Industrial and Organizational Psychology, **1** (2008), 105–111.

Mend It, Don't End It: An Alternate View of Assessment Center Construct-Related Validity Evidence

WINFRED ARTHUR, JR.
Texas A&M University

ERIC ANTHONY DAY
The University of Oklahoma

DAVID J. WOEHR
The University of Tennessee

Why the Current State Is What It Is

The unitarian conceptualization of validity serves as the conceptual and logical basis for the so-called assessment center (AC) "construct-related validity paradox." Within the unitarian framework, 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). And because ACs do not appear to do so, we have the resultant AC construct-related validity paradox. So, accepting the premise that the unitarian view is conceptually and logically sound, what is the explanation for the paradox? Why do AC dimension ratings appear not to "work" in terms of construct-related validity evidence? At a broad conceptual level, we present a view that is contrarian to Lance's (2008) view of "why ACs don't work the way they're supposed to" and subsequently what to do about them.

Our contrarian view is based on two key points, namely that the vast majority of the empirical AC research to date—particularly that which serves as the basis for calls for the "redesign of ACs toward task- or role-based ACs and away from traditional dimension-based ACs" (Lance, 2008, p. 84)—is based on (a) espoused as opposed to actual constructs and (b) flawed analysis resulting from an overemphasis on postexercise dimension ratings as measures of AC dimensions. In our view, ACs in practice appear to be effectively designed to representatively sample from the job content domain and also predict criteria of interest, but they are woefully deficient in their construct explication and development. Consequently, we do not concur with Lance's interpretation of the extant literature and his conclusions concerning what to do about it. Our position is that the issue is not one of a failure in "AC theory" but rather a failure to engage in appropriate tests of said theory. Until such tests have been undertaken, we think it is premature to abandon

Correspondence concerning this article should be addressed to Winfred Arthur, Jr. E-mail: wea@psyc.tamu.edu

Address: Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235

Winfred Arthur, Jr., Department of Psychology, Texas A&M University; Eric Anthony Day, Department of Psychology, The University of Oklahoma; David J. Woehr, Department of Management, The University of Tennessee.

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

a dimension-based approach to ACs for a task- or role-based focus. We expand on the basis for our position below.

Espoused Constructs Versus Actual Constructs—The "Elephant in the Room"

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). But as predictors, what is it that ACs measure? "A predictor is a specific behavioral domain, information about which is sampled via a specific method. Thus depending on one's focus, predictors can be represented in terms of *what* they measure and *how* they measure what they are designed to measure" (Arthur & Villado, in press).

Traditional AC theory specifies the behavioral domain of ACs in terms of psychological constructs (i.e., "dimensions").[1] That is, AC theory is based on an interpretation of dimensions as constructs that are presumed to underlie performance on the job, and the operationalization of these dimensions is via performance on a number of AC exercises. Thus, the overt behaviors observed by assessors are best conceptualized as a sample set of indicators of the constructs of interest. Consistent with classical psychometric theory, which has been developed to deal with the measurement of latent constructs, this conceptualization is an important one because it has implications for the psychometric rigor applied to the development of tests and measures such as ACs. Consequently, a crucial aspect of test development is establishing that a test or measure actually assesses what it claims to measure. The fundamental issue here is one of construct validity and an emphasis on the fact that merely labeling data as reflecting a particular construct (espoused construct) does not mean that is the construct that is being assessed (actual construct). Yet, for some unexplainable reason, this practice appears to be the norm in AC research and practice where statements about what exercises measure (e.g., stress tolerance, social competence, factual argumentation, activity, imaginativeness) are by self-proclamation with rarely any systematic psychometric test development evidence presented to support these assertions (Woehr & Arthur, 2003).

Hence, the AC literature appears to be a domain in which assertions about the dimensions being measured are rarely, if ever, subjected to the psychometric standards that characterize test development in other domains (e.g., general mental ability testing, personality). Rather claims about what constructs/dimensions are being measured are taken at face value. And given the particularly esoteric nature of some of the construct labels (e.g., sensitivity, self-direction, inspiring trust, seasoned judgment, personal breadth; see table 2 of Arthur, Day, McNelly, and Edens, 2003, for additional examples), it is not surprising that we see little evidence supporting these latent factors. We have always been baffled as to why ACs and other method-based predictors (e.g., situational judgment tests and interviews) are not held to widely recognized and common psychometric test development standards. Proper test development involves an iterative sequence of content formulation, evaluation, and refinement—a process that typically entails multistage data collection and refinement efforts before a test is put into operational or research use. Validity must be built at the onset of test construction and before operational use (Anastasi & Urbina, 1997). Accordingly, the formulation and explication of construct definitions should not be approached casually, and support for the construct representation of test content should be perspicuously demonstrated. Instead, we are

1. Whereas our preference is to use the term *construct* to refer to what ACs measure, where warranted, we also use the term *dimension* to be consistent with common usage in AC literature.

