7-1998

Factors which improve the construct validity of assessment centers: A review

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg
DOI: https://doi.org/10.1111/1468-2389.00085

Published in International Journal of
Selection and Assessment, 1998 July,
Volume 6, Issue 3, Pages 141-152

Factors which Improve the Construct Validity of Assessment Centers: A Review

Filip Lievens*

This article reviews 21 studies which manipulated specific variables to determine their impact on the construct validity of assessment centers. This review shows that the studies regarding the impact of different observation, evaluation, and integration procedures yielded mixed results. Conversely, dimension factors (number, conceptual distinctiveness, and transparency), assessor factors (type of assessor and type of assessor training), and exercise factors (exercise form and use of role-players) were found to moderate construct validity. On the basis of the review, practical recommendations are derived to maximize the probability that practitioners design and administer an assessment center with construct validity. Finally, new perspectives for future research are identified.

Key words: Assessment centers, construct validity, design recommendations

Over the past 40 years assessment centers have established themselves as popular procedures which can serve a variety of human resource functions such as selection and development. It is well established in the literature that assessment centers possess high criterion-related validity (Gaugler *et al*. 1987) and face validity (Macan *et al*., 1994). In the early 1980s, however, questions were raised whether assessment center dimensions did indeed represent meaningful constructs.

Most studies investigated the construct validity of within-exercise dimension ratings through the multitrait-multimethod approach. The following results were consistently found: Ratings on the same dimension across exercises correlated lowly (i.e., low convergent validity), and ratings on different dimensions in a single exercise correlated highly (i.e., low discriminant validity). Similar results were obtained from a more powerful construct validation approach such as confirmatory factor analysis: ratings clustered according to exercises rather than dimensions (see Kauffman *et al*. 1993, for a review of previous studies).

Given these troublesome findings, Klimoski and Brickner (1987) called for (quasi) experimental research on assessment center construct validity, stating:

> In terms of priorities, however, given the real need of organizations to assess potential (apart from competencies), it would seem most important to establish if, or under what conditions, assessment centers can be made to produce valid measures of constructs (...) Specifically, numerous and potentially relevant variables could be experimentally manipulated to determine their impact on discriminant and convergent validities of staff ratings. (p. 255)

A number of studies have followed Klimoski and Brickner's (1987) suggestions and have attempted to increase the convergent and discriminant validities by modifying specific assessment center features. A wide variety of variables have been manipulated. To date, a systematic and comprehensive review of the effects of these different interventions is not present. Nevertheless, identifying under which conditions assessment centers possess construct validity is both of conceptual and practical importance. For instance, it may lead to a reconceptualization of some basic components of assessment centers. Practitioners may benefit from the implications of the available research evidence, as it may give them concrete guidelines on how to increase the quality of construct measurement of assessment centers. Such recommendations may also reduce the considerable variability in center design and administration (see Schmitt *et al*. 1990).

Therefore, this article reviews studies which experimentally manipulated specific variables to determine their impact on the construct validity of assessment centers. Resulting from this review, we aim (1) to propose a set of practical recommendations which should increase the probability to find dimensions with construct

*Parts of this article were presented at the 25th International Congress of Assessment Center Methods, London, UK (1997). Filip Lievens works as a research assistant of the Fund for Scientific Research – Flanders (FWO) at the Department of Personnel Management and Work and Organizational Psychology (University of Ghent).
Address for Correspondence: Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, University of Ghent, Henri Dunantlaan 2, 9000 Ghent, Belgium. Electronic mail may be sent via Internet to [filip.lievens@rug.ac.be].

validity; and (2) to identify new perspectives for future research.

Objectives and Design of Review

To be included in the review, a study must have compared different approaches in design or administration of basic assessment center features. The dependent variable must have provided information about the construct validity of within-exercise dimension ratings.

On the basis of these inclusion criteria, we conducted a search using a number of computerized databases (i.e., Psychlit, the Social Science Citation Index, and Current Contents). Additionally, we scrutinized reference lists from obtained studies to find other published and unpublished studies. At last, researchers in the assessment center domain were contacted to retrieve more unpublished papers.

Description of Studies

Twenty-one studies, dating from 1976 to 1997, conformed to the stated criteria. Fifteen studies were published articles, four were unpublished dissertations and two were conference presentations. Appendix A presents these studies. Per study the assessment center situation and the operationalization of the independent variable are described. For comparison, the results of each study are also synthesized.

Ten studies used a quasi experiment to investigate the impact of specific variables on construct validity, eight studies a laboratory experiment, and three studies a field experiment. In company assessors were used in 13 studies. The total number of assessors ranged from 9 to 179 ($Mdn = 22$). In all studies assessors were trained. Training program length ranged from 3 weeks to 90 minutes ($Mdn = 12$ hours). Eighteen studies used actual assessees, three studies hypothetical assessees. The number of assessees ranged from 2 to 1,758 ($Mdn = 117$). In half of the studies these assessees were first or second line managers, in the other half assessees were students. The median number of dimensions used was 6. The median number of exercises was 3. Assessment center construct validity was investigated by evaluating the degree of convergent and discriminant validity present in the multitrait-multimethod matrix (Campbell and Fiske, 1959) of within-exercise dimension ratings in fourteen studies. Seven studies analyzed this matrix with confirmatory factor analysis, two studies with exploratory factor analysis. Six studies relied on analysis of variance. Finally, three studies used dimensional accuracy (i.e., stereotype and differential accuracy) to gauge some information of construct validity. These accuracy components are said to be related to construct validity (Murphy and Cleveland, 1995, p. 294). Some studies combined two or more of these analyses. Gaugler and Thornton (1989), for example, used multitrait-multimethod correlations, analysis of variance, and accuracy measures.

