IPA
International Journal of Psychology
Vol. 14, No.2, Summer & Fall 2020
PP. 218-247

Iranian **P**sychological
Association

A Critical Approach to the Construct -Related Validity of Assessment Centers: A Study of the Assessment Center of the National Iranian Oil Company

Majid Salimi, PhD*
Training and Development
Manager of the National Iranian Oil
Company
ma_salimi@atu.ac.ir

Hossein Shokrkon, PhD
Department of Industrial and
Organizational Psychology
Shahid Chamran University of
Ahvaz, Ahvaz, Iran

Mohammad Montakhab Yeganeh, PhD
Expert for Designing the Competency Model
of the National Iranian Oil Company

Received: 10/ 12/ 2020 Revised: 9/ 8/ 2021 Accepted: 14/ 8/ 2021
Doi: 10.22034/ijpb.2021.261549.1211
Dor: 20.1001.1.20081251.2020.14.2.8.2

The purpose of this study was to investigate the construct validity of the Assessment and Development Center of Managers in the National Iranian Oil Company. The research method is based on the applied purpose and in terms of data collection is descriptive-analytical. The statistical sample includes data obtained from the evaluation of 384 managers who were selected from the database of managers by simple random sampling. These data were analyzed based on the multi-trait-multi-method and the confirmatory factor analysis. The overall fit indices showed that the data fit the model very well. The examination of the individual parameter estimates indicates that while both dimension and exercise factors contribute to the ratings, the estimates for exercise parameters are substantially larger in all cases than estimates for dimension factors (mean parameter estimates of .39 and .68 for the dimension and exercise parameters, respectively). In addition, inter-correlations among the latent dimension factors ranged from 0.98 to 1.0, providing evidence of no discriminant validity across dimensions. Thus, these results support and are consistent with those obtained from the correlational analyses. To put it short, the exercise factors are the primary determinants of assessment center ratings.

Keywords: assessment centers, construct validity, discriminant validity, dimension, exercise

Assessment centers (ACs) are widely used for selection and development purposes. Usually, they consist of several exercises (e.g., role-plays, presentations, or group discussions) that simulate relevant job-related tasks in which participants' performance is repeatedly rated on different job-related performance dimensions (Kleinmann & Ingold, 2019). Ratings from the different exercises are then combined, resulting in overall dimension ratings, which represent candidates' overall performance for each of the different performance dimensions, or in an overall assessment rating (OAR), which represents candidates' overall performance across all the exercises and dimensions in the entire AC (Wirz, Melchers, Kleinmann, Lievens, Annen, Blum & Ingold, 2020).

Over the past several decades, assessment centers have enjoyed increasing popularity. They are currently used in numerous private and public organizations to assess thousands of people each year. The validity of assessment centers is undoubtedly partially responsible for their popularity. Evidence supporting the criterion-related validity of assessment center ratings has been consistently documented (Woehr & Arthur, 2003). In addition, content-related methods of validation are also regularly used in assessment center development in an effort to meet the professional and legal requirements. Evidence for the construct-related validity of assessment center dimensions, however, has been less promising. Specifically, assessment centers are designed to evaluate individuals on specific dimensions of job performance across situations or exercises. Research, however, has indicated that exercises rather than dimension factors emerge in the evaluation of assessees (Schneider & Schmitt, 1992). Previous research has shown that ratings from ACs predict future performance and show incremental validity beyond cognitive ability and personality (Sackett, Shewach & Keiser, 2017).

Assessment centers used to be valuable diagnostic tools—at times they were run as procedures integrating a broad methodological diversity (Schuler, 2008). The history of AC begins with the German, British, and Australian military officer selection efforts in the 1930s and 1940s (Thornton & Byham, 1982). Quoted from Lance (2008), the most commonly accepted date for the development of a historical frame of reference for this process goes back to the 1940s and the work of the Office of Strategic Services (OSS). The 1950s saw the adaptation of these assessment techniques to managers in the private sector in AT&T's Management Progress Study (Moses & Byham, 1977) and the Michigan Bell Operational Program (Dunnette, 1971). These early efforts combined personality assessment, business games, situational tests, intelligence testing, and interviews with assessments, often resulting in a dozen or more dimensions. More recently, ACs have moved away from their personality origins toward a primary emphasis on assessing candidate behavior in situational exercises according to the relevant behavior-related performance dimensions (Thornton & Byham, 1982).

