ОБЗОРЫ

Evidence That Assessment Center Judgments Measure Dimensions of Management Performance

THORNTON George C. III
PhD, Colorado State University, Fort Collins, CO, USA, George.Thornton@colostate.edu

Ratings from the assessment center method show considerable evidence of construct validity to measure managerial performance dimensions. Analyses of post-exercise dimension ratings demonstrate the effects of performance dimensions, as well as the effects of specific simulation exercises. Analyses of final dimension ratings aggregated across exercises show correlations with expected cognitive ability, related personality characteristics, and performance on related job criteria. This article summarizes very recent studies as well as historical evidence.

Keywords: assessment centers, construct validity.

A controversy has raged for over 25 years about whether assessment centers (ACs) measure intended dimensions of managerial performance. Buttressed by some research which shows that ACs do not have some forms of construct validity, critics have advocated that dimensions be abandoned as the basic architecture of ACs and be replaced by alternative rating structures such roles or exercises. Logic and current research suggests that much of that earlier research does not represent the rationale of the AC method, has limited value, and has even been misguided. More appropriate approaches to study both dimension ratings after each exercise and final dimension ratings aggregated across exercises show considerable evidence of construct validity. Evidence to support construct validity of ACs consists of analyses of ratings within ACs, as well as external analyses of final exercise ratings in relation to other measures of cognitive abilities, personality characteristics, and job performance. This article summarizes research findings, including very recent studies, and concludes that well-run ACs do, in fact, measure an array of performance dimensions important to managerial performance.

The Assessment Center Method

The assessment center (AC) method is a comprehensive and flexible procedure to assess applicants and to assess and develop employees (Thornton & Rupp, 2006). For over 50 years, ACs have been used to assess such dimensions as Problem Solving, management skills such as Planning and Organizing, Oral and Written Communications, and areas of interpersonal effectiveness such as Leadership. An AC involves multiple, trained assessors observing overt behavior relevant to several performance dimensions in simulations of work-related activities. The essential features of an AC (International Task Force, 2009[1]) include:

1 Перевод нормативов опубликован в журнале «Организационная психология»: Международная комиссия по нормативам создания и проведения Центра оценки (2011). Нормативы и этические принципы создания и проведения Центра оценки.

- multiple assessors including line managers, HR staff, industrial/organizational psychologists, or other persons trained in behavioral observation procedures,
- multiple simulation exercises with high fidelity to the target job(s) including such activities as group discussion, presentation, written case analysis, and interviews with a subordinate, supervisor, or client,
- a systematic process for observing and recording behaviors, classifying behaviors into relevant performance-related dimensions or other categories, and
- integration of behavioral observations across assessors and exercises via consensus discussion or arithmetic combination.

For some applications, a final integration yields an overall assessment rating, and/or a set of developmental recommendations.

Evidence of Validity

Validity for any measure has been defined in at least two ways. For many years validity has been defined as sets of evidence that support the inferences to be made from test scores (AERA et al., 1999). Sets of evidence may consist of (a) demonstrations that the content of the test represents content in the performance domain of interest (sometimes referred to as "content validity"), (b) correlations of test scores with criterion measures of performance (referred to as "criterion validity"), or (c) more complex sets of studies showing that test scores demonstrate convergent and discriminant relationships with similar and dissimilar measures (referred to as "construct validity"). Other sources (Thornton & Mueller-Hanson, 2004; Thornton & Rupp, 2006, in press) provide extensive evidence of content and criterion validity. The present article focuses on evidence of construct validity.

The second way to demonstrate that a test is valid to measure an attribute is to show that the attribute exists and variations in test scores produce variation in outcomes (Borsboom, Mellenberegh, & van Heerden, 2004). In this regard, the present article reviews studies that show that AC ratings of dimensions relate to measurement outcomes.

Analyses of Post Exercise Dimension Ratings (PEDRs): "Internal Evidence"

Analyses of ratings on dimensions after each exercise, sometimes called post-exercise dimension ratings (PEDRs), within the AC show evidence of construct validity to measure managerial performance dimensions. Such analyses include (a) correlations of a given dimension assessed in different exercises, (b) multi-trait multi-method (MTMM) analyses, (c) factor analyses, (d) variance components analyses, and (e) generalizability analyses. Discussions of the results and interpretation of such analyses, both critical and supportive of the AC method, can be found in the March 2008 issue of the journal *Industrial and Organizational Psychology*. Results of internal studies show that PEDRs show, to varying degrees, convergence across exercises and common variance due to dimensions. Some studies show that larger portions of variance are due to the simulation exercises. The pattern showing that exercise variance is larger than dimension variance led many to advocate abandoning dimensions as the framework for AC architecture and assessor ratings.