Categorization of Studies

The studies were sorted into five content categories: dimensions, situational exercises, assessor characteristics, systematic observation and evaluation procedures, and integration of results. These categories were chosen because they represented the corner stones of assessment centers (Thornton, 1992). These five general categories were described as follows:

(1) *Dimensions* consisted of research on the effects of characteristics of the dimensions on the quality of construct measurement in assessment centers. Studies 6, 8, 9, and 10 of Appendix A fell under this category.

(2) *Situational exercises* consisted of research on the effects of exercise factors (i.e., exercise form, exercise content, exercise instructions, role-player standardization, etc.) on the quality of construct measurement. Studies 12, 18, and 21 of Appendix A fell under this category.

(3) *Assessor characteristics* consisted of research on the effects of assessor characteristics and different assessor training approaches on the quality of construct measurement. This content category included studies 4, 13, and 17 of Appendix A.

(4) *Systematic observation and evaluation procedures* consisted of research on the effects of different observation and evaluation approaches on the quality of construct measurement. This fourth category consisted of studies 1, 2, 3, 5, 12, 14, 15, 16, and 20 of Appendix A.

(5) *Integration of results* consisted of research on the effects of consensus discussion and different consensus discussion formats on the quality of construct measurement. This last category covered studies 7, 11, and 19 of Appendix A.

Results of Review

Dimensions

A first series of studies examined how construct validity was affected by the number, the distinctiveness, the nature, and the definition of the assessment center dimensions.

Gaugler and Thornton (1989) demonstrated that the number of dimensions wielded effects on the level of convergent validity. Discriminant validity was not affected. In this study assessors were able to give convergent valid and accurate ratings on three dimensions. When assessors had to deal with six or nine dimensions, this was not possible. These results illustrate that assessors possess limited capacities to process information. Therefore, we concur with Gaugler and Thornton (1989) that assessment center users should limit the number of dimensions to be evaluated. This recommendation is especially relevant for centers conducted for hiring purposes. In another study Kleinmann *et al.* (1995) found higher discriminant validity, when assessors rated assessees on conceptually distinct dimensions. With interchangeable dimensions, assessors provided interdependent ratings which did not differ meaningfully from each other. Therefore, dimensions which are merely variations on the same theme (e.g., flexibility, tenacity, decisiveness, etc.) should be avoided.

Another issue concerns the type of constructs used in assessment centers. In practice there exists a curious similarity across organizations in the set of dimensions used. Hence, conclusions about assessment center construct validity are almost exclusively based on dimensions such as sensitivity, problem analysis, or leadership. These constructs are less stable across situations than, for example, the Big Five personality constructs. Related to this, Russell and Domm (1995) experimented with an assessment center where assessors rated candidates on seven role requirements of the target position. For example, they defined the dimension *initiative* as "the degree to which behaviors influence events to achieve goals by originating action rather than merely responding to events as required on the job of store manager" (p. 30). Similarly, Joyce *et al.* (1994) compared the traditional dimensions to a set of constructs based on the functional structure of managerial work (e.g., internal contacts, performance management, etc.). Nevertheless, within-exercise ratings on these task-oriented dimensions exhibited also weak evidence of convergent and discriminant validity.

The definition and operationalization of dimensions is a last important aspect of construct measurement. This issue will be dealt with at length in the context of behavioral checklists.

In sum, research showed that the quality of construct measurement in assessment centers was affected by the number and the conceptual distinctiveness of dimensions. In particular, limiting the number of dimensions increased convergent validity, and using conceptually distinct dimensions yielded positive effects on discriminant validity. Using tasks as organizing categories in assessment centers did not lead to substantial benefits in terms of construct validity. Related to this, future studies are needed to link the implicit constructs which assessors use in their spontaneous cognitions of managerial effectiveness to the assessment center constructs (see Klimoski, 1993).

Trained Assessors

It has often been emphasized that the quality of assessment centers depends mainly on the quality of the assessors. Surprisingly, only a few studies have investigated the effects of assessor characteristics on construct validity. Sagie and Magnezy (1997) discovered that type of assessor (i.e., managers vs. psychologists) significantly impacted assessment center construct validity. In the ratings of psychologists all five predetermined dimensions were represented. Managers' ratings yielded only two dimensional factors. These findings highlight that psychologists should play a key role in assessor teams. For example, they could serve as coach of line managers or as chair of the discussion session.

Regarding assessor training, operational centers vary greatly in both length and type of training given (Spychalski *et al.* 1997). Dugan (1988) demonstrated that neither length of assessor training nor amount of refresher courses led to a more differential use of the various dimensions. Lorenzo (1984) reported lower dimensional accuracy for assessors who had attended assessor training and had been serving as full-time assessors for at least three months (compared to novice assessors). Thus, the amount of assessor training given does not seem to be an important variable. We were not able to locate studies that investigated the influence of type of assessor training on construct validity. Woehr (1994), however, argues that assessment centers can benefit considerably from experimental research on the effects of rater training strategies. In the following we present the rater training approaches which are generally distinguished (Woehr and Huffcutt, 1994). First, raters may be trained to avoid rating effects (e.g., halo, leniency, etc.). Although this rater error training approach is frequently used in the studies listed in Appendix A, it is inadequate. Research shows that training to avoid rater effects does not lead to more accurate ratings (Bernardin and Buckley, 1984; Woehr and Huffcutt, 1994). A second training approach, behavior observation training, is strongly related to the previous one. This training focuses on strategies to improve observation. It also places a heavy emphasis on avoiding (observational) errors. A third alternative is performance dimension training.