Early assessment centers were designed primarily to do one thing. The initial Office of Strategic Services (OSS) center, the early AT&T operational centers, and the centers at IBM, Sohio, and Sears all focused on prediction, but their underlying operational rationale was not just to identify those who would succeed but, as important, to eliminate those likely to fail (Moses, 2008).

In contrast, the climate for conducting centers today has shifted significantly. Retention and development are often the primary use for this technique. The competition for talent means that there are far fewer candidates in most organizational pipelines, and assessment centers are frequently used as a reward rather than as a hurdle to overcome (Moses, 2008). As Lance (2008) quoted, two things happened along the way that changed the AC theoretical landscape: (a) the evolution of within-exercise ‘‘post-exercise dimension ratings’’ PEDRs (Sackett & Dreher, 1984) as an intermediate step in the evaluation process (Howard, 2008; Rupp et al., 2008) and (b) the equation of the resulting Dimensions _ Exercises ratings matrix with the multitrait–multimethod (MTMM) methodology ‘‘in which dimensions serve as traits and exercises as methods’’ (Sackett & Dreher, 1982).

Today, as Lance (2008) noted, the International Task Force on Assessment Center Guidelines (2000) considers a number of components essential in order for a process to be considered an AC, including job analysis to identify critical job performance elements, classification of candidate behaviors into meaningful categories or dimensions, use of multiple assessment techniques that measure critical behaviors, use of multiple trained assessors, and systematic procedures for recording, integrating, and summarizing candidates’ behaviors in a reliable and valid fashion. Often, ACs are designed to assess candidate performance on multiple performance dimensions as they are assessed in multiple exercises (Bowler & Woehr, 2006). The secret of

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

success was that this assessment center resulting in high predictive validity included tools adding incremental value to work sample kinds of tasks, for example, tests, interviews, and biographical questionnaires (Schuler, 2008).

There are some approaches to evaluation in ACs, for example, dimensional performance is most often rated only after the completion of all exercises (within-dimension rating method). In this approach, assessors describe participants' behavior in exercise reports that they read aloud in integration sessions; they rate the dimensions after all the reports are heard using any behavioral evidence that is relevant. In the other approach, after the completion of each exercise, assessors are expected to rate dimensions (within-exercise rating method or post-exercise dimension ratings (PEDRs)), in which assessors often use to form consensus-based final dimension ratings (as in the ‘‘within dimension’’ method), as well as at the end of the process and summary overall ratings (Howard, 2008; Lance, 2008; Arthur, Day & Woehr`s, 2008).

The crossing of dimensions as assessors assess in various exercises resembles a multitrait–multimethod (MTMM) design. In the context of ACs, the MTMM approach is operationalized such that dimensions are viewed as traits and exercises as methods. Indeed, this mapping of dimensions and exercises has provided the basis for this research on AC construct validity (Lance, 2008).

The Construct Validity Problem

Arthur, Day & Woehr (2008) noted that at a theoretical level, if a measurement tool demonstrates criterion-related and content-related validity evidence, as is widely accepted with ACs, then it should also be expected to demonstrate construct-related validity evidence (Binning & Barrett, 1989). ACs have well-documented criterion-related (e.g., Hardison & Sackett, 2004) and content validity (Thornton & Mueller-Hanson, 2004) but appear not to measure the constructs that were intended to be measured (i.e., dimensions) (Lance, 2008). So, because ACs do not appear to do so, we have the resultant AC construct-related validity paradox. This alleged paradox ‘‘is reflected in the idea that assessment center ratings demonstrate (a) content-related validity, (b) criterion-related validity, and (c) a lack of construct-related validity–’’ defined in terms of ‘‘–a lack of convergent and discriminant validity with respect to assessment center dimensions’’ (Arthur, Day & Woehr, 2008) as assessed in this quasi-MTMM framework (Lance, 2008). Also, some meta-analyses found negative correlations between validities and years of publication; that is, there is a rather continuous decline of assessment center validity over the past 40 years (Schuler, 2008). Historically, ACs have been designed with the intent of measuring behavioral dimensions, but according to Lance (2008), assessment centers (ACs), as they are often designed and implemented, do not work as they are intended to work. In other words, after a quarter of century of research, it is now clear that ‘‘exercises and not dimensions are the currency of assessment centers’’ (Howard, 1997). This is the crux of what has been called the AC construct validity problem (Lance, 2008).

