HUMAN PERFORMANCE, *15*(4), 325–337

Impact of Common Rater Variance on Construct Validity of Assessment Center Dimension Judgments

Nanja J. Kolk
Department of Work and Organizational Psychology
Vrije Universiteit, Amsterdam

Marise Ph. Born
Department of Social Sciences
Erasmus University, Rotterdam

Henk van der Flier
Department of Work and Organizational Psychology
Vrije Universiteit, Amsterdam

In an assessment center (AC), assessors generally rate an applicant's performance on multiple dimensions in just 1 exercise. This rating procedure introduces common rater variance within exercises but not between exercises. This article hypothesizes that this phenomenon is partly responsible for the consistently reported result that the AC lacks construct validity. Therefore, in this article, the rater effect is standardized on discriminant and convergent validity via a multitrait–multimethod design in which each matrix cell is based on ratings of different assessors. Two independent studies ($N = 200$, $N = 52$) showed that, within exercises, correlations decrease when common rater variance is excluded both across exercises (by having assessors rate only 1 exercise) and within exercises (by having assessors rate only 1 dimension per exercise). Implications are discussed in the context of the recent discussion around the appropriateness of the within-exercise versus the within-dimension evaluation method.

Requests for reprints should be sent to Nanja Kolk, Berenschot, Europalaan 40, 3526 KS Utrecht, The Netherlands. E-mail: n.kolk@berenschot.com

In judging an applicant's managerial potential, the assessment center (AC) involves multiple methods (exercises) and multiple traits (dimensions). These traits are evaluated by multiple assessors (psychologists, managers). The AC has received convincing, empirical support for predicting a variety of future job criteria, such as performance, promotion, and salary growth (Gangler, Rosenthal, Thornton & Bentson, 1987; Jansen & Stoop, 2001). However, at the same time, researchers have not been able to show strong evidence for construct validity of the AC dimensions.

Typically, AC construct validity is studied by means of Campbell and Fiske's (1959) multitrait–multimethod (MTMM) matrix, which jointly considers several validity criteria. Evidence for convergent validity is demonstrated when the values on the validity diagonal (monotrait–heteromethod [MTHM]) are significant and large enough to warrant further examination. Evidence for discriminant validity is established when MTHM correlations are larger than the correlations among ratings of different dimensions measured in different exercises (heterotrait–heteromethod [HTHM]). The more rigorous criterion is for the correlations among ratings of different dimension measured in the same exercise (heterotrait–monomethod [HTMM]) to be smaller than the MTHM values on the validity diagonal (Silverman, Dalessio, Woods, & Johnson, 1986). Two decades of research on the AC's construct validity have yielded a rich body of literature, yet all studies seemed to come to the same pessimistic conclusion that different dimensions within exercises correlate higher than corresponding dimensions across exercises, and that construct validity is therefore lacking (e.g., Brannick, Michaels, & Baker, 1989; Chan, 1996; Sackett & Dreher, 1982). This pervasive result has led many of these scholars to disclaim the usefulness of the dimensions as the heart of the AC method (e.g., Robertson, Gratton, & Sharpley, 1987). This article considers an as yet unexamined methodological explanation for the lack of construct validity of the dimensions.

This methodological explanation applies to the vast majority of studies on the AC's construct validity until now, as these have had assessors rate dimensions immediately following each exercise (within-exercise rating method). This generally involves a rotation scheme in which assessors observe each candidate only once. The advantage of this procedure is that it minimizes bias resulting from prior knowledge of the applicant's performance and thus ensures interexercise independence (Andres, Kleinman, 1993; Jones, 1997, p.176; Lievens, 1998). However, it also introduces an inequality in the MTMM matrix among the sources (i.e., the assessors) from which the ratings are obtained, namely the fact that HTMM correlations are based on ratings given by the same assessor (which inflates the correlations), whereas MTHM correlations are based on ratings given by different assessors (which deflates the correlations). Campbell and Fiske (1959) noted that "In practice all that can be hoped for is evidence for relative validity, that is, for common variance specific to a trait, above and beyond shared method variance" (p. 84).

