DOCUMENT RESUME

ED 393 92· TM 024 937

AUTHOR Donahue, Lisa M.; And Others
TITLE An Analysis of the Effects of Untranslated Behavioral Checklists on the Psychometric Properties of Assessment Centers.
PUB DATE Jun 95
NOTE 41p.; Paper presented at the Annual Meeting of the International Personnel Management Association Assessment Council Conference (New Orleans, LA, June 25-29, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Assessment Centers (Personnel); *Behavior Rating Scales; *Check Lists; *Evaluation Methods; Field Tests; Occupational Tests; Personnel Evaluation; Personnel Selection; *Police; Psychometrics; Scores; Simulation; *Validity

ABSTRACT

A field study with 178 candidates for a police promotional examination was conducted to investigate the effects of "untranslated" behavioral checklists on certain psychometric properties of an assessment center. The untranslated checklist used all behavioral responses elicited by the assessment center exercises, not just those that met a retranslation criterion of categorizing into dimensions. The study examined whether similar convergence of dimensions across exercises could be obtained across four job simulation exercises that varied greatly in content. The reliability of the behavioral checklist and criterion-related validity were evaluated by comparing the checklist to a conventional graphic rating scale format. The results suggest that the untranslated behavioral checklists improved the discriminant validity and reliability of dimension scores over a traditional graphic rating scale, but did not have a corresponding effect on the convergent validity of dimension scores. In addition, the untranslated behavioral checklist did not yield a significant relationship with performance. It is suggested that behavioral checklists have many benefits, and thus are very appropriate as a method of evaluating assessment center exercises. An appendix lists definitions of assessment center dimensions. (Contains 7 tables and 21 references.) (Author/SLD)

An Analysis of the Effects of Untranslated Behavioral Checklists on the Psychometric Properties of Assessment Centers

Lisa M. Donahue, Melanie C. Jones

New Orleans Civil Service Department

Donald M. Truxillo, Ph.D.

Portland State University

Nancy B. Goldstein

Tulane University

Paper presented at the annual International Personnel Management Association (IPMA) Assessment Council Conference on Public Personnel Assessment, New Orleans, LA, June 1995.

Abstract

A field study was conducted to investigate the effects of "untranslated" behavioral checklists on certain psychometric properties of an