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

willing to simply accept the authors' claim that their ACs measure the espoused constructs—something that we rightly are unwilling to and do not accept in the context of other tests and measures.

We posit that we cannot have a meaningful discussion of the construct-related validity of AC ratings in the absence of any evidence or demonstration that they do indeed measure the intended constructs. Phrased another way, the types of tests and analyses presented in the AC construct-related validity debate are all predicated on the assumption that the exercises are actually measuring the espoused constructs and that these constructs are distinct from each other. Yet, these constructs are derived by simply labeling dimensions in an espoused manner without clear evidence to support their construct representation.

It is our view that the reliance on espoused constructs is one important reason for the construct-related validity issues that are observed with ACs, especially at the level of primary studies. For instance, one result of the espoused self-proclamations is the lengthy, extensive list of dimensions purported to be measured by ACs, such that Arthur et al. (2003) extracted 168 dimension labels from 34 ACs and Woehr and Arthur (2003) extracted 129 from 48 ACs. Although human behavior is certainly complex, it seems unlikely that 129–168 different dimensions are required to explain managerial performance. Thus, the reliance on espoused constructs has perpetuated a lack of dimension distinctiveness within primary studies. Do we really actually expect to find evidence of construct discrimination in ACs simultaneously tapping dimensions like analysis, judgment, and problem solving?

In contrast, three recent meta-analyses have demonstrated fairly favorable construct validity evidence for AC dimensions after collapsing the myriad of dimensions found in the literature into a smaller, conceptually distinct set of dimensions. Specifically, using Arthur et al.'s (2003) seven-dimension taxonomy, these meta-analyses provide evidence for the criterion-related validity of AC dimensions (Arthur et al., 2003), the impact of dimension factors (Bowler & Woehr, 2006), and the differentiation from and incremental criterion-related validity of AC dimensions over cognitive ability and personality (Meriac, Hoffman, Fleisher, & Woehr, 2007). Thus, it would seem that these recent studies provide strong evidence for the construct validity of the core set of AC *dimensions* posited by Arthur et al. and that Lance's call for the abandonment of dimensions is at best premature.

In summary, in spite of dimensions (as constructs) being the central focus of traditional AC development, there are low expectations in the published literature on how dimensions are to be derived and defined. The typical study includes a few perfunctory statements about the use of job analysis to identify appropriate exercises and the dimensions to be extracted as well as lists of the exercises and dimensions. In *some* instances, the list of dimensions includes a one-sentence definition of each dimension. In the absence of any description of how the dimensions were operationalized or what procedures were used to ensure that each dimension was adequately represented without construct contamination or deficiency, the focus of measurement in the typical AC is dubious at best.

As psychologists, our focus should be on constructs (Landy, 1986), and thus, we should take great care in how our constructs are conceptualized and operationalized. In fact, it is astonishing that despite the robust literatures and scientific disciplines devoted to areas like judgment and decision making, influence, communication, and leadership, we have not seen the developers of ACs consulting these literatures when defining their similarly labeled dimensions. Thus, we contend that the current state of the AC literature could be characterized as overly empirical, given the profusion of dimension labels and paucity of dimension explication. We believe that the AC literature is in great need of more rigorous procedures for both the conceptualization and the operationalization of dimensions. We should not be simply labeling AC constructs/dimensions in an espoused manner. Instead,

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

we should be applying the standard test development and psychometric approaches and practices to demonstrating that they are actually measuring the intended constructs *before* we proceed with their use and also the comparative evaluation of dimensions and exercises. In other words, researchers and practitioners should undertake formative evaluations of ACs before proceeding with summative evaluations like determining the relative contribution of dimension versus exercise variance in AC ratings.

Overemphasis on Postexercise Dimension Ratings

An important issue that is often overlooked in the current debate is the fact that the overwhelming majority of literature examining AC construct-related validity has exclusively focused on postexercise dimension ratings. Indeed, this is vividly highlighted in Lance's (2008) abstract (and elsewhere throughout the article) where he states that "after 25 years of research it is now clear that *AC ratings that are completed at the end of each exercise* (commonly known as post exercise dimension ratings) substantially reflect the effects of the exercises in which they were completed and not the dimensions they were designed to reflect" (p. 85, italics added). 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 (see Lance, Woehr, & Meade, 2007, for additional problems with the MTMM approach).

