Industrial and Organizational Psychology, **1** (2008), 131–133.
Copyright © 2008 Society for Industrial and Organizational Psychology. 1754-9426/08

Back to Basics of Test Construction and Scoring

MICHAEL T. BRANNICK
University of South Florida

I want to make two main points in response to Lance's (2008) paper. First, we should pay more attention to the psychometrics of our simulations, particularly the reliability of exercise scores related to candidate actions. Poor reliability limits convergent validity and undermines the quality of feedback given to candidates. Second, I suggest 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.

Behavioral Consistency

Behavior in assessment center exercises is so obviously related to job performance that the psychometric considerations seem nearly irrelevant. The beauty of the assessment center's face validity has turned our heads from the mundane calculation of reliability. One of the main reasons that the convergent validities (same dimension–different exercise [SDDE] correlations) are small across exercises is that the behavior of the candidates is inconsistent even within exercises. Social psychologists label our tendency to attribute individual behavior to stable personal characteristics rather than situational determinants as "the fundamental attribution error" (e.g., Jones & Harris, 1967). We believe that a candidate's behavior in an exercise is more informative about character (stable personal characteristics) than is actually the case.

Evidence for poor behavioral consistency can be found by examining correlations both between and within exercises. First, correlations (SDDE) between exercises that are apparently quite similar are often nearly as poor as correlations between exercises that are apparently quite different (e.g., Bycio, Alvares, & Hahn, 1987, found that the average SDDE correlation between two group discussion exercises was .38, whereas the average SDDE correlation between the group discussions and the in-basket was .35). If behavior were determined by the task or role demands of the exercises, and if the exercises provided similar demands, then the correlation between exercises should be large unless, of course, the reliability of the individual exercise scores is low.

Evidence about the consistency of behavior within exercises is surprisingly hard to find and often not estimated even when the study design would allow authors to do so (e.g., Donahue, Truxillo, Cornwell, & Gerrity, 1997; Reilly, Henry, & Smither, 1990). One study that provided such estimates is Brannick, Michaels, and Baker (1989), who devised objective scoring keys for two in-basket forms that could be used to score

Correspondence concerning this article should be addressed to Michael T. Brannick. E-mail: mbrannic@luna.cas.usf.edu

Address: University of South Florida, PCD 4118G, Tampa, FL 33620-7200

Michael T. Brannick, Psychology Department, University of South Florida.

https://doi.org/10.1111/j.1754-9434.2007.00025.x Published online by Cambridge University Press

responses to each item in each in-basket. They found that correlations computed on a subset of items within an in-basket provided correlations similar to those computed between in-baskets and that both were low. Such a result strongly supports the conclusion that actual behavior of candidates is inconsistent. Of course, the in-basket is only one exercise, and such findings bear replication in other exercises.

Instead of the typical assessment center dimension feedback, suppose we move to exercise feedback. For example, suppose we interpret a score on an in-basket for feedback about performance on in-baskets in general. To inform a candidate how well they do on in-baskets generally, we would need to have a generalizable (reliable) score from our in-basket exercise. But if exercises correlate .3 (approximately the typical convergent validity), then we would need 21 in-baskets to achieve a reliability of .90 for a total score used to give accurate feedback. Therefore, the interpretation of exercise scores may be nearly as problematic as the interpretation of dimension scores unless several, perhaps many, of the same type of exercise are included in an assessment center.

At a minimum, 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 exercise 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 (see, e.g., Smith-Jentsch, Salas, & Brannick, 2001). For another example, instead of having one 30-minute performance discussion with a problem clerk, we might have five different 6-minute assessments where the candidate is given instructions to react to specific problems for each subordinate. Or we might have one longer discussion and several shorter ones.

Test Design and Validation Design

Other than poor reliability, the main reason for the poor MTMM results is the mismatch between assessment center exercises and the exercise scoring system. Exercises are typically based on job content, and the scoring system is typically based on traits or constructs. Aligning the exercises or stimulus content with the scoring system will likely improve the pattern of results.

In the assessment center, the exercises are based on work samples, but the inferences are based on knowledge, skill, ability, other characteristics (KSAOs). Work samples are usually designed to assess performance on the task; KSAOs that are needed for the task are presumed to be sampled. In work samples, we typically measure the outcome of the task, not the standing on a number of constituent traits needed for task performance. We do not score a welding test for knowledge of materials, perceptual speed and accuracy, and steadiness of hand (the constituent traits), even though all may be needed. Instead, we score the quality of the weld (the task outcome). Suppose we ask a judge to observe a work sample of welding and judge the candidate for knowledge, perceptual speed, and hand steadiness on the basis of a single weld. We should not be surprised to find that the ratings of the three traits are correlated on a sample of candidates because the *same instance of behavior*, success or failure on the weld, must form the basis of all three judgments.

The reason that the assessment center does not work as it is supposed to is that we have a work sampling approach coupled with a trait-based scoring system. In each exercise, assessors see a limited number of exercise-relevant behaviors by the candidate, and these same few behaviors are the basis for inferring standings on multiple traits. Note that the tasks (exercises) were not designed to separate the traits or to measure specific traits exclusively. If we really wanted to measure traits such as managerial decision making, for example, we would devise a large number of independent decision-making tasks, each designed to tap decision making as well and as *exclusively* as possible. The score on decision making would be the sum of the independent items, and the variance of the scores would depend chiefly

https://doi.org/10.1111/j.1754-9434.2007.00025.x Published online by Cambridge University Press

upon the covariance of the items; that is, the score would represent whatever the decision-making items have in common rather than decision making in a specific context.

Solutions

Assessment center results would be less troublesome if we better aligned the scoring system with the stimulus construction. The simplest solution is to dispense with the assessment center dimensions and score exercises directly (as Lance and others have suggested). If we could measure each exercise with sufficient reliability, we could interpret an in-basket score, a performance discussion score, and so forth. In the validation of such an assessment center, we might, in addition to checking the relations of such scores to overall job performance, compute an MTMM matrix in which tasks were considered traits and methods were a source of ratings (assessment center vs. job). That is, we might have scores on in-baskets, performance discussions, group problem solving meetings, and so forth from both assessment center and job and compare those in an MTMM.

An alternative fix for the troubling correlations is to begin with KSAOs and devise means to measure each. The emphasis here would be on attempts to measure each of the KSAOs as exclusively as possible. For example, persuasive communication might involve short, independent simulations with a boss, peer, subordinate, and customer. Each such discussion would be aimed clearly at persuasion. It is likely that the units or items used to elicit behavior under such an approach might be much smaller (less time consuming) than they are in the traditional assessment center. Advantages to this approach are that the test developer would likely address the internal consistency (or at least alternate forms) of the measures of the traits, and a relatively small number of behaviors would not be the basis of inferences about a relatively large number of traits.

References

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*, 957–963.

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.

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

Jones, E. E., and Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*, 1–24.

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

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

Smith-Jentsch, K., Salas, E., & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support and team climate. *Journal of Applied Psychology, 86*, 279–292.

https://doi.org/10.1111/j.1754-9434.2007.00025.x Published online by Cambridge University Press