Because in normal practice exercises and assessors are confounded, the AC not only suffers from common exercise variance, but also from common rater variance. This point has been mentioned by Robertson et al. (1987):

> Thus correlations of the same dimensions across exercises represent to a large extent the level of agreement between different raters. The correlations between different dimensions within an exercise, by contrast, are derived from scores produced by the same raters. Certain kinds of rating errors (specifically halo effects) and low interrater reliability would serve to inflate the correlations of different dimensions within an exercise and depress the correlations of the same dimensions across exercises. (pp. 189–190)

Therefore, the possibility of finding construct validity gets suppressed by the traditional rating scheme. For practical reasons, we do not correct for this inequality of common rater variance between and within exercises, as it would involve at least one assessor for each dimension per exercise (Robie, Osburn, Morris, Etchegaray, & Adams, 2000). Nonetheless, for research purposes, it is essential to determine how pervasive this distortion is to be able to estimate its effect on construct validity.

Another motivation to conduct this study is that it directly relates to the manipulation of the traditional AC evaluation process, as Howard (1997), among others, recommended. Specifically, the within-dimension evaluation process is postulated to be conceptually more adequate than the traditional within-exercise process. As said, the within-exercise method requires assessors to observe each exercise and evaluate these independently immediately afterward. The within-dimension method involves postponing evaluation until all exercises have been completed, at which time assessors discuss each candidate's performance per dimension across exercises in a consensus meeting and only then make their independent rating. This thinking was introduced by Silverman et al. (1986), who posited that because the within-exercise method requires assessors to rate each dimension in one exercise, the assessors are forced to process information according to the exercises. Likewise, in the within-dimension method, assessors process information according to the dimensions. Silverman et al. showed that the within-dimension method did indeed result in higher construct validity than the within-exercise method. Later, Harris, Becker, & Smith (1993) challenged this outcome, maintaining that the within-dimension method in the Silverman et al. study artificially led to higher consistency across exercises, because the assessors first arrived at an overall rating and then determined the overall rating per dimension, whereas the assessors in the Harris et al. study first gave a rating per dimension and only then made an overall assessment rating. This study did not result in increased construct validity.

Recently, researchers seem to have regained confidence in the AC's construct validity by applying the within-dimension evaluation method. One study em-

ployed two different analytical procedures to obtain the within-dimension versus the within-exercise method (Adams, 1997). Using the same data set, this study analyzed all the dimension ratings in one exercise and just the ratings on one dimension across exercises, respectively. Results showed that the within-dimension rating procedure resulted in dimension factors, whereas the within-exercise procedure resulted in exercise factors. In a more direct test, Robie et al. (2000) also found that when each assessor truly rated only one dimension in each exercise (within-dimension method), construct validity increased significantly compared to when each assessor rated all dimensions in only one exercise (within-exercise method). Similar findings using the within-dimension method have been found by Lammers and Holling (2000) and Arthur, Woehr, and Maldegen (2000), who had assessors rate each dimension across exercises (c.f., Silverman et al., 1986).

These are important results because they show that construct validity can be much improved by manipulating the rating procedure (either analytically or experimentally). Although these studies all examine slightly different modifications of the within-dimension method proposed by Silverman et al. (1986), for this study we highlight one important element that they have in common, which is that assessors evaluate the candidates' performance in more than one exercise. The most obvious example is the Robie et al. (2000, p. 365) study in which assessors rate one dimension across exercises, thereby eliminating the exercise effect. Although this evidently benefits construct validity, it sacrifices the independence between the exercises. In fact, the within-dimension method introduces common rater variance across exercises either instead of (e.g., Robie et al., 2000) or in addition to (e.g., Silverman et al., 1986) within exercises. For this reason, it remains unclear whether the within-dimension method generally yields better construct validity results than the within-exercise method because the assessors are better able to form their hypotheses clustered on dimensions rather than on exercises, or whether rating candidates across exercises introduces (additional) common rater variance. As for the Robie et al. study, it is not clear whether the increase in construct validity is due to lesser cognitive demands because assessors rate only one dimension per exercise or to cross-exercise bias because assessors rate candidates across exercises, because both features differed between the within-exercise and the within-dimension method.

When common rater variance is introduced across exercises, the estimated correlations between exercises may not only refer to convergent validity, but also to rater-related factors such as halo and cross-exercise bias. If common method variance cannot be avoided, its influence should favor falsification of one's construct-related validity hypothesis, rather than its confirmation. Enhancing convergent validity by allowing common rater variance interferes with this notion. Another confounding methodological issue is that between-exercise correlations depend on interrater reliability of the same dimensions across exercises, which

may deflate convergent validity (Kudish, Ladd & Dobbins, 1997; Lievens, 1999; Robertson, Gratton & Sharply, 1987). The within-dimension procedure obviously avoids this confusion, yet causes another methodological problem at the same time, as it introduces unwanted common rater variance across exercises.

