen, Day, & Christiansen, 2006). That is, convergence

was stronger between exercises that both provided opportunity to observe behaviour related to the same trait. Findings further showed that trait activation worked best for dimensions related to Extraversion and Conscientiousness. In addition, discrimination among ratings within exercises was better for dimensions that were not expressions of the same underlying traits.

So, trait activation theory provides a psychologically deeper look into the construct-related validity findings of ACs. Trait activation posits that we should expect only strong convergence among dimension ratings between exercises that are both high in trait activation potential for that trait. In addition, trait activation theory predicts only good discriminant validity coefficients between dimensions that do not share the same underlying trait.

METHODOLOGICAL IMPLICATIONS

In the past, the MTMM approach was typically used for examining the within-exercise ratings made in assessment centres. Over the last years, however, there is increased recognition that the assumptions underlying the MTMM approach are too stringent for assessment centres. As demonstrated in the section on empirical advancements, recent empirical studies have shown that exercise variance is not unwanted method variance. Similarly, the assumption that different methods (in this case assessment centre exercises) should always converge in their measurement of a specific construct seems untenable in light of the theoretical advancements generated by trait activation theory. Haaland and Christiansen (2002) succinctly summarized the current thinking about the use of the MTMM approach in assessment centre research by stating:

> We believe what needs to change is not the approach for evaluating construct validity, but the inference that something is wrong when situations (exercises) intended to be dissimilar do not converge more than dimensions that are understood to share relations with the same individual difference constructs. (p. 160)

Apart from the MTMM approach, many studies have also used confirmatory factor analysis (CFA) for examining the construct-related validity of assessment centres. In that case, a model comprised of correlated dimensions and correlated exercises is often tested. This model is attractive because it dovetails the two key components of the assessment centre framework. However, this CFA model typically has serious estimation problems, resulting in ill-defined solutions reflected by parameter estimates outside the permissible range, such as factor correlations greater than one (Bagozzi & Yi, 1991). Recent simulation research (Conway, Lievens,

Scullen, & Lance, 2004) has shown that this is less likely to occur with large samples and with many methods and constructs being measured in the MTMM matrix. This research has also demonstrated that the so-called correlated uniqueness model (Marsh, 1989) is not a good alternative to the CFA model consisting of correlated dimensions and correlated exercises as it inflates construct variance.

A limitation inherent in all analytical alternatives heretofore mentioned is that they recognize only two sources of systematic variance, namely dimensions and exercises. Exercise variance is then necessarily a combination of variance due to different exercises and variance due to different assessors (see the second explanation in the section on empirical advancements). We believe studies are needed which decompose variance according to the three main sources of variance in assessment centre ratings: dimensions, exercises, and assessors. Dimension, exercise, and assessor variance can only be disentangled if a fully crossed design is available. This means that multiple assessors (e.g., three assessors) evaluate each assessee in each exercise. In operational centres, this is sometimes difficult to accomplish. If such a fully crossed design is available, two analytical approaches can be fruitfully applied, namely generalizability analysis (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and hierarchical confirmatory factor analysis (Lance, Teachout, & Donnelly, 1992). Both approaches allow decomposing the observed variance into the three major sources of variance (exercises, dimensions, and assessors) that may affect assessment centre ratings. Although several studies have already used generalizability analysis for investigating assessment centre ratings (e.g., Arthur, Woehr, & Maldegen, 2000; Jackson et al., 2005; Kolk et al., 2002; Lievens, 2001b, 2002b), no studies have used hierarchical confirmatory factor analysis.

PRACTICAL IMPLICATIONS

Besides the empirical, theoretical, and methodological advancements, one might wonder whether the construct-related validity debate has had any impact on how practitioners are actually designing and implementing assessment centres. In other words, has this whole debate merely been a case of scientific navel-gazing or are practitioners indeed running assessment centres differently from what was done in the past? Our answer sways towards the latter. Substantial practical implications can be seen in three broad domains.