Contrary to those findings, other studies of PEDRs show substantial evidence of construct validity to measure intended dimensions. For example, Guenole and colleagues (2011) found strong

Организационная психология. 1(2). С. 26–44. URL: http://orgpsyjournal.hse.ru/2011--2/44652053.html. – *Прим. ред.*

factor loadings of the expected dimensions in each exercise. These results were obtained when the assessors were trained thoroughly and certified competent to assess a clear set of dimensions, and after taking typical scale development steps. Hoffman et al (2012) found evidence that dimension loadings from factor analyses of PEDRs are equivalent across exercises, and thus it is meaningful to combine PEDRs into across-exercise dimension ratings.

More importantly, Kuncel and Sackett (2012) acknowledged that the typical stream of research on PEDRS, in large part launched by Sackett and Dreher's influential article 30 years ago (Sackett & Dreher, 1982), does not reflect the rationale of the assessment center method and led the field astray. Observations of behavior and ratings of dimension performance after a single exercise are just bits of information that lead to evaluations of overall dimension performance. Kuncel and Sackett conducted analyses which aggregated ratings of a given dimension across exercises to yield overall dimension ratings. They then studied the composition of variance due to dimensions and exercises. When PEDRs are aggregated across exercises, dimension variance dominates over exercise variance when ratings are aggregated across as few as three exercises.

In summary, while performance in ACs is influenced by the exercises which elicit behavior, performance is also influenced by individual difference constructs represented by the dimensions. Thus, while some critics have said that analyses of PEDRs within the AC lead to the conclusion that ACs do not have construct validity to measure managerial dimensions, recent research leads to a very different conclusion: namely, evidence internal to the AC shows that ACs do have construct validity to measure managerial performance dimensions.

Analyses of Final Dimension Ratings: "External Evidence"

Whereas the previous section dealt with evidence internal to the AC, this section deals with external evidence. It examines construct validity evidence of relationships of AC ratings with other measures external to the AC, including evidence related to measurement outcomes.

Final dimension ratings aggregated across multiple exercises show evidence of construct validity in relation to a variety of measures external to the AC. Final dimension ratings can be derived by consensus discussion among assessors, or by arithmetic combination of PEDRs. Evidence includes the correlation of final AC ratings of dimensions in relation to other methods of measuring the same and related dimensions. For example, whereas highly cognitively loaded dimensions (e.g., Decision Making, Planning) correlate with cognitive ability tests, Interpersonal and Behavioral Style dimensions tend to correlate with related personality measures (Shore, Thornton, & Shore, 1990). Comparing validities with sets of ability tests, Thornton et al. (1997) found that assessors' ratings of Routine Problem Solving, Complex Problem Solving, Decision Making, Written Communication, Oral Communication, and Interpersonal Relations correlated more highly with tests of comparable abilities than with non-comparable tests. Those studies were bolstered by two subsequent meta-analyses of other studies (Dilchert & Ones, 2009; Meriac, Hoffman, Woehr, & Fleisher, 2008). Using a set of six core dimensions derived by Arthur, Day, McNelly, and Edens (2003), these studies reported operational and corrected meta-analytic correlations among dimensions, cognitive ability, and Big 5 personality traits. At the construct level, these results show that the dimensions are moderately correlated with either cognitive ability or personality, which is to be expected given the nature of dimensions. For dimensions that were more like cognitive ability, correlations with cognitive ability were shown to be larger than correlations with personality. For dimensions that were more personality-like, correlations with personality were shown to be larger than correlations with cognitive ability. For example, Dilchert and Ones (2009) found the dimension Problem Solving to

relate to cognitive ability (r = .32), but minimally with Big 5 personality traits. The opposite trend was observed for dimensions such as Drive, Influencing Others, and Consideration of Others, which were unrelated to cognitive ability but more strongly related to the personality traits. In a meta-analysis of 65 studies, Meriac and Woehr (2012) found that three factors of AC dimensions correlated in different patterns with external measures: Administrative dimensions correlated more strongly with general mental ability (GMA) than with personality characteristics. In addition, GMA correlated more strongly with Administrative dimensions than with Interpersonal and Activity dimensions.