This study uses a design that bypasses this inequality between within-exercise and across-exercise ratings. In doing so, we control for the methodological artifact of common rater variance, which is generally present within exercises but absent across exercises. It is expected that when this inequality between HTMM and MTHM correlations is eliminated, the Campbell and Fiske (1959) MTMM criterion will be met to a larger extent. For research purposes, we propose a design in which each cell in the MTMM matrix is based on ratings given by a different assessor. In doing so, we standardize the rater effect on discriminant and convergent validity. Therefore, unwanted common rater variance is not only excluded across exercises (across-exercise ratings are independent) but also across traits within exercises (within-exercise ratings are independent).

This standardizing of the rater effect on discriminant and convergent validity will be accomplished by means of two independent studies. In the first study, an experiment is carried out in which assessors in the control condition rate the applicants in the traditional way (on each dimension in one exercise), whereas assessors in the experimental condition rate the applicants on only one dimension in one exercise, such that each dimension is rated by another assessor. In the second independent study, we examine an MTMM matrix through two alternative analytical procedures. First, we use the traditional procedure for analyzing the MTMM matrix, using all dimension ratings of each assessor per exercise. Second, we employ an alternative analytical procedure, by using just one dimension rating per assessor per exercise.

We hypothesize that the HTMM correlations (discriminant validity) will be significantly lower in the experimental conditions in both studies, leading to smaller differences with MTHM correlations (convergent validity) and thus to improved construct validity. Because we do not manipulate across exercise dependence in either study, we do not expect changes in convergent validity.

STUDY 1

Method

Post-exercise dimension ratings (PEDRs) for 200 Dutch job applicants (67% men, mean age 35, $SD = 8$) were collected in 2000 and 2001 at a psychological consulting firm. These ratings were part of an AC for evaluating managerial potential for several companies. The AC was developed in accordance with the Guidelines (Task Force on Assessment Center Guidelines, 1989). The assessors rated appli-

cants on a day-to-day basis and were frequently trained on dimensions, exercises, and rater errors as well as on usage of a common frame of reference. The following types of exercises were used: interview simulation exercises, in which the applicant had a one-on-one talk with a subordinate (i.e., a trained role player) about a performance problem; and analysis and presentation exercises, in which applicants were required to study a complex business problem, present the best solution to this problem to the board of directors (i.e., two confederates), and defend this solution during a discussion. The interview simulation exercise tapped the dimensions sensitivity, judgment, and tenacity; and the analysis and presentation exercise tapped sociability, judgment, and tenacity. Although it would have been methodologically preferable if the dimensions sensitivity and sociability were the same in both exercises, we felt that it was justifiable to compute a correlation between these dimensions across exercises, because they are both "interpersonal style dimensions" (cf. Shore, Thornton, & Shore, 1990). The exercises used in this study are common for AC practice (Thornton, 1992). Each exercise took 15 min to prepare and another 15 min to play.

Depending on the target job, the applicants either took part in two interview simulations (with a different content) or in one interview simulation and one analysis and presentation exercise. The first and second exercise both consisted of 25% analysis and presentation exercises and 75% interview simulations. Therefore, the influence of exercise type is equally present in both exercises. This ratio was approximately the same for the control and for the experimental group.

Research design and rating procedure. The assessors in the control group rated applicants in the traditional way, assessing all three dimensions per exercise, whereas the assessors in the experimental group were asked to rate applicants on only one dimension. The applicants' performance was rated on the target dimensions immediately following each exercise. The assessors rotated after each exercise, so that each assessor observed each participant only once. The applicants in the control group were rated by two assessors, which is normal practice, whereas the applicants in the experimental condition were rated by three assessors, which was necessary to cover each dimension. For comparison reasons, the analyses were based on the ratings of one assessor per dimension in both groups. The assessors were blind to the true purpose of the study.