A first spin-off of the construct-related validity debate has been the increased emphasis on ''good'' assessment centre design. As noted earlier, one of the dominant research streams consisted of uncovering design considerations that might improve the quality of construct measurement

(see Table 1). Although it is now generally acknowledged that good assessment centre design is only one—albeit an important—part of the construct-related validity puzzle, this strand of studies has generated several important practical guidelines. One example has been the increased use of frame-of-reference training. Frame-of-reference training (Bernardin & Buckley, 1981) aims to impose a shared performance theory on assessors, providing them with common standards as a reference for evaluating assessee performance. Research demonstrated that this training approach, in comparison to other assessor training formats, resulted in higher discriminant validity, higher interrater reliability, and higher rating accuracy (Lievens, 2001a; Schleicher, Day, Mayes, & Riggio, 2002). Another design consideration deals with the type of assessors used. Research further showed that both the use of managers and psychologists yields advantages (Lievens, 2001a, 2002b; Sagie & Magnezy, 1997). Discriminant validity was higher when psychologists served as assessors, whereas higher accuracy was established with managers. These results corroborate the use of a mixed assessor pool consisting of both managers and psychologists.

Clearly, an increased emphasis on "good" design among assessment centre users is to be recommended because various design considerations have also been found to moderate the criterion-related validity of assessment centres (Gaugler, Rosenthal, Thornton, & Bentson, 1987), although this does not necessarily mean that all design considerations that increase construct-related validity will also augment criterion-related validity (see earlier).

Second, task-based assessment centres constitute another practical development of the construct-related validity research stream in assessment centres. Task-based assessment centres are composed of several work samples in which general exercise performances rather than complex constructs such as traits are assessed (Lowry, 1997). Until today, research is rather scarce. An exception is Jackson et al. (2005), who compared the psychometric characteristics of task-based assessment centres with those of traditional dimension-based assessment centres. Both models yielded similar psychometric qualities, although only the task-based model provided an acceptable fit for the data, suggesting that this model offers a better conceptualization of assessment centres.

Third, we believe the application of trait activation theory provides a window of opportunities for improving assessment centre practice. Although such trait activation "practice" has not yet been implemented, the potential implications of trait activation theory are straightforward. Trait activation theory does not need to be reserved as a descriptive theoretical framework. If desired, one should go even further and might use trait activation theory as a prescriptive framework in assessment centre design. Prior to presenting some examples, we want to emphasize that trait activation theory does *not* mean that assessors should directly rate

personality traits and that dimensions should be removed from assessment centres. 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 behaviour, 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 (Lievens, Sanchez, & de Corte, 2004).

One way to use the logic of trait activation in practice concerns the development of exercises (Lievens, 2008). In current assessment centre 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 behaviours that might all be expressions of Openness). Accordingly, trait activation might help in dimension selection and exercise design.

Another concrete example is that assessment centre users might fruitfully build on trait activation theory when constructing role-player instructions. In current assessment centre 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 behaviour from candidates (Thornton & Mueller-Hanson, 2004). Trait activation might help identifying which specific traits can be evoked by specific role-player stimuli (i.e., specific statements or actions).

Trait activation theory might also have implications regarding assessment centre feedback. There has been some debate about whether assessment centre feedback reports should be built around dimensions versus exercises (Thornton, Larsh, Layer, & Kaman, 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. However, such 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''), as suggested by proponents of the task-based assessment centre approach. Yet, this feedback lacks depth as it generalizes to only one specific situation (one exercise). The interesting point is that trait activation theory can be situated 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'').