Table 1. Criterion-Related Validity of AC Dimensions

	Arthur et al., 2003 Meta-analysis[a]	**Meriac et al., 2008 Meta-analysis[a]**	**Connelly et al., 2008 N = 3100**	**Lievens et al., 2009**
Communication	.26/.33	.25/.27	.10	X
Consideration/ Awareness of Others	.20/.25	.22/.24	.16	.21
Drive	.24/.31	.15/.16	.22	X
Influencing Others	.30/.38	.29/.31	.24	.08
Planning and Organizing	.29/.37	.33/.35	X	.15
Problem Solving	.30/.39	.31/.32	.25	.18
Stress Tolerance	X	.16/.18	X	X

Note: [a] Sample-weighted correlation/Corrected value. X = not assessed in this study

Furthermore, the personality characteristic of extraversion correlated more strongly with the dimension of Activity than with the dimension of Administrative. Focusing just on personality correlates, in a study of ACs in Russia, Simonenko, Thornton, Gibbons, and Kravtsova (2012) found that whereas the competency Leadership correlated with empathy and dominance, the competency Thoroughness of Execution correlated with intellectance and conscientiousness. For more information on the application of ACs in Russia, see Simonenko (2011).

Equally important or even more so in terms of demonstrating validity in relationship to measurement outcomes, other studies have examined correlations with non-test measures. Final dimension ratings correlate with and job-related criterion measures such as supervisory ratings of job performance, promotion, and salary. Table 1 summarizes the results of four studies. Arthur et al. (2003) used meta-analysis to collapse the AC dimension labels from 34 articles into six categories. Meriac et al. (2008) employed this dimension framework to meta-analyze data from 48 samples. As shown in Table 1, corrected and uncorrected sample-weighted average correlations were similar in nature, and the results of these two studies appear consistent. Each of the dimensions (with the exception of Stress Tolerance) shows at least a moderately high correlation with job performance, and explains variance beyond each other.

In two subsequent studies, variance in predicting salary due to ratings of dimensions remained after accounting for variance due to exercises. Connelly, Ones, Ramesh, and Goff (2008) derived criterion-related validity estimates from 3100 managers. The correlations for Drive, Influencing Others, and Problem Solving are comparable to the meta-analyses. Lievens, Dilchert, and Ones (2009) made similar estimates from a sample of 520 employees. The correlations were significant but some were lower than those in the other studies cited here.

The unique contributions of dimensions has been demonstrated in studies which show that final dimension ratings add incremental validity in prediction of managerial performance beyond test

measures of cognitive ability and personality. Dilchert and Ones (2009) found that Problem Solving and Communication added the highest incremental value over personality, whereas Influencing Others added the highest value over cognitive ability. When Meriac et al. (2008) regressed job performance onto seven AC dimensions after entering general mental ability and personality measures, the percent of variance accounted for in job performance increased from 20% to 30% (ΔR^2 = .10). Forward selection of the individual dimensions into the regression equation showed that the following ordered six dimensions added significant accuracy: Organizing and Planning, Problem Solving, Drive, Influencing Others, Consideration, and Stress Tolerance. Lievens et al (2009) found that even though four exercise factors explained more variance in the criterion (R = .49) than 4 dimension factors (R = .22), the dimension factors added unique explanatory power.

In summary, these studies involving evidence outside the AC show that ACs have construct validity to measure intended managerial performance dimensions. They build evidence of a logical nomological network of relationships of AC ratings with other measures of related constructs, including important criterion outcomes.

Conclusions and Implications

Theoretically, this body of evidence, and other recent studies, confirms the need to articulate the rationale of the assessment center, the model for constructing the AC for its intended purpose, and the nature of dimensions and competencies being assessed. Subsequently, researchers should conduct studies designed to address questions relevant to those goals. And, analytical and statistical methods that represent the model of assessment method being employed should be used. Then, in evaluating the overall construct validity of the AC, we need to examine evidence relevant to that application.

A number of practical implications are also clear. Foremost, practitioners can be assured the research evidence supports the use of the AC method to assess and diagnose managers' strengths and weaknesses. Such diagnoses are often used to lay out development programs, develop managerial skills, place candidates in optimal job settings, and support various talent management programs.