Here, the objective consists in familiarizing assessors with the rating dimensions. Hence, the definition and operationalization of the rating dimensions are the central ingredients. Research shows that performance dimension training leads to more differentiated dimensional ratings (Woehr, 1992). The fourth type of training, also known as frame-of-reference training, elaborates on performance dimension training. Besides increasing understanding of the dimensions, frame-of-reference training attempts to provide raters with the same evaluative standards as a reference for judging performance. Research shows that compared to the other training types frame-of-reference training leads to the largest increase in accuracy (Woehr and Huffcutt, 1994). Furthermore, there is evidence that frame-of-reference training helps raters develop consistent categorization schemes that result in improved dimensional (i.e., stereotype and differential) accuracy (Stamoulis and Hauenstein, 1993). As already mentioned, these two accuracy indexes are said to be closely related to construct validity (Murphy and Cleveland, 1995, p. 294). Most assessor training programs are a mixture of the four training types. Still, emphasis is often laid on the observation oriented part. In order to increase the probability to find construct validity we propose a shift in focus in assessor training programs. We recommend that assessor training builds more on the logic behind frame-of-reference training. More efforts should be undertaken to impose consistent categorization schemes on assessors.

In sum, research revealed that both type of assessor and type of assessor training seriously affected the quality of construct measurement. The research evidence, however, is sparse. Therefore, more research is needed on how construct validity is affected by assessor characteristics. We are also not aware of studies dealing with the influence of assessor selection (e.g., via identification of assessors most likely in need of training in the assessor population) on construct validity.

Situational Exercises

Under this rubric researchers explored how the content, form, and level of standardization of assessment center exercises impacted on the quality of construct measurement.

Assessment center exercises are developed to carefully represent the most important elements of the target job (see e.g., Ahmed *et al*. 1997). On the one hand, this contributes to the job relatedness and predictive power of assessment centers. On the other hand, assessees often perform in a very diverse set of exercises (regarding exercise content, form, and subject matter). According to Schneider and Schmitt (1992) this exercise variance partly explains why assessees perform inconsistently across exercises. Schneider and Schmitt proved that variance due to the form of the exercise in particular (e.g., one-to-one exercises vs. group exercises) prompted assessees to perform differently on the same dimension across exercises. These situationally dependent assessee performances for their part led to ratings which were said to be non convergent valid.

For assessment center developers and users it is important to be aware of these results. When constructing situational exercises, they should always make a trade-off between a large number of structurally different exercises which sample the broad and complex job domain and a smaller number of exercises which may fail to capture all elements of the domain but which generate a large number of dimension relevant behaviors. This potential of exercises to elicit a large number of dimension-related behaviors from assessees has been identified as an important factor with respect to measuring valid constructs. For instance, Reilly *et al*. (1990) proved that there existed a close relationship between the number of behavioral observations generated per exercise and convergent and discriminant validity.

Besides the exercise itself, trained role-players are often used to evoke dimension-related behavior from assessees and to limit unintentional exercise variance. Tan (1996) found empirical support for this practice. Higher convergent and discriminant validity were reported when role-players performed an active role. When the role-player remained rather passive (i.e., did not seek to elicit dimension-related behavior), these validities were very low.

Finally, there has been research on the effects of exercise instructions. Kleinmann (1993) and Kleinmann *et al*. (1996) revealed to assessees which dimensions were measured in the exercises. Assessees were also informed which behaviors were relevant per dimension. Kleinmann and his colleagues assumed that informed assessees would orient themselves more towards the given dimensions and would demonstrate more clearly and consistently the accompanying behaviors. This would enable assessors to differentiate among the dimensions and to rate assessees consistently across the exercises. These hypotheses were empirically supported: Within-exercise ratings for assessees who oriented their behaviors towards the dimensions showed convergent and discriminant validity. Consequently, we highly recommend disclosure of the dimensions to participants of developmental assessment centers. This practice of divulging the dimensions could also be considered in centers for selection purposes.

In sum, recent research confirmed that assessment center construct validity was increased by limiting unintentional exercise variance and by giving assessees increased opportunities to display dimension-related behaviors. The latter may be accomplished by using trained role-players, who actively attempt to elicit dimension-related behaviors and by making the dimensions transparent to assessees. Because of these promising results, future studies should continue to explore the effects of other exercise factors. For instance, the bandwidth of situational exercises could be a very salient exercise factor. Whereas some exercises elicit behaviors relevant to many dimensions, other exercises are more dimensionally pure. Future studies should compare assessment center ratings for both types of exercises in terms of construct validity.

Systematic Observation and Evaluation Procedures

The majority of studies in this category tried to ascertain whether construct validity was affected by the rating format used (e.g., graphical rating scales, behavioral checklists, etc.). The rationale is that these rating aids reduce the cognitively complex task faced by assessors. For example, because behavioral checklists list per dimension (e.g., cooperation) the relevant behavioral observations (e.g., picking up other's ideas and opinions, sharing successful results with others, etc.) for each exercise (in this example a group discussion), assessors do not have to decide anymore on the dimension-relevance of behaviors.

Research on the effects of behavioral checklists on construct validity yielded equivocal results. Reilly *et al*. (1990) reported positive findings: ratings made via behavioral checklists demonstrated higher convergent and somewhat higher discriminant validity. In other studies behavioral checklists only enhanced discriminant validity (Campbell, 1986; Donahue *et al*. 1997). In the latter study convergent validity was even lower. Finally, some researchers concluded that behavioral checklists did not influence construct validity (Louiselle, 1986, March; Sweeney, 1976).