In order to collect AC ratings for decision-making and research, most commonly, assessors observe candidates in each simulation exercise and then determine scores for each dimension once the exercise has been completed. These scores are known as post-exercise dimension ratings (PEDRs). Although there are other ways to combine and investigate AC ratings, the majority

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

of construct-related validity research used PEDRs as the unit of analysis. In addition, confirmatory factor analysis (CFA) remains one of the most popular techniques to use in AC construct-related validity research. But it is here that research has produced results that are often considered problematic for ACs that are designed with a dimension-based perspective in which AC designers and users target dimension-related information. Specifically, factor analytic studies typically found that most of the variance in PEDR scores is indicative of exercise factors and not of dimension factors (Buckett, Becker, Melchers & Roodt, 2020).

According to the traditional theory supporting AC architecture, dimensions represent relatively stable behavioral categories that should be (a) reasonably distinct within exercises and (b) reasonably consistent across exercises (Woehr & Arthur, 2003). If true, this state of affairs would produce the same dimension–different exercise (SDDE) correlations (sometimes referred to as ‘‘convergent validities’’; e.g., Woehr & Arthur, 2003) that are large relative to the different dimension–different exercise (DDDE) correlations and different dimension–same exercise (DDSE) correlations (sometimes referred to as ‘‘discriminant validities,’’ e.g., Woehr & Arthur, 2003) that are relatively low, as correlations among the DDSE correlations would reflect upon the distinctness or discriminability of the dimensions being measured within each exercise. Furthermore, traditional AC theory would anticipate that factor analyses of correlation matrices would result in factors that represent the dimensions being measured and not the exercises (methods) used to measure them (Lance, 2008).

Nonetheless, as several narrative reviews (e.g., Howard, 1997; Lievens & Klimoski, 2001; Sackett & Tuzinski, 2001) and large-scale empirical summaries of existing findings on AC construct validity have shown (e.g., Bowler & Woehr, 2006; Lance, Lambert, et al., 2004; Woehr & Arthur, 2003), these expectations have not been supported. Instead, DDSE correlations are almost always larger than SDDE correlations (and usually substantially so), and (both exploratory and confirmatory) factor analyses almost always support robust exercise factors and not dimension factors. That is, the accumulated evidence to date indicates little or no evidence for convergent or discriminant validity of AC dimensions and strong and robust method (exercise) effects.

Why the Construct Validity is Low?

Lance (2008) argued that the AC construct validity problem has arisen from the misapplication of multitrait–multimethod (MTMM) design to test what in hindsight were unjustified hypotheses concerning AC candidate behavior that is inherently cross-situation specific and that tends to be accurately evaluated by assessors. Rupp, Thornton and Gibbons (2008), also acknowledged that multitrait–multimethod (MTMM) approach for establishing construct validity of assessment center (AC) ratings is inappropriate. However, they mentioned that this assertion is only supportable under a narrow, incomplete, and outdated definition of construct validity and an exclusive reliance on MTMM-based analyses of within-exercise dimension ratings.

Howard (2008) noticed that a lamentable reconceptualization of the assessment center model took place with the application of the multitrait–multimethod (MTMM) approach, so that misuse of this model has been a serious distraction to understanding the architecture of assessment centers. He argued that MTMM misrepresents assessment center design by assuming that all exercises are equally capable of measuring each dimension marked with an X on a dimension_ exercise coverage grid. An exercise that might shed a little light on a dimension, or elicit only one of several key behaviors that are included in the definition of the dimension, is suddenly given equal status with an exercise that was designed specifically to measure that dimension. It is no wonder that dimensions fail to hang together in statistical analyses that rest on this dubious assumption.

Lance (2008) concluded that assessment centers do not measure dimensions at all but only situationally specific exercise performance. Howard (2008) also noted that another problem with MTMM is the misconception that its terminology creates. Assessment center simulations do not—or should not—measure traits. They measure observable behaviors that are logically organized into categories related to job success. as Howard (1997) mentioned, the population of dimensions or competencies is a muddled collection of learned skills, readily demonstrable behaviors, basic abilities, attitudes, motives, knowledge, and other attributes, including traits, that are often ambiguously defined and difficult to rate. The assessment center guidelines clearly state that competencies can only be used as assessment center dimensions if they can be ‘‘defined precisely and expressed in terms of behaviors observable on the job or in a job family and in simulation exercises’’ (International Task Force on Assessment Center Guidelines, 2000).