Results

Table 1 shows means, standard deviations, and intercorrelations for the control and the experimental group in Study 1. Visual examination of these MTMM matrices reveal that the HTMM correlations are higher in the control group than in the experimental group (the mean difference was .16). MTHM correlations are approximately equal in both groups. The difference between HTMM and

TABLE 1
Means, Standard Deviations, and Intercorrelations Among Ratings for Study 1

Groups	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*
Experimental								
1. Sensitivity[a]	26.60	8.76	—					
2. Judgment	28.30	8.50	.49**	—				
3. Tenacity	31.85	8.84	.28**	.33**	—			
4. Sensitivity	27.13	9.15	.26*	.37**	.30**	—		
5. Judgment	27.67	9.50	.10	.26**	.17	.52**	—	
6. Tenacity	31.14	8.45	.18	.04	.31**	.16	.44**	—
Control								
1. Sensitivity[a]	30.05	9.42	—					
2. Judgment	27.63	9.28	.71**	—				
3. Tenacity	31.24	9.36	.38**	.54**	—			
4. Sensitivity	27.52	9.89	.28**	.29**	.23*	—		
5. Judgment	28.02	9.49	.20*	.29**	.12	.60**	—	
6. Tenacity	30.40	9.34	.21*	.23*	.39**	.36**	.59**	—

Note. $N = 100$ for both groups.

[a]This dimension is labeled "sociability" in the case of the analysis and presentation exercise. Mean scale ranges from 10 (*poor performance*) to 50 (*good performance*).

*$p < .05$. **$p < .01$.

MTHM correlations was .21 in the control group and .09 in the experimental group. We tested for significance of the differences between the correlation matrices displayed in Table 1 using LISREL 8.20 (Jöreskog & Sörbom, 1989).[1] The HTMM correlations differed significantly between groups, $\chi^2(6, N = 200) = 14.66$, $p < .05$. The results therefore lend support to our expectation that the experimental method, where each rating is given by a different assessor, has im-

[1]First we conducted a confirmatory factor analysis (CFA) on the two groups. The variances of the methods (i.e., the exercises) were set equal throughout the models, for these can be assumed to be roughly similar. The results supported the hypothesis that the experimental group yields higher construct validity. However, we came across severe estimation problems in several of the CFA models, such as out-of-range estimates and convergence problems. These problems are common for studies examining MTMM data (Lance et al., 2000). An alternative to the traditional trait *x* method model is not specifying method factors, but allowing the errors within methods to correlate (i.e., the correlated uniqueness model [CUM]). A limitation of the CUM is that factor loadings and intercorrelations tend to increase spuriously, thus artificially overestimating convergent and underestimating discriminant validity (Kenny & Kashy, 1992). This methodological artifact would be especially unfavorable for this study, as it cannot be determined whether the strength of the overestimation or underestimation is the same across groups. For this reason, we did not consider the CUM as a solution to the empirical underidentification problems we encountered in our initial analyses.

proved discriminant validity over the control group, where ratings within exercises are given by the same assessors.

Discussion

In line with our expectations, we see that there is less distance between HTMM and MTHM correlations in Study 1 in the experimental group than in the control group. Specifically, the HTMM correlations (discriminant validity) are significantly smaller in the experimental group. Therefore, discriminant validity improves when within-exercise ratings, from which these correlations are computed, were obtained by different assessors, compared to the control group, in which these ratings were obtained from the same assessors. We neither hypothesized nor found a significant difference in convergent validity.

Robie et al. (2000, p. 366) acknowledged that cognitive load differed across conditions in their study and therefore suggested examining a design that controls for cognitive load. In Study 1, a difference in cognitive load might be an alternative explanation for this increase in discriminant validity because assessors in the experimental condition rated only one dimension per exercise, whereas assessors in the control condition rated three dimensions per exercise. On the other hand, the assessors in this study may have found it difficult to observe only one dimension per exercise because they observe all dimensions per exercise in their normal day-to-day practice. To be sure, Study 2 examines the same hypothesis, yet using the same regular data set twice: once using each dimension rating from each rating source and once using only one dimension rating per rating source. Therefore, the cognitive demands are the same in both conditions.