A possible explanation for these mixed results is that the procedures to develop behavioral checklists varied across studies. Usually, subject master experts generated dimension-related behaviors per exercise and agreed upon the retranslation of behaviors to dimensions. This retranslation process served to eliminate 'fuzzy' behaviors and, hence, might impact on convergent and discriminant validity. Some studies developed behavioral checklists without a retranslation procedure (e.g., Donahue *et al*. 1997). The differences in construct validity between the various studies may also relate to the ordering of the behavioral statements. Recently, Binning *et al*. (1997, April) found that the discriminant validity of behavioral checklists only increased when the items were ordered in naturally occurring clusters. The discriminant validity of a randomly-ordered checklist was low. Unfortunately, most studies did not report how the statements were ordered. A final explanation for the inconsistencies relates to the number of behavioral items in checklists. Reilly *et al*. (1990) empirically determined that the optimal number of statements per dimension varied between six and twelve. Outside this range no substantial gains in construct validity should be expected. Hence, only the key behaviors should be listed. Most studies did not report on the number of listed statements.

With regard to systematic evaluation procedures in assessment centers, Sackett and Dreher (1982) point out that two evaluation procedures exist. In the traditional behavior reporting method "evaluation is postponed until the completion of all exercises, at which time the assessors share their observations and rate the candidates on a series of dimensions" (p. 402). According to the within-exercise rating method candidates are rated on each dimension upon completion of each exercise. Silverman *et al*. (1986) argued that rating dimensions after each exercise forces assessors to process information in terms of exercises. A variant of the behavior reporting method, the within-dimension method, showed higher convergent validity and somewhat higher discriminant validity. Recently, this study's methodological adequacy has been criticized. Furthermore, two independent studies failed to replicate the differences found (Harris *et al*. 1993; Kleinmann *et al*. 1994). Hence, it is difficult to give conclusive recommendations about the superiority of either one of the evaluation procedures.

Another element under the rubric of systematically observing and evaluating involves the rotation of assessors through the various exercises. In light of construct validity, it is important to identify a rotation scheme which minimizes rating biases. Andres and Kleinmann (1993) developed a rotation system, which attempts to reduce information overload, contrast effects, halo effects, and sympathy effects. The rotation of assessors then builds upon the following principles: (1) Each assessor observes each assessee exactly once; (2) assessees meet at least twice and not more than four times; and (3) each pair of assessors meet at least twice and not more than four times. Theoretical considerations guided the development of this optimal rotation scheme.

Research which empirically demonstrates the incremental value of this scheme in terms of construct validity is needed.

Finally, Ryan *et al*. (1995) concluded that the impact of videotaping assessees on ratings is minimal. For example, rewinding and pausing the videotape did not increase the dimensional accuracy of assessors. Thus, neither indirect nor controlled observation seem relevant modifications to raise the quality of construct measurement in assessment centers.

In sum, research on the effects of different observation and evaluation procedures has yielded mixed and disappointing results in the assessment center domain. Similar conclusions regarding rating format research have been drawn in the broader performance rating field (see Landy and Farr, 1980; Murphy and Cleveland, 1995).

Integration Procedure

Contrary to research regarding the impact of the integration procedure on predictive validity, only a small number of studies have investigated how the integration procedure influences construct validity. Russell (1985) concluded that factor analyses of across-exercise ratings prior to and after discussion yielded the same underlying structure. Other researchers investigated whether the discussion format carried out effects on assessment center construct validity. Ratings of assessor teams who discussed ratings by exercise were compared to ratings of assessor teams who discussed ratings by dimension. No differences in terms of construct validity were reported (Harris *et al*., 1993; Kleinmann *et al*., 1994).

Discussion

In 1987, Klimoski and Brickner called to investigate under what conditions assessment centers could be made valid measures of constructs. This article reviewed 21 studies that followed Klimoski and Brickner's suggestions. Various factors were found to moderate the construct validity of assessment centers. Hence, on the one hand, it is possible to make practical design recommendations to maximize the probability of finding dimensions with construct validity. On the other hand, such design recommendations may also be interpreted as representing 'cosmetic' changes which do not really advance our understanding of the internal workings of assessment centers. This highlights the need for new research areas and possibilities.

Practical Recommendations

Table 1 summarizes practical recommendations which are derived from the results of the studies reviewed. These recommendations should increase the probability that assessment center dimensions reflect those constructs they are purported to represent. In Table 1 the design considerations are presented per assessment center feature. Nonetheless, it is clear that the recommendations are complementary. For example, a frame-of-reference training helps assessors attend to behaviors listed in behavioral checklists.

Although contact with operational centers confirms that some of the recommendations have already been implemented (e.g., use of a smaller number of dimensions), most of them have yet to be embodied in practice. These recommendations are valuable to assessment center users to benchmark their operational center. Assessment center developers may rely on the design considerations right from the beginning.

Directions for future research

Recently, Landy *et al*. (1994) have called for a roadmap of future research on assessment center construct validity. Based on this review, five research routes seem to emerge. First, the studies of this review focused on factors to improve the construct validity. It is equally important to examine how these factors affect the criterion-related validity. Consider, for example, a recommendation such as using a smaller number of rating dimensions. If practitioners follow this recommendation, dimensions may be eliminated that would allow the entire job domain to be represented. This may impair the predictive power of the center. Another example is the recommendation to disclose the dimensions to candidates. If candidates know on which dimensions and behaviors they will be rated, they could play-act and their behavior could become less representative of their true performance. In short, future research should apply both construct and criterion-related validation strategies to the ratings of a single sample. Related to this, Murphy and Cleveland (1995) have argued for the need to examine simultaneously the construct validity and the accuracy of ratings. Virtually all of the studies listed in Appendix A concentrated solely on construct validity.