In the context of ACs, the MTMM approach is operationalized such that dimensions are viewed as traits and exercises as methods. Thus, in order to examine the magnitude of dimension and exercise effects, ratings are required for each dimension within each exercise (i.e., postexercise dimension ratings). Here, it is important to note that two primary evaluation approaches have been identified across ACs (Robie, Adams, Osburn, Morris, & Etchegaray, 2000; Sackett & Dreher, 1982; Woehr & Arthur, 2003). In the *within-exercise* approach, assessees are rated on each dimension after completion of each exercise. In the *across-exercise* (i.e., within-dimension) approach, evaluation occurs after all the exercises have been completed, and dimension ratings are based on performance from all the exercises.

However, common to both these rating approaches is the use of post consensus dimension ratings to assess dimensions. Post consensus dimension ratings represent the combination (either clinical or mechanical) of postexercise dimension ratings into a summary representing dimension-level performance information across multiple exercises and raters. From a traditional psychometric perspective, post exercise dimension ratings may be viewed as item-level information, whereas post consensus dimension ratings represent scale-level information. The increase in reliability associated with the aggregation of item-level information sets a higher upper bound for the validity of the measurement tool. Unfortunately, however, the use of post consensus dimension ratings does not allow the evaluation of internal construct validity in the form of typical convergent and discriminant validity coefficients. Thus, although it represents the dimension score, it is not the data used in MTMM analyses—instead, the level of analysis is at the item level, that is postexercise dimension ratings.

Regarding discussions of AC construct-related validity, this is an important and critical distinction. That is, postexercise dimension ratings are not the final measure of dimensions but rather an item in a composite measure. Thus, Lance's (2008) position is at one level a straw man argument because anyone would be hard pressed to disagree with the observation that AC ratings that are completed at *the end of each exercise* reflect exercise effects more so than dimension effects. Items on any composite measure typically reflect small proportions of true score variance, yet, the final composite based on these items may still provide a reliable measure of the construct of interest. Consequently, the pertinent issue that Lance did not address is whether measures

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

based on across-exercise composite ratings also reflect exercise effects more so than they do dimension effects. Results, such as those presented by Arthur, Woehr, and Maldegen (2000) and Woehr and Arthur (2003), would lead us to conclude that they do not.

What Are We to Do About the Current State of the Literature?

We obviously hold quite different views from Lance about the reasons for the current state of the AC construct-related validity literature, and subsequently, what to do about it. We believe that giving up the traditional focus of ACs on human requirements and turning to task- or role-based ACs are (a) premature, given how poorly dimensions have been conceptualized, operationalized, and scored, and (b) scientifically untenable, given that identifying the human requirements to effective task and role performance is fundamental to the field of industrial–organizational (I–O) psychology. Indeed, we would argue that psychology in general focuses on constructs as the subject of study (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994); there is no reason why the same should not be true for I–O as well.

Our position is that AC theory has not been adequately and satisfactorily tested, and as a result, it is premature to abandon it. Consequently, it should not come as a surprise that we find the solution recommended by Lance to be extremely problematic on several grounds. Not the least of these is the scientific versus technical focus of I–O psychology as a discipline. Distinguishing between predictor constructs and predictor methods (Arthur & Villado, in press) has important implications for the debate on the construct-related validity associated with AC dimension scores. Measurement tools are by definition a means of operationalizing constructs. Accordingly, the position that ACs measure "exercises" and not constructs is not *scientifically* very tenable in that it is analogous to stating that scores on some paper-and-pencil tests represent performance in paper-and-pencil situations.

We find it to be particularly ascientific to resign ourselves to a position that (to paraphrase) states that "ACs measure something, we just do not know what it is" or that the scores represent exercises when exercises are not constructs. We submit that the onus is on us as members of a scientific discipline to investigate and determine what ACs measure and the boundary conditions within which they may or may not do so effectively—a point that was also made over 20 years ago by Zedeck (1986). This endeavor can be best facilitated by taking a construct-oriented approach from the outset of test construction. A focus on ACs as work samples does not advance the literature in any scientifically or theoretically meaningful manner—a method focus or even a task or role focus only moves us toward becoming more of a technician-oriented discipline.[2]

In addition to simply being better science, a construct-oriented approach is also consonant with and has implications for recent and recurring discussions on defining I–O psychology as a field and a discipline (e.g., Gasser, Butler, Waddilove, & Tan, 2004; Highhouse & Zickar, 1997; Ryan, 2003). We submit that a construct-oriented focus that recognizes and rigorously maintains the predictor construct and predictor method distinction serves as a point of departure that distinguishes I–O psychologists as scientists from other practitioner-oriented human resource management fields and disciplines and is a step in addressing concerns about how well I–O psychology fits within the broader field of psychology. It is difficult to overstate the centrality of constructs to psychology as a scientific discipline.