STUDY 2

Method

Fifty-two participants (84% men, mean age 24, $SD = 2$) with a master's degree in business economics took part in a 1-day AC developed for evaluating applicants for the position of trainee for a large accountancy firm. The AC consisted of several tests and inventories as well as three exercises: an interview simulation exercise with a subordinate, in which participants were to persuade a subordinate to put in overtime; a simulated customer interview, in which participants were required to first soothe an unsatisfied client and then work together toward a solution; and an analysis and presentation exercise. After all exercises had been completed, the assessors gathered in an evaluation meeting where an overall assessment rating was established, yet for the analyses we used PEDRs. In this study, the dimensions for the interview with a customer and the interview with a subordinate were sensitiv-

TABLE 2
Mean Dimension and Exercise Intercorrelations Among Ratings for Study 2

Items	*Traditional Method*	*Alternative Method*
Dimension (monotrait-heteromethod)		
1. Sensitivity	.31	.36
2. Persuasiveness	.20	.14
3. Tolerance for stress	.17	.18
4. Mean (convergent validity)	.23	.23
5. Mean heterotrait–heteromethod	.14	.15
Exercise (heterotrait–monomethod)		
1. Interview simulation subordinate	.56	.35
2. Client interview	.59	.28
3. Analysis and presentation exercise	.59	.48
4. Mean (discriminant validity)	.58	.37

ity, persuasiveness, and tolerance for stress; whereas for the analysis and presentation exercise, they were persuasiveness, tolerance for stress, and analytical skills. Because not all dimensions were similar across exercises, only the correlations between similar dimensions across exercises were computed.

Research design and rating procedure. The assessors were asked to evaluate the applicants' performance on all dimensions. Different assessor teams were used in each of the exercises. Each assessor team consisted of one psychologist and two managers who were about two levels above the target position. We analyzed this data set in two ways. First, in the traditional way, using the ratings from all dimension ratings from each assessor (i.e., Dimensions 1, 2, and 3 in exercise A from Assessor 1; Dimension 1, 2, and 3 in exercise B from Assessor 2). Second, the experimental way, in which correlations were computed by utilizing only one dimension rating per assessor (i.e., Dimension 1 from Assessor 1, Dimension 2 from Assessor 2, Dimension 3 from Assessor 3, etc.).

Results

First of all, we computed the 27 × 27 MTMM matrix (3 exercises, 3 dimensions, 3 assessors). For brevity's sake, we do not report this matrix here in its entirety,[2] although Table 2 summarizes part of it as an illustration. Table 2 clearly shows that HTMM correlations are higher when they are obtained within assessors (traditional method) than when they are obtained across assessors (alternative method). As expected, MTHM and HTHM correlations are very much the same for both

[2]These data can be obtained from Nanja J. Kolk.

methods. The difference between HTMM and MTHM correlations was .35 when analyzing the MTMM matrix in the traditional way and .14 when analyzing it in the experimental way. As an omnibus test for the complete matrix, we then compared the within-exercise correlations from similar and different assessors, respectively, using LISREL 8.30. The results revealed a significant difference between HTMM correlations obtained from within-assessor ratings as compared to between-assessor ratings, $\Delta\chi^2(9, N = 52) = 83.90$, $p < .001$.

GENERAL DISCUSSION

This article contributes to the literature on AC construct-related validity by showing that the negative difference between MTHM and HTMM correlations often reported in literature (e.g., Sackett & Dreher, 1982) is due in part to common rater variance that is present within but not between exercises. We have accomplished this by using a study design that employs as many assessors as there are dimensions, as suggested by Robie et al. (2000). In this way, common rater variance is excluded within and across exercises. Two independent studies tested this hypothesis: first, through a direct manipulation of the rating procedure (by having assessors rate only one dimension in one exercise); and second, by a manipulation of the analytical procedure (by using only one dimension per assessor per exercise). Both studies lend support to the hypothesis that common rater variance spuriously decreases discriminant validity.

The studies presented here were both modifications of previous studies conducted to scrutinize the benefits of the within-dimension evaluation method over the traditional within-exercise method (Adams, 1997; Arthur et al., 2000; Harris et al., 1993; Robie et al., 2000; Silverman et al., 1986). Although all but one of these studies showed higher construct validity for the within-dimension method than for the within-exercise method, up to this point it could not be determined whether this was due to an actual and meaningful increase or to the methodological artifact of common rater variance across exercises. For instance, Robie et al. concluded in their study that, " the exercise effect … was eliminated due to the fact that the same assessor did not rate all dimensions within one exercise" (p. 365). However, they were not able to determine whether exercise interdependence may not also be responsible for their result.