Brannick (2008) noted that a main reason for the troubling multitrait– multimethod (MTMM) results is a mismatch between the inferences to be made based on the scores and the construction of the exercises. Construct validity evidence is poor because the exercises are based on tasks sampled for content rather than chosen or designed for illuminating individual differences on the constructs. In other words, exercises are typically based on job content and work samples, and the scoring system is typically based on knowledge, skill, ability, and other characteristics (KSAOs) or traits.

Arthur, Day & Woehr`s (2008) position is that the issue is not one of a failure in ‘‘AC theory’’ but rather a failure to engage in appropriate tests of the said theory. In fairly broad terms, construct validity pertains to an assessment of whether a test is measuring what it purports to measure, how well it does so, and the appropriateness of inferences that are drawn from the test’s scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society for Industrial and Organizational, Inc., 2003). They also mentioned that the problem is that this focus is largely an artifact of the requirements of multitrait–multimethod (MTMM)-based approaches to construct-related validity rather than the way in which dimension ratings are typically operationalized.

Given the importance of candidates’ cross-situationally inconsistent performance across exercises, Lievens (2008) argued that research should also pay attention to the assessees. First, we need to better understand which individual differences variables affect candidate performance across exercises. For example, people who are high on social effectiveness constructs are typically able to ‘‘read’’ situations better than others and flexibly adapt their interpersonal behavior in line with the cues gathered. Rupp, Thornton and Gibbons (2008) stated that Consistency of behavior and differentiation of performance across dimensions can and should be viewed at the individual level of analysis. Connelly, Ones, Ramesh and Goff (2008) also suggested that (a) there are some determinants of assessment performance that are

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

common across exercises and (b) these determinants are stable characteristics of assessees. Stable dimensions have important effects on behavior in assessment center exercises. However, psychometric factors attenuate dimension effects, making assessment center behavior appear more situationally specific than it truly is.

Lievens (2008) also mentioned that we know little about how variations in exercise instructions and exercise design might influence performance; so we need to find out which exercise characteristics are ‘‘incidentals’’ (i.e., surface exercise characteristics that do not determine performance) and which ones are ‘‘radicals’’ (i.e., structural exercise characteristics that determine performance). In a related domain (situational judgment tests), research has shown that even minor variations in the situations presented to candidates might affect performance (Lievens & Sackett, 2007). Lievens (2008) also noted, the interaction between individual differences variables and exercise characteristics are important. In this context, an interaction theory like trait activation theory, might help to better understand factors that affect candidate performance variations across exercises. For example, trait activation theory might help to identify which exercise factors trigger and release trait-relevant candidate behavior versus which ones impede trait-relevant candidate behavior (Tett & Burnett, 2003).

Schuler (2008) suggestion is that assessment centers often perform poorly because too simplistic methods are employed. Their attractiveness for managers and practitioners in personnel departments is connected with concentrating on ‘‘exercises’’ such as a group discussion, roleplay, and presentation, which allow for behavioral observations and a lively personal impression formation but are essentially non-psychometric tools. Plausible

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

reasons about why assessment centers, notwithstanding their considerable expenditure, have low validity, are that their exercises are not really based on task requirements, not really job or organization specific, and not developed as structured, reliable tasks in a process equivalent to usual test development. The same is true for assessment center observers who are psychological laypersons in most cases.

As stated earlier, evidence for the construct-related validity of assessment center dimensions has been less promising. Specifically, assessment centers are designed to evaluate individuals on specific dimensions of job performance across situations or exercises. Research, however, has indicated that exercises rather than dimension factors emerge in the evaluation of assessees. So, the purpose of this article is to reconsider the proven status of the assessment center construct validity problem and to propose some solutions.

Method

The design of the present study is a correlation design based on a multidimensional-multivariate matrix as well as confirmatory factor analysis. In the present study, the statistical population was all of the employees of the National Iranian Oil Company who were evaluated from the beginning of 1391s.c* (2012) to 1398 s.c (2019) in the Assessment and Development Center. In order to select a sample from the population of the employees evaluated, the method of simple random sampling was used. The

* Solar Calender

number of employees evaluated in the assessment center between 1391 s.c (2012) to 1397 s.c (2019) was 9,000. According to the Cochran sampling formula for a specific population (with a permissible error value of 5% and z equal to 1.96), the sample size was calculated to be 384 people which formed the sample of this study.