A second route for future research pertains to the statistical technique used to analyze multitrait-multimethod data in assessment centers. Most of the studies in this review used Campbell and Fiske's (1959) eyeball approach. An important limitation of this approach is that

Table 1: Summary of research-based recommendations to increase assessment center construct validity

Dimensions
Use a small number of dimensions, especially when assessment centers are conducted for hiring purposes.
Choose dimensions which are conceptually distinct (i.e., which are relatively unrelated to each other).
Define the dimensions in a concrete and job related way.
Assessors
Psychologists should play a key role in assessor teams (e.g., as coach of line manager-assessors).
Focus on the quality of training provided to assessors (instead of the length of training).
Besides other training approaches, ensure to incorporate the ideas behind frame-of-reference training in the training program: Familiarize assessors with the dimensions, performance levels, and impose consistent categorization schemes on them.
Situational exercises
Try to develop dimensionally pure assessment center exercises. Thus, pick exercises which generate a large amount of dimension-related assessee behaviors. Try to avoid 'fuzzy' exercises which elicit behaviors potentially relevant to many dimensions.
Train and try to standardize role-players in order to limit exercise variance.
Use role-players who actively seek to elicit dimension-related assessee behaviors.
Reveal the dimensions (and related behaviors) to the assessees, especially when a developmental assessment center is conducted.
Systematic observation, evaluation and integration procedures
Provide assessors with an observational aid (e.g., behavior checklists that list dimension-related behaviors per exercise).
Operationalize each dimension in the checklists with a minimum of six and a maximum of twelve behaviors. Thus, include only the key behaviors.
Group the checklist behaviors in naturally occurring clusters.
Use a rotation system which minimizes rating biases. For example, the one proposed by Andres and Kleinmann (1993).
Video technology and consensus discussion format do not seem to influence assessment center construct validity.

it does not yield any definite criterion for evaluating the size of convergent and discriminant validity (Bagozzi *et al.* 1991). Confirmatory factor analysis overcomes most of the limitations of Campbell and Fiske's (1959) eyeball approach. Seven of the 21 studies employed this powerful confirmatory approach to test the dimensional model of assessment centers. Unfortunately, several problems have also been noted with this procedure, resulting in ill-defined solutions (Bagozzi *et al.*, 1991; Marsh, 1989). Therefore, future assessment center research should capitalize more on research on the application and development of structural equation modeling techniques to the analysis of multitrait-multimethod matrices. For instance, recently several alternative ways of modeling multitrait-multimethod data by means of confirmatory factor analysis have been proposed. Examples include the correlated uniqueness model (Marsh, 1989), the direct product model (Wothke and Browne, 1990), and the hierarchical confirmatory factor analysis model (Lance *et al.* 1992). Future studies should analyze assessment center construct validity according to these more appropriate models (e.g., Sagie and Magnezy, 1997).

Another interesting technique for understanding sources of variance in assessment center ratings may be generalizability analysis (Cronbach *et al.* 1972). Generalizability analysis may provide construct-related evidence of assessment center validity because it aims to decompose, in any measurement, the observed

variance into components attributable to the underlying attributes (real variance) or components attributable to measurement error (error variance) (Kane, 1982). As opposed to classical test theory, generalizability theory regards this measurement error as *multi*faceted. In this manner it permits the simultaneous estimation of different sources of error (e.g., assessors, exercises, etc.) that may affect assessment center ratings. To our knowledge, no authors have used generalizability theory to garner construct-related evidence for assessment center validity.

Third, up to this point, research which tested for the effects of relevant variables on construct validity has been done in quasi or laboratory experiments. In particular, ten studies of this review designed a quasi experiment, and eight studies conducted a laboratory experiment. Whereas the latter designs were typically vulnerable to external validity concerns, the former designs often lacked control for potential confounds (e.g., assessors were confounded with exercises, true performance levels of candidates were not available, etc.). To side-step both of these pitfalls, future studies could design simulation experiments. A simulation experiment verses the participant into a high-fidelity reconstruction of a real-life situation (see Sackett and Larson, 1991). In the case of assessment centers, this implies that researchers design a simulation of an assessment center which embodies the essential elements of operational centers, without sacrificing the high degree of control inherent in a laboratory study. To date, only one study (Gaugler and Thornton, 1989) designed an assessment center simulation to examine construct validity.

A fourth need is the need for a systematic list of variables to be manipulated. This review illustrates that the studies on the impact of different observation and evaluation procedures yielded rather inconsistent results. Therefore, we believe that future research should primarily focus on how assessor factors (e.g., type of assessor and type of assessor training) and exercise factors (e.g., bandwidth of situational exercises) affect construct validity. After all, previous studies in these areas were promising. In addition, two other factors warrant attention: contextual factors and assessee factors. Regarding the former, the rating purpose could be manipulated. Perhaps, ratings given for developmental purposes will be less dominated by a general factor than ratings for a yes/no decision. Recent research (Kleinmann, 1993) suggests also that assessee characteristics (e.g., impression management strategy, degree of self-monitoring, etc.) may explain why assessees perform differently across exercises. There is virtually no research on how candidate characteristics affect the size of convergent and discriminant validities.