CONCLUSION

This review demonstrates that our thinking about the construct-related validity puzzle in assessment centres has considerably evolved over the years. At the start, the findings were seen as ''troublesome'' and as the ''Achilles' heel of assessment centres''. In fact, in 1982, Sackett and Dreher concluded ''The findings suggest severe problems for assessment centres: In all three centres, method variance predominates'' (p. 406). Nowadays, the typical construct-related validity results are no longer seen as troublesome, as reflected in the statement of Lance, Foster, et al. (2004): ''There may be nothing wrong with assessment centre's construct validity after all'' (p. 23, see also Lance, 2008) or Ployhart (2006): ''Research may be close to solving the construct-related validity question for assessment centres'' (p. 881). Put briefly, one can state that the construct-related validity findings have essentially remained the same over the years. However, how we interpret them has substantially changed.

Another important empirical development is that we now have a much better picture of the reasons behind the construct-related validity findings. Specifically, we know that careful assessment centre design and high interrater reliability among assessors are necessary albeit insufficient conditions to establish assessment centre construct-related validity. The nature of candidate performances is another key factor. Only when candidates perform consistently across exercises and heterogeneously across dimensions can evidence of construct-related validity be established. This condition is not only difficult to accomplish in assessment centres, it is also often not to be recommended.

As a consequence of these empirical advancements, the other sections of this article demonstrated that we currently conceptualize, analyse, and design assessment centres differently than we did in the past. Table 2 illustrates this paradigm shift clearly by comparing our thinking about the key players in

TABLE 2
Description of paradigm shift in assessment centres

	Old paradigm	*New paradigm*
Dimensions	Stable traits	Conditional dispositions
Exercises	Parallel measures	Trait-relevant situational cues
Candidates	Stability	Variability
Assessors	Flawed	Relatively accurate

assessment centres over the years. As indicated in Table 2, dimensions measured in assessment centres are no longer seen as stable traits. Instead, they are conceptualized as conditional dispositions (Mischel & Shoda, 1995). This means that stable candidate performances on dimensions can be expected only when the exercises elicit similar trait-relevant situational cues. This also means that exercises are no longer viewed as parallel measures. Finally, assessors are no longer regarded as being fundamentally flawed raters but as individuals who might make relatively accurate ratings (provided that the assessor training and the assessment centre design were adequate).

Across the sections of this article, we have already mentioned various directions for future research. To end this article, we repeat the two directions that we see as most vital for advancing the field of assessment centres. First, we believe that trait activation theory should be used proactively. The critique that assessment centres are atheoretical is long overdue. Therefore, we need to take this chance of providing assessment centres with a stronger theoretical background. When we use trait activation in a more prescriptive way, the selection of dimensions, the design of exercises, and the development of feedback reports are only some of the components of the assessment centre framework that might be modified.

Second, future studies should use comprehensive validation designs. This means that construct-related *and* criterion-related validity should be examined in tandem (Woehr & Arthur, 2003). Such a broad validation design enables to integrate these research streams that have evolved apart from each other. In fact, it enables to ascertain whether specific assessment centre design interventions positively affect both construct-related *and* criterion-related validity. It also allows investigating how proportions of dimension versus exercise variance relate to criterion-related validity. For instance, do assessment centres with high exercise variance and low dimension variance show more criterion-related validity than assessment centres with low exercise variance and high dimension variance? Or is the highest criterion-related validity obtained for assessment centres that show both high exercise variance and high dimension variance?

REFERENCES

Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology*, *91*, 555–566.

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., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment confer 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.

Bagozzi, R. P., & Yi, Y. J. (1991). Multitrait–multimethod matrices in consumer research. *Journal of Consumer Research*, *17*, 426–439.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*, 205–212.

Borteyrou, X. (2005). *Intelligence, personnalité, mises en situation et prédiction de la réussite professionnelle: La construction d'un centre d'évaluation pour des officiers de marine*. Unpublished doctoral dissertation, Université Victor Segalen Bordeaux, Bordeaux, France.

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.

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 Organizational Psychology*, *69*, 167–181.

Conway, J. M., Lievens, F., Scullen, S. E., & Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, *11*, 535–559.