Assessors provided ratings for each participant in their group on each observable dimension in each exercise. We analyzed the ratings on two exercises and four dimensions. The selected exercises were the interview and the managerial game. The selected dimensions were leadership, planning, innovation, and decision-making. We inspected the correlations between the dimensions and exercises and conducted a Confirmatory Factor Analysis (FCA) test.

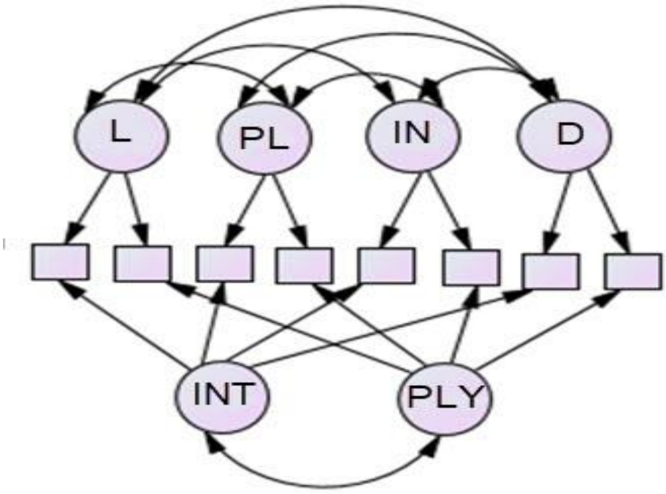
Over the past decade, Campbell and Fiske's (1959) MTMM paradigm has become one of the most frequently applied methods for investigating construct validity. The MTMM paradigm is based on the correlations betwee observable variables. The correlational analyses involved comparing the mean of the within-dimension and across-exercise correlations with the mean of the within-exercise and cross-dimension correlations. Higher values for the former relative to the latter would be indicative of the convergent/discriminant validity.

Although correlational analysis provides some evidence of both dimension and exercise effects, it does not allow for an overall test of these effects. Consequently, we used CFA to evaluate a model representing both exercise and dimension factors (i.e., we used a traditional CFA approach to MTMM data). The model evaluated six latent variables (shown in Figure 1). Four of the latent variables represented four factors (analogous to trait factors in MTMM analysis) and two of the latent variables

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

represented exercise factors (analogous to method factors in MTMM analysis). The overall measures of fit for this model indicate how well a model specifying the four dimension and the two exercise factors corresponds to the data. In addition, a comparison of the magnitude of the individual parameter estimates of the dimension factors on the ratings versus parameter estimates of the exercise factors on the ratings provides an indication of the relative magnitude of the dimensions and exercises. Specifically, large dimension factor loadings indicate the existence of convergent validity, large exercise factor loadings indicate the existence of exercise effects, and large dimension correlations indicate a lack of discriminant validity (Marsh & Grayson, 1995).



Note: L = Leadership; PL = Planning; IN = Innovation; D = Decision; INT = Interview; PLY = Playing.

Figure 1. CFA model

Results

Fit values for the default model, saturated model and Independence model. A saturated model is a model in which all the possible parameters are added to it. That is, all the relationships between the variables are plotted. Such a model has a perfect fit and its reproduced matrix is equivalent to the observed matrix, so the remaining matrix will be zero. The purpose of this model is to estimate the variance - covariance of variables in the population. Sometimes this model is used as a basis for determining the success of the developed model (indicators closer to it but with fewer parameters).

The independence model or zero model is a base model for the comparison in which no non-free parameters (such as covariance between variables) are defined. In other words, it lacks any one-way or two-way relationship between variables. Comparative fit indices how far the model has been able to distance itself from the independence model. The greater the distance, the better the fit of the model.

Based on the content of Table 1, it is concluded that the developed model has greatly reduced the chi square of an independence model (more than 3900). It can be said that, the rejection of the independence model, methodologically justifies the development of a research model.