Fifth, in the majority of studies theory-based hypotheses were not formulated to anticipate how specific variables might affect construct validity. Yet, several theoretical models may be used to ground hypotheses (see Lord and Maher, 1990, for a more thorough discussion). For instance, assessment center architects assumed a rational assessor model. Assessors were expected to observe and record all relevant assessee behaviors, to classify them correctly into dimensions, and to combine them objectively into ratings. Conversely, research (e.g., effects of behavioral checklists and number of dimensions) has been most consistent with a model of assessors as limited information processors. Another fruitful model suggests that experienced assessors rely upon highly organized and extensive knowledge structures to rate assessees. Future studies could apply this expert assessor model to predict effects of assessor training approaches. In this vein, training provides novice assessors with accurate and valid schemata to effectively identify and categorize dimension-related behaviors.

References

Ahmed, Y., Payne, T. and Whiddett, S. (1997) A process for assessment exercise design: a model of best practice. *International Journal of Selection and Assessment*, **5**, 62–68.

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

Bagozzi, R.P., Yi, Y. and Philips, L. (1991) Assessing construct validity in organizational research. *Administrative Science Quarterly*, **36,** 421–436.

Baker, T.A. (1986) *Multitrait-multimethod analysis of performance ratings using behaviorally anchored and behavioral checklist formats*. Unpublished Master's thesis, Old Dominion University, Norfolk.

Bernardin, H.J. and Buckley, M.R. (1984) Strategies in rater training. *Academy of Management Review*, **6**, 205–212.

Binning, J.F., Adorno, A.J. and Kroeck, K.G. (1997, April) *Validity of behavior checklist and assessor judgmental ratings in an operational assessment center*. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, St Louis, MO.

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

Campbell, W.J. (1986) *Construct validation of role-playing exercises in an assessment center using BARS and behavioral checklist formats*. Unpublished Master's thesis, Old Dominion University, Norfolk.

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972) *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, John Wiley.

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

Dugan, B. (1988) Effects of assessor training on information use. *Journal of Applied Psychology*, **73,** 743–748.

Gaugler, B.B. and Thornton, G.C. (1989) Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, **74**, 611–618.

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

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

Joyce, L.W., Thayer, P.W. and Pond, S.B. (1994) Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology*, **47**, 109–121.

Kane, M.T. (1982) A sampling model for validity. *Applied Psychological Measurement*, **6,** 125–160.

Kauffman, J.R., Jex, S.M., Love, K.G. and Libkuman, T.M. (1993) The construct validity of assessment centre performance dimensions. *International Journal of Selection and Assessment*, **1**, 213–223.

Kleinmann, M. (1993) Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, **78,** 988–993.

Kleinmann, M., Kuptsch, C. and Köller, O. (1996) Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An international Review*, **45,** 67–84.

Kleinmann, M., Andres, J., Fedtke, C., Godbersen, F. and Köller, O. (1994) The influence of different rating procedures on the construct validity of assessment center methods [In German]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **41,** 184–210.

Kleinmann, M., Exler, C., Kuptsch, C. and Köller, O. (1995) Independence and observability of dimensions as moderators of construct validity in the assessment center [In German]. *Zeitschrift für Arbeits und Organisationspsychologie*, **39**, 22–28.

Klimoski, R.J. (1993) Predictor constructs and their measurement. In N. Schmitt and W.C. Borman. (eds.), *Personnel Selection in Organizations*, 99–135. San Francisco, Jossey-Bass.

Klimoski, R.J. and Brickner, M. (1987) Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, **40**, 243–260.

Lance, C.E., Teachout, M.S. and Donnelly, T.M. (1992) Specification of the criterion construct space: an application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, **77**, 437–452.

Landy, F.J. and Farr, J.L. (1980) Performance rating. *Psychological Bulletin*, **87**, 72–107.

Landy, F.J., Shankster, L.J. and Kohler, S.S. (1994) Personnel selection and placement. *Annual Review of Psychology*, **46**, 261–296.

Lord, R.G. and Maher, K.J. (1990) Alternative information-processing models and their implications for theory, research, and practice. *Academy of Management Review*, **15,** 9–28.

Lorenzo, R.V. (1984) Effects of assessorship on managers' proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology*, **37**, 617–634.

Louiselle, K.G. (1986, March) *Confirmatory factor analysis of two assessment center rating procedures*, Paper presented at the IO/OB Graduate Student Conference, Minneapolis, MS.

Macan, T.H., Avedon, M.J., Paese, M. and Smith, D.E. (1994) The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, **47**, 715–738.

Marsh, H.W. (1989) Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, **13,** 335–361.

Murphy, K.R. and Cleveland, J.N. (1995) *Understanding performance appraisal*. Thousands Oaks, Sage.

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

Russell, C.J. (1985) Individual decision processes in an assessment center. *Journal of Applied Psychology*, **70**, 737–746.

Russell, C.J. and Domm, D.R. (1995) Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology*, **68**, 25–47.

Ryan, A.M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T. and McCormick, S. (1995) Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology*, **80**, 664–670.

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

Sagie, A. and Magnezy, R. (1997) Assessor type, number of distinguishable categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology*, **70**, 103–108.

Schmitt, N., Schneider, J.R. and Cohen, S.A. (1990) Factors affecting validity of a regionally administered assessment center. *Personnel Psychology*, **43**, 1–12.

Schneider, J.R. and Schmitt, N. (1992) An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, **77**, 32–41.