The assessment center method may not be appropriate to assess some complex competencies. AC can't do everything! Recognize that some competencies are really complex constructs that can be assessed in the form of more than one dimension.

Practitioners should assess only a limited number of dimensions. Assessors seem to have difficulty assessing more than a few dimensions in the typical AC with limited time and resources. How many dimensions is too many? There is no definite answer: some would say ratings from AC method cannot differentiate more than 3 or 4 major constructs; others would say 5 to 7 separate dimensions can be assessed.

Thorough assessor training is necessary, especially if you wish to assess more than a limited number of dimensions. Frame of reference training and certification of the competence of assessors are important, along with the generally recognized need to clearly define dimensions and provide assessment aides.

In conclusion, recent studies have added considerable evidence to the already large pool of evidence (for summaries, see Povah & Thornton, 2011; Thornton & Gibbons, 2009; Thornton & Rupp, 2006, in press) that assessment centers have validity to measure managerial performance constructs. This evidence consists of what has been called content validity, criterion validity, and construct validity. Recent studies reported in professional conferences, recognized journals, and

book chapters written by authorities around the world bring up-to-date evidence that the assessment center method is a valuable tool to promote and advance talent management in global organizations.

References

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

Arthur, W., Jr., 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.

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

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

Dilchert, S., & Ones, D.S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytical incremental validity. *International Journal of Selection and Assessment, 17*, 254–270.

Guenole, N., Chernyshenko, O., Stark, S., Cockerill, T., & Drasgow, F. (2011). We're doing better than you might think: A large scale demonstration of assessment centre convergent and discriminant validity. In N. Povah & G.C. Thornton III. (Eds.). *Assessment and development centres: Strategies for global talent management*. (pp. 15–32). Farnham, England: Gower.

Hoffman, B.J., Baldwin, S.P., Guenole, N., & Cockerill, T. (April 26, 2012). *Resolving the assessment-center construct-validity problem*. In D.J.R. Jackson & B.J. Hoffman, Dimension, task, and mixed-model perspectives on assessment centers. 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

International Task Force on Assessment Center Guidelines (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243–254.

Kuncel, N.R., & Sackett, P.R. (April 26, 2012). *Resolving the assessment-center construct-validity problem*. In D.J.R. Jackson & B.J. Hoffman, Dimension, task, and mixed-model perspectives on assessment centers. 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Lievens, F., Dilchart, S., & Ones, D.S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examination of construct-related and criterion-related validity. *Human Performance, 22*, 375–390.

Meriac, J.P., Hoffman, B.J., Woehr, D.J., & Fleisher, M.S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052.

Meriac, J.P., & Woehr, D.J. (April 26, 2012). *Broad assessment center dimensions: A nomological network examination of validity*. In D.J.R. Jackson & B.J. Hoffman, Dimension, task, and mixed-model perspectives on assessment centers. 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Povah, N., & Thornton, G.C.III. (Eds, 2011). *Assessment and development centres: Strategies for global talent management*. Farnham, England: Gower.

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

Shore, T. H., Thornton, G. C. III, & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101–116.

Simonenko, S.I. (2011). The use of assessment centers and development centres in Russia. In N. Povah & G.C. Thornton III. (Eds.). *Assessment and development centres: Strategies for global talent management*. (pp. 429–439). Farnham, England: Gower.

Simonenko, S.I., Thornton, G.C., Gibbons, A.M., & Kravtsova, A. (April 27, 2012). *Correlates of assessment center consensus dimension ratings: Evidence from Russia*. In A.M. Gibbons, Inside assessment centers: New insights about assessors, dimensions, and exercise. 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Thornton, G.C. III, & Gibbons, A.M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 19*, 169–187.

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

Thornton, G.C. III, & Rupp, D.R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, G.C. III, & Rupp, D.R. (in press). Research into dimension-based assessment centers. In D.J.R. Jackson, C.E. Lance, & B.J. Hoffman. *The psychology of assessment centers*, pp. . New York, NY: Routledge.

Thornton, G.C. III, Tziner, A., Dahan, M., Clevenger, J.P., & Meir, E. (1997). Construct validity of assessment center judgments: Analyses of the behavioral reporting method. *Journal of Social Behavior and Personality, 12*, 109–128.