There is also applied utility to a focus on dimensions. For instance, as noted by

2. The AC literature—both academic and practitioner—has focused primarily on the method and technique of ACs and less so on the constructs or what they measure as an area of study. Situational judgment tests and interviews are two other predictor methods that have had and continue to have a technique and method focus, and interestingly, one thing they have in common with ACs is a lack of clarity and confusion about what they measure and exactly how well they do so.

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

Howard (1997), "There are a number of practical reasons why assessment center users prefer dimensions based on human attributes rather than tasks. Lists of tasks can be long and generalize to fewer situations. Tasks are an unnatural way to describe people and are less meaningful than attributes for developmental feedback. Important to psychologists, task descriptions have little explanatory power. Landy and his colleagues made a similar argument about job analysis; psychologists should be studying human qualities, not tasks" (p. 28).

What Is the Next Step? Where Do We Go From Here?

We would like to conclude by putting forward a number of recommendations on how to move the research and discussion forward on the construct-related validity issue. First, as previously noted, as researchers, practitioners, editors, reviewers, and educators, we need to pay closer attention to the espoused versus actual construct issue. We need to hold AC researchers and others (who purport to speak to or contribute to the construct-related validity discussion) to the same psychometric test development standards to which we hold all other test developers. This is a prerequisite to a meaningful discussion of the construct-related validity issue.

Second, we need to design and implement ACs that properly represent the constructs they are intended to measure in a manner that is as free as possible from construct-irrelevant variance (Messick, 1995). There is now a rich and reasonably large body of research that indicates what the desirable features of ACs are (Lievens, 1998; Woehr & Arthur, 2003). Thus, recognizing that as methods, they are only as good as their design and implementation, in *addition to* the espoused versus actual construct issue above, we should again hold AC researchers and others to the incorporation of these features in their ACs.

Third, we need to broaden our evidential basis beyond MTMM. 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 (e.g., Meriac et al., 2007). This approach is consonant with Woehr and Arthur's (2003) finding that the results of studies examining the relationship of AC dimension ratings with convergent and discriminant constructs measured by other methods such as paper-and-pencil measures tended to provide evidence for both convergent and discriminant validity. In addition, research that incorporates the *comparative* criterion-related validities of dimension versus exercise scores in conjunction with the nomological net approach would also be very informative.

Finally, upon amassing a body of research that addresses the issues noted above, if we still have a construct-related validity paradox, then we think we can start to have a more grounded discussion of the "theory of ACs" and why a framework that seems so conceptually sound is generating empirical data that are so at odds with the prevailing unitarian view of validity. Until the above issues are commonly addressed and incorporated into the extant empirical literature, we think it is premature to abandon a construct-oriented (i.e., dimension-based) conceptualization of ACs.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NY: Prentice-Hall.

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

Arthur, W., Jr., & Villado, A. J. (in press). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*.

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

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the

https://doi.org/10.1111/j.1754-9434.2007.00019.x Published online by Cambridge University Press

inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494.
Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*, 1114–1124.
Gasser, M., Butler, A., Waddilove, L., & Tan, R. (2004). Defining the profession of industrial-organizational psychology. *Industrial-Organizational Psychologist, 42*(2), 15–20.
Highhouse, S., & Zickar, M. J. (1997). Where has all the psychology gone? *Industrial-Organizational Psychologist, 35*(2), 82–88.
Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality*, *12*, 13–52.
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.
Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods, 10*, 430–448.
Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.
Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141–152.
Meriac, J. P., Hoffman, B. J., Fleisher, M., & Woehr, D. J. (2007, April). *Expanding the nomological net surrounding assessment center dimensions: A meta-analysis*. Paper presented at the 22nd Annual Conference of the Society for Industrial-Organizational Psychology, New York.
Messick, S. J. (1995). The validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Robie, C., Adams, K. A., Osburn, H. G., Morris, M. A., & Etchegaray, J. M. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance*, *13*, 355–370.
Ryan, A. M. (2003). Defining ourselves: I-O psychology's identity quest. *Industrial-Organizational Psychologist, 41*(1), 21–33.
Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401–410.
Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: SIOP.
Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231–258.
Zedeck, S. (1986). A process analysis of the assessment center method. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior* (Vol. 8, pp. 259–296). Greenwich, CT: JAI Press.