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

Spychalski, A.C., Quinones, M.A., Gaugler, B.B. and Pohley, K. (1997) A survey of assessment center

practices in organizations in the United States, *Personnel Psychology*, **50**, 71–90.
Stamoulis, D.T. and Hauenstein, N.M.A. (1993) Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, **78**, 994–1003.
Sweeney, D.C. (1976) *The development and analysis of rating scales for the Chicago Police Recruit Assessment Center*. Unpublished manuscript, Bowling Green State University, Psychology Department.
Tan, M. (1996) *The effects of role-player standardization on the construct validity of dimensions in assessment exercises* [in Dutch]. Unpublished doctoral dissertation, University of Amsterdam.
Thornton, G.C., III (1992) *Assessment Centers in Human Resource Management*. Reading, Addison-Wesley Publishing Company.
Woehr, D.J. (1992) Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior*, **13**, 357–367.
Woehr, D.J. (1994) Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, **79**, 525–534.
Woehr, D.J. and Huffcutt, A.I. (1994) Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organisational Psychology*, **67**, 189–205.
Wothke, W. and Browne, M.W. (1990) The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrica*, **55**, 255–262.

Appendix A

Summary of studies which investigated assessment center construct validity under different conditions

	Author	Assessment center situation	Description of independent variable	Results
(1)	Baker (1986)	12 trained students rated videotaped assessees on 5 dimensions in 2 group discussions.	Effects of different rating formats (laboratory experiment): GROUP 1: Assessors used BARS to rate assessees. GROUP 2: Assessors were given behavioral checklists to rate assessees.	Discriminant validity for GROUP 1 was higher for the assigned-role group discussion. For GROUP 2 discriminant validity was higher for the nonassigned-role group discussion.
(2)	Binning *et al*. (1997, April)	16 trained assessors rated 1758 assessees on 3 dimensions in 1 group discussion.	Effects of item ordering in behavioral checklists (quasi experiment): GROUP 1: Assessors used a checklist where behaviors were empirically ordered (as determined by their factor loadings). CONTROL: Assessors used a checklist where the behaviors were randomly ordered.	Discriminant validity of GROUP 1 ratings was much higher than CONTROL ratings. Convergent validity was not studied.
(3)	Campbell (1986)	12 trained students rated videotaped assessees on 5 dimensions in 2 role-plays.	Effects of different rating formats (laboratory experiment): GROUP 1: Assessors used BARS to rate assessees. GROUP 2: Assessors used behavioral checklists.	Discriminant validity was higher for GROUP 1 than for GROUP 2, no effect on convergent validity.
(4)	Donahue *et al*. (1997)	41 trained police captains and majors rated 188 candidates for a police promotional exam on 9 dimensions in 4 exercises.	Effects of different rating formats (quasi experiment): GROUP 1: Assessors used 'untranslated' behavioral checklists. GROUP 2: Assessors used graphical rating scales.	In both GROUPS exercise factors rather than dimension factors were found. GROUP 1 ratings showed lower convergent validity but higher discriminant than GROUP 2.
(5)	Dugan (1988)	23 trained third level managers and contract assessors rated 522 assessees on 17 dimensions in 5 exercises.	Effects of length of training (quasi experiment): GROUP 1: Assessors received 2 weeks of training. GROUP 2: Assessors got 3 weeks of training.	No difference between the groups on the extent to which assessors distinguished among the dimensions.
(6)	Gaugler and Thornton (1989)	131 trained students provided within- and across-exercise dimensional ratings of 3 videotaped hypothetical assessees in 3 exercises.	Effects of number of dimensions (laboratory experiment): GROUP 1: Assessors were instructed to rate assessees on 3 dimensions. GROUP 2: Assessors had to rate on 6 dimensions. GROUP 3: Assessors had to rate on 9 dimensions.	Ratings of all GROUPS showed poor discriminant validity. GROUP 1 ratings showed higher convergent validity and dimensional accuracy (i.e. stereotype and differential accuracy) than GROUPS 2 and 3.

Appendix A (continued)