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

Table 1

Model Fit Summary (CMIN)

Model	NPAR*	CMIN**	DF***	P****	CMIN/DF*****
Default model	37	9.346	7	.229	1.335
Saturated model	44	.000	0		
Independence model	16	3981.449	28	.000	142.195

*number of distinct parameters (q) being estimated **minimum discrepancy

degrees of freedom **** P-Value ** minimum discrepancy per degree of freedom

Table 2
Baseline Comparisons

Model	NFI* Delta1	RFI** rho1	IFI*** Delta2	TLI**** rho2	CFI*****
Default model	.998	.991	.999	.998	.999
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

*normed fit index **relative fit index ***incremental fit index

****Tucker–Lewis index *****comparative fit index

The Poor fit index and its confidence level indicates that the fit is desirable and differs greatly from the independence model.

[DOI: 10.22034/ijpb.2021.261549.1211]

Table 3
RMSEA Index

Model	RMSEA*	LO 90**	HI 90**	PCLOSE***
Default model	.018	.000	.045	.977
Independence model	.373	.363	.383	.000

* Root Mean Square Error of Approximation

**the columns labeled LO90 and HI90 contain the lower limit and the upper limit of a 90% confidence interval for the population value of RMSEA

*** p of Close Fit (This measure is a one-sided test of the null hypothesis that the RMSEA equals .05, which is called a close-fitting model).

All adaptive indices show values higher than .9, which means the model is able to distance itself from the independence model and approach the saturation model.

The inter-correlations among the ratings for each dimension derived from each exercise are presented in Table 4. These correlations provide evidence suggesting that this assessment center demonstrates method (exercise) factors as opposed to trait (dimension) factors. As indicated in Table 4, the overall mean correlation among the ratings of the same dimension across exercises was .18 compared with a mean overall correlation among ratings of different dimensions within the same exercise of .60.

Table 4
Dimension and Exercise Intercorrelations

		Playing				Interview			
		L	**PL**	**IN**	**D**	**L**	**PL**	**IN**	**D**
playing	**L**	-							
	PL	.75	-						
	IN	.68	.66	-					
	D	.71	.72	.61	-				
interview	**L**	.19	.17	.12	.16	-			
	PL	.23	.18	.12	.21	.62	-		
	IN	.25	.21	.16	.24	.47	.47	-	
	D	.20	.19	.14	.19	.61	.59	.43	**-**

Mean Within-Dimension Across-Exercises *r*:	Mean Within-Exercise Across-Dimensions *r*:
Leadership = .19	playing = .65
planning = .18	interview = .55
innovation = .16	
Decision = .19	
Overall Mean = .18	Overall Mean = .60

IN = Innovation; L = Leadership; PL = planning; D = Decision

Confirmatory Factor Analysis

We used a CFA application of Amos 24 to evaluate the fit of the model presented in Figure 1. Covariances among the 8 ratings (1 rating for each of the four dimensions based on each of the two exercises) served as the input to the program. The overall fit indices indicate that the model provides an excellent representation of the data (χ^2 [7] = 9.54 ns, GFI = 1.0, AGFI 5 .98, RMSR = .02, NFI 5 .99, CFI 5 1.0). Examination of the individual parameter estimates, presented in Table 5, indicates that while both dimension and exercise factors contribute to the ratings, the estimates for the exercise parameters are substantially larger in all of the cases than estimates for dimension factors (mean parameter estimates of .39 and 0.68 for the dimension and exercise parameters, respectively). In addition, inter-correlations among the latent dimension factors ranged from .98 to 1.0, providing evidence of no discriminant validity across dimensions. Thus, these results support and are consistent with those obtained from the correlational analysis. That is, the exercise factors are the primary determinants of assessment center ratings.

Table 5
Standardized Parameter Estimates from the MTMM CFA Model

	Dimensions				**Exercises**	
	Leadership	**planning**	**Innovation**	**Decision**	**Playing**	**Interview**
PLY, L	.58				.68	
PLY, PL		.42			.77	
PLY, IN			.39		.65	
PLY, D				.68	.61	
INT, L	.19					.79
INT, PL		.27				.73
INT, IN			.34			.51
INT, D				.23		.71
Mean Dimension Loading = .39				Mean Exercise Loading = .68		

Note: L = Leadership; PL = Planning; IN = Innovation; D = Decision; INT = Interview; PLY = Playing.