Study 1 circumvented this confounding situation by having assessors rate only one dimension in one exercise. This study showed that within-exercise correlations decreased compared to the control group where assessors rated all dimensions in one exercise. An alternative explanation for these results may be sought in a difference in cognitive demands between the two groups, rather than in the elimination of common rater variance, because assessors in the experimental group rated only one dimension per exercise, whereas the assessors in the control group rated all di-

mensions per exercise. To rule out this possibility, Study 2 examined the same AC data according to two different analytical procedures: one using each dimension rating from each rating source, and one using only one dimension rating per rating source. Here again, we see that within-exercise correlations differed significantly between groups, while at the same time we can ascertain that this result is not due to a difference in cognitive demands.

Implications

This study shows that common rater variance partly accounts for the pervasive result reported in AC literature that correlations between different dimensions within exercises exceed correlations between similar dimensions across exercises (e.g., Sackett & Dreher, 1982). However, in both studies, HTMM correlations are still higher than MTHM correlations, albeit that the difference is much smaller when common rater variance is no longer present. This may be explained by the fact that although we corrected for common rater variance in the experimental groups, common exercise variance is still present. We also acknowledge that, besides the methodological artifact of common method bias, other factors that have been suggested in literature may also influence discriminant and convergent validity. These are, for instance, conceptual similarity between dimensions, differences in educational background and experience of the assessors (managers vs. psychologists), type of exercise, and so forth (for an overview of these factors, see Lievens, 1998). True performance levels of candidates may also provide an explanation for the typical within- and across-exercise correlation pattern (Lievens, 1999).

It does not follow from this study that future AC architecture should be altered to meet the stringent independence requirements in this study. Evidently, it would be too costly to have as many assessors as there are dimensions. However, a direct implication for researchers is to not be too surprised when their HTMM correlations exceed their MTHM correlations, despite many sound and fruitful manipulations, as this phenomenon is at least partly due to the presence of common rater variance. In view of these results, it is surprising how little is reported in AC literature on the specifics of assessor rotation schemes. To be able to make valid interpretations of the outcomes of construct validity studies, we suggest that future researchers make clear how their AC is designed; whether different assessors are used in different exercises; which rating procedure is applied; and, if an applicant's performance is assessed by more than one rater, what the interrater reliability is.

These results call for a re-evaluation of the appropriateness of the within-dimension method. Because the benefits of this method in terms of construct validity have been established in numerous studies, and because it also seems to be a practical way of evaluating participants' performance, it seems even more valuable to ascertain that common rater variance is not responsible for the increase in construct validity. For this reason, further research is needed that will directly assess the im-

pact of interexercise dependence. We would encourage any study that also incorporates a criterion-related measurement to assess whether interexercise dependence helps or hurts the predictive validity of the AC.

ACKNOWLEDGMENTS

Nanja J. Kolk is now at Berenschot Utrecht, The Netherlands.

This article was presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA, April 2001 in M. Ph. Born (Chair), *Assessment Center Dimension Validation: Are We Asking the Wrong Questions?*

We thank LTP for helping us collect the data for both studies.

REFERENCES

Adams, K. A. (1997). *The effect of the rating process on construct validity: Reexamination of the exercise effect in assessment center ratings.* Unpublished master's thesis, University of Houston, TX.

Andres, J., & Kleinmann, M. (1993). Development of a rotation system for assessors' observations in the assessment center. *Zeitschrift für Arbeits- und Organisationspsychologie, 37,* 19–25.

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

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational & Organizational Psychology, 69,* 167–181.

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

Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78,* 675–678.

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

Jansen, P. G. W., & Stoop, B. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology, 86,* 741–753.

Jones, R. G. (1997). A person perception explanation for validation evidence from assessment centers. *Journal of Social Behavior and Personality, 12,* 169–178.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS Inc.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior & Personality, 12,* 129–144.

Lammers, F., & Holling, H. (2000). Assessor rotation and construct validity of assessment centers. *Zeitschrift für Differentielle und Diagnostische Psychologie, 21,* 270–278.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. B. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13,* 323–353.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152.

Lievens, F. (1999). *An examination of factors which affect the construct validity of assessment centers.* Unpublished doctoral dissertation, University of Gent, Belgium.

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

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13,* 355–370.

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

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

Task Force on Assessment Center Guidelines. (1989). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 18,* 457–470.

Thornton, G. C. (1992). *Assessment centers in human resource management.* Reading, MA: Addison-Wesley.