	Author	Assessment center situation	Description of independent variable	Results
(7)	Harris *et al.* (1993)	165 trained in company assessors rated 793 assessees on 7 dimensions in 6 exercises.	Effects of different consensus discussion formats (quasi experiment): GROUP 1: After taking notes in all exercises and discussing ratings by exercise, assessors reached consensus on dimensional ratings within exercises and then on overall dimensional ratings. GROUP 2: After taking notes in all exercises and discussing ratings by dimension, assessors reached consensus on dimensional ratings across exercises and then on overall dimensional ratings.	Discriminant validity was low for both GROUP 1 and 2. Convergent validity was low for both GROUP 1 and 2.
(8)	Joyce *et al.* (1995)	Trained in company assessors rated 152 middle level managers on 7 dimensions in 4 exercises.	Effects of the level of abstraction of dimensions (quasi experiment): GROUP 1: Assessors rated assessees on 7 person oriented dimensions (attributes). GROUP 2: Assessors rated on 7 task oriented dimensions (managerial functions).	Both GROUP 1 and 2 showed weak evidence for both convergent and discriminant validity. Factor analyses of ratings of both groups yielded exercise factors.
(9)	Kleinmann (1993)	9 trained graduate psychologists and psychology students rated 56 business students on 12 dimensions in 5 exercises.	Effects of transparency of dimensions for assessees (quasi experiment): GROUP 1: Assessees did not recognize an identical dimension in two exercises. GROUP 2: Assessees recognized an identical dimension in only 1 of 2 exercises. GROUP 3: Assessees recognized an identical dimension in both exercises.	Assessor ratings of GROUP 1 and GROUP 3 showed more convergent validity than GROUP 2 ratings.
(10)	Kleinmann *et al.* (1995)	15 trained psychology students rated 115 students on 4 dimensions in 4 exercises.	Effects of dependency and observability of dimensions (laboratory experiment): GROUP 1: Assessors rated on independent and poorly observable dimensions. GROUP 2: Assessors rated on independent and highly observable dimensions. GROUP 3: Assessors rated on dependent and poorly observable dimensions. GROUP 4: Assessors rated on dependent and highly observable dimensions.	Effect of dependency of dimensions: GROUP 1 and GROUP 2 ratings showed higher discriminant validity and somewhat lower convergent validity than GROUP 3 and GROUP 4 ratings. No effect of observability.
(11)	Kleinmann *et al.* (1994)	33 trained psychology students rated 60 videotaped students on 3 dimensions in 3 group discussions.	Effects of different consensus discussion formats (laboratory experiment): GROUP 1: After taking notes in all exercises and discussing ratings by exercise, assessors reached consensus on dimensional ratings within exercises. GROUP 2: After taking notes in all exercises and discussing ratings by dimension, assessors reached consensus on dimensional ratings across exercises.	No difference between GROUP 1 and GROUP 2. Both GROUP 1 and GROUP 2 produced a satisfactory convergent and discriminant validity.
(12)	Kleinmann *et al.* (1996)	12 trained postgraduate students rated 119 students on 3 independent and easily observable dimensions in 3 exercises.	Effects of divulging the dimensions to assessees (laboratory experiment): GROUP 1: Assessees knew the dimensions and which behaviors were required to perform well. CONTROL: No such instructions were given.	GROUP 1 ratings (in particular of those assessees who adhered to the instructions) showed both discriminant and convergent validity as opposed to CONTROL.

Appendix A (continued)

Author	Assessment center situation	Description of independent variable	Results
(13) Lorenzo (1984)	80 first and second level managers rated 4 videotaped hypothetical first level managers on 8 dimensions in 1 exercise.	Effects of assessorship (field experiment): GROUP 1: Managers who had attended assessor training and had been serving as full-time assessors for at least 3 months. CONTROL: Managers who had neither attended training nor been working as assessors.	Dimensional accuracy for GROUP 1 was lower than CONTROL but no significant difference.
(14) Louiselle (1986, March)	16 trained assessors rated 60 assessees on 5 dimensions in 3 role-plays.	Effects of behavioral checklists (quasi experiment): GROUP 1: Assessors used behavioral checklists. CONTROL: No use of behavioral checklists.	CONTROL ratings showed evidence for both exercise and dimension factors. No interpretable solution was found for GROUP 1 ratings.
(15) Reilly *et al.* (1990)	10 trained in company assessors rated 355 assessees on 8 dimensions in 8 exercises.	Effects of retranslated behavioral checklists (quasi experiment): GROUP 1: Assessors used retranslated behavioral checklists. CONTROL: No use of behavioral checklists.	GROUP 1 ratings showed a large increase in convergent validity and somewhat higher discriminant validity than CONTROL ratings.
(16) Ryan *et al.* (1995)	179 trained psychology students rated videotaped performances of 2 hypothetical assessees on 6 dimensions in 1 group discussion.	Effects of video technology (laboratory experiment): GROUP 1: Assessors viewed a live group discussion. GROUP 2: Assessors viewed a videotape of the same discussion. GROUP 3: Assessors had also opportunities to rewind and pause the tape.	No effect of use of video technology on dimensional accuracy.
(17) Sagie and Magnezy (1997)	105 trained assessors rated 425 students on 5 dimensions in 3 exercises.	Effects of assessor type (quasi experiment): GROUP 1: 39 psychologists. GROUP 2: 66 senior managers.	Factor analyses of GROUP 1 ratings yielded the 5 dimension factors. Only two dimension factors were found for GROUP 2 ratings.
(18) Schneider and Schmitt (1992)	21 recruited and trained assessors rated 89 videotaped students in 3 dimensions in 4 exercises.	Effects of exercise form and exercise content (laboratory experiment): Two levels of FORM (role play vs. group discussion) were crossed with two levels of CONTENT (competitive vs. cooperative)	Exercise FORM accounted for a significant proportion of method bias (i.e., lack of dimensionality in ratings). No effect of exercise CONTENT.
(19) Silverman *et al.* (1986)	24 trained in company assessors rated 90 assessees on 6 dimensions in 3 exercises.	Effects of different scoring methods (field experiment): GROUP 1: After taking notes in all exercises and discussing ratings by dimension, assessors reached consensus on overall dimensional ratings and then gave privately dimensional ratings per exercise. CONTROL: Directly after each exercise, assessors independently gave dimensional ratings.	GROUP 1 ratings showed higher convergent validity than CONTROL. For GROUP 1 discriminant validity was also somewhat higher.
(20) Sweeney (1976)	Trained police officer and civilian assessors rated 186 police recruits on 7 dimensions in 3 individual and 3 group exercises.	Effects of different rating formats (quasi experiment): GROUP 1: Assessors used graphic rating scales to rate the assessees. GROUP 2: Assessors used behavior checklists.	For both GROUP 1 and GROUP 2 convergent and discriminant validity was low.
(21) Tan (1996)	Trained assessors rated 48 candidates on 4 dimensions in 2 exercises.	Effects of role-player standardization (field experiment): GROUP 1: Role-player was more passive. GROUP 2: Role-player played a more assertive role.	For GROUP 1 convergent and discriminant validity was very low. Validities were higher for GROUP 2.