Discussion

The purpose of this article is to reconsider the proven status of assessment center construct validity problem and to propose some solutions. As the results show, the inter-correlations among the ratings for each dimension derived from each exercise provide evidence that this assessment center demonstrates method (exercise) factors as opposed to trait (dimension) factors. As indicated in Table 4, the overall mean correlation among ratings of the same dimension across exercises was .18 compared with a mean overall correlation among ratings of different dimensions within the same exercise of .60. These results are in line with the Lance (2008) assertion about the assessment center construct validity.

In this context, Howard (2008) states that the problem lies with misguided approaches to the assessment center research and practice. The problem areas include (a) questionable theory and models underlying the experimental tests, (b) misinterpretation and/or misuse of dimensions, (c) misunderstanding the practical uses of assessment centers, and (d) a simplistic and outdated view of assessment center design.

Researchers argue that there are certain design characteristics of ACs as they are typically implemented that, if reengineered, should lead to increased construct validity of AC. This argument is based on the ideas that in typical ACs, assessor cognitive demands are excessive, number of dimensions are extensive, target candidate behaviors and dimensions are not defined sufficiently concretely, poor and non-psychometric tools and methods are performed, assessors are not sufficiently skilled, and/or certain rating strategies (e.g., the ‘‘within-exercise’’ method) engender systematic rating biases (Lievens & Klimoski, 2001; Lance, 2008; Arthur, Day & Woehr, 2008; Howard, 2008; Schuler, 2008).

Following this line of reasoning, a number of design fixes have been studied in attempts to increase construct validity, including targeting the key behaviors by assessment designers that define each dimension to be rated and create simulations that will elicit these behaviors (Howard, 2008), reducing the number of dimensions to be rated (e.g., Arthur, Day & Woehr, 2008; Howard, 2008), better defined dimensions (Howard, 2008), providing behavioral checklists that specifically anchor what the dimension includes to aid in observing and recording candidate behavior (e.g., Hennessy, Mabey & Warr, 1998; Howard, 2008), making dimensions transparent to candidates (Kleinmann & Koller, 1997; Kolk, Born & van der Flier, 2003), applying the standard test development and psychometric approaches and practices (Arthur, Day & Woehr, 2008; Schuler, 2008), extending methodological diversity (Schuler, 2008), building design characteristics into exercises that might elicit specific trait related behavior (Lievens, 2008); using alternative rating methods such as the within-dimension or across-exercise (vs. within exercise) method (Arthur, Woehr & Maldegen, 2000; Robie et al., 2000), aligning the exercises or stimulus content with the scoring system (Brannick, 2008), using expert (vs. nonprofessional) assessors (e.g., Lievens, 2002; Schuler, 2008), and providing assessors with various types of training (e.g. Howard, 2008; Lance, 2008).

Howard (2008) suggests that assessors should be aware of situational differences and take them into account. For example, when assessors observe notably different behaviors in two different exercises—as when they rate building relationships ‘‘4’’ in a customer exercise and ‘‘2’’ in a peer exercise— they should rate the final dimension as 4 or 2 rather than a compromise 3.

Lance (2008) argues that cross-situational variance in dimensions is not necessarily error; it can also be argued that inter-correlated dimensions are not necessarily error, particularly those that are in similar domains, such as interpersonal skills (Howard, 2008).

Also, we need to pay closer attention to the espoused versus actual construct issue. We need to hold AC researchers to the same psychometric test development standards to which we hold all other test developers. In addition, they suggested that We need to move beyond a reliance on only internal structure and instead include tests of external construct-related validity that examine the nomological network of post consensus dimension ratings (Arthur, Day & Woehr, 2008).

Rupp, Thornton and Gibbons (2008) argued that within-exercise dimension ratings should not be used as the unit of analysis when exploring the construct validity of the AC method. The Management Progress Study, and many of the applied ACs that followed it, generated overall dimension ratings only after hearing reports of a candidate’s performance in all exercises (Howard, 1997).

Assessment centers should routinely compute estimates of the reliability of candidate performance within exercises. One-way to do so is to deliberately introduce multiple dimension-relevant items or problems within the exercises and to score such items. For example, if we want to assess assertiveness, we should design at least three such problems (not just one) as part of a single exercise. For another example, instead of having one 30-minute performance discussion, we might have five different 6-minute assessments where the candidate is given instructions to react to specific problems for each subordinate (Brannick, 2008).
Lievens (2008) suggested that as we know little about how individual differences and variations in exercise instructions and design might influence performance, research should pay more attention to and scrutinize exercise characteristics, individual differences variables and factors that affect candidate performance variations across exercises. Brannick (2008) too mentioned that we should pay more attention to the psychometrics of our simulations, particularly the reliability of the exercise scores related to candidate actions.