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

Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment center dimensions, personality and aptitudes. *Journal of Occupational Psychology*, *63*, 211–216.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289–303.

Funder, D. C. (1989). Accuracy in personality judgment and the dancing bear. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 210–223). New York: Springer-Verlag.

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

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.

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

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

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment*, *9*, 114–123.

Hoffman, C. C., & Thornton, G. C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*, 455–470.

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

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

Julian, E. R., & Schumacher, C. F. (1988, March). *CBT pilot examination: Results and characteristics of CBX*. Paper presented at the conference of the National Board of Medical Examiners on Computer-based Testing in Medical Education and Evaluation, Philadelphia.

Klein, S. P. (1992). *The effect of content area and test type on bar exam scores*. Paper presented at the National Conference of Bar Examiners.

Kleinmann, M., & Koller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality*, *12*, 65–84.

Kleinmann, M., Kuptsch, C., & Koller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review*, *45*, 67–84.

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*, 325–337.

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. (2008). Why assessment centers (ACs) don't work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 87–100.

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, A. (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.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, *77*, 437–452.

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. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255–264.

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, *22*, 203–221.

Lievens, F. (2002a). An examination of the accuracy of slogans related to assessment centres. *Personnel Review*, *31*, 86–102.

Lievens, F. (2002b). 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. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 117–120.

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., Sanchez, J. I., & de Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, *57*, 881–904.

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

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality*, *12*, 53–62.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335–361.

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

Mischel, W. (1977). The interaction of person and situation. In D. Magnussen & N. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*, 868–897.

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

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centers: Dimensions into exercises won't go. *Journal of Occupational Psychology*, *60*, 187–195.

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.

Robinson, D. D. (1981). Content-oriented personnel selection in a small business setting. *Personnel Psychology*, *34*, 77–87.

Rolland, J. P. (1999). Construct validity of in-basket dimensions. *European Review of Applied Psychology*, *49*, 251–259.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401–410.

Sackett, P. R., & Tuzinski, K. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.), *How people evaluate others in organizations* (pp. 111–129). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Schmitt, N., & Ostroff, C. (1986). Operationalizing the behavioral consistency approach: Selection test development based on a content-oriented strategy. *Personnel Psychology*, *39*, 91–108.

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.

Shavelson, R. J., Baxter, G. P., Pine, J., Yure, J., Goldman, S. R., & Smith, B. (1991). Alternative technologies for large-scale science assessment: Instrument of education reform. *School Effectiveness and School Improvement*, *2*, 1–8.

Shavelson, R. J., Mayberry, P., Li, W., & Webb, N. (1990). Generalizability of job performance measurements: Marine Corps rifleman. *Military Psychology*, *2*, 129–144.

Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology*, *39*, 565–578.

Stratford, P. W., Thomson, M. A., Sanford, J., Saarinen, H., Dilworth, P., Solomon, P., et al. (1990). Effect of station examination item sampling on generalizability of student performance. *Physical Therapy*, *70*, 31–36.

Terpstra, D. E., Mohamed, A. A., & Kethley, R. B. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment*, *7*, 26–34.

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, Larsh, S., Layer, S., & Kaman, V. S. (1999, May). *Reactions to attribute- versus exercise-based feedback in developmental assessment centers*. Paper presented at the Assessment Centers, 21st Century: New Issues, and New Answers to Old Problems symposium conducted at the 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: Lawrence Erlbaum Associates, Inc.

Thornton, G. C., III, & Rupp, D. E. (2005). *Assessment centers in human resource management*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Trippe, D. M., & Foti, R. J. (2003, April). *An evaluation of the construct validity of situational judgment tests*. Paper presented at the conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, *17*, 71–93.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376.

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.

Wu, Z. M., & Zhang, H. C. (2001). The construct validity and structure modeling of assessment center. *Acta Psychologica Sinica*, *33*, 372–378.

Original manuscript received February 2007
Revised manuscript received March 2008
First published online June 2008