We suggest that future research uses the overall dimension ratings of the selection ACs, which are assumed to evoke maximum performance and the dimension ratings from other maximum performance situations. This might enhance the chances that the dimension factors can be found for the AC overall dimension ratings and the external dimension ratings. As another example, a parallel selection AC or a selection interview that targets the same dimensions might be suitable for the maximum performance situations. Still another example is the comparison of the overall dimension ratings from the developmental ACs with the external measures of the same dimensions in order to evaluate the convergence of these dimension ratings under the conditions that ratings of both sources might more strongly reflect typical performance.

Acknowledgments

The authors would thank the managers, assessors and staff of the National Iranian Oil Company Assessment Center for their help in conducting the research and gathering information.

And in memory of a worthy graduate of industrial and organizational psychology who soon passed away (**Mohammad Montakhab Yeganeh**).

References

Arthur, Jr., W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, *26*(4), 813-835.

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*(1), 105-111.

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*(3), 478.

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*(5), 1114.

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

Buckett, A., Becker, J. R., Melchers, K. G., & Roodt, G. (2020). How different indicator-dimension ratios in assessment center ratings affect evidence for dimension factors. *Frontiers in Psychology*, 11, 459.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology*, *1*(1), 121-124.

Dunnette, M. D. (1971). The assessment of managerial talent. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 2). Palo Alto, CA: Science and Behavior Books.

Hardison, C. M., & Sackett, P. R. (2004). Assessment center criterion-related validity: A meta-analytic update. *Unpublished Manuscript*.

Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment centre observation procedures: An experimental comparison of traditional, checklist and coding methods. *International Journal of Selection and Assessment*, *6*(4), 222-231.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12*(5), 13–52.

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

International Task Force on Assessment Center Guidelines. (2000). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment*, *17*(3), 243-253.

Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of assessment centers: A conceptual review. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*, 349-372. doi:10.1146/annurev-orgpsych-012218-014955.

Kleinmann, M., & Köller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality*, *12*(5), 65.

Kolk, N. J., Born, M. P., & Der Flier, H. V. (2003). The transparent assessment centre: The effects of revealing dimensions to candidates. *Applied Psychology, 52*(4), 648-668.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. Industrial Organizational Psychology: *Perspectives on Science and Practice, 1*, 84–97.

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

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*(2), 377.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, *87*(4), 675.

Lievens, F. (2008). What does exercise-based assessment really mean?. *Industrial and Organizational Psychology*, *1*(1), 112-115.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now?.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, *92*(4), 1043.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data.

Moses, J. (2008). Assessment centers work, but for different reasons. *Industrial and Organizational Psychology*, *1*(1), 134-136.

Moses, J. L. & Byham,W. C. (Eds.). (1977). Applying the assessment center method. New York: Pergamon.

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*(4), 355-370.

Rupp, D. E., Thornton, G. C., & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology*, *1*(1), 116-120.

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

Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology*, 69, 187–190.

Sackett, P. R., & Tuzinski, K. (2001). The role of dimensions in assessment center judgment. In M. London (Ed.), How people evaluate others in organizations. Mahwah, NJ: Erlbaum.

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102,* 1435-1447.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, *77*(1), 32.

Schuler, H. (2008). Improving assessment centers by the trimodal concept of personnel assessment. *Industrial and Organizational Psychology*, *1*(1), 128-130.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500.

Thornton, G. C., III, & Byham,W. C. (1982). Assessment centers and managerial performance. New York: Academic.

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]

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

Wirz, A., Melchers, K. G., Kleinmann, M., Lievens, F., Annen, H., Blum, U., & Ingold, P. V. (2020). Do overall dimension ratings from assessment centers show external construct-related validity?. *European Journal of Work and Organizational Psychology*, *29*(3), 405-420.

Woehr, D. J., & Arthur, Jr., 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*(2), 231-258.

[DOR: 20.1001.1.20081251.2020.14.2.8.2]

[DOI: 10.22034/ijpb.2021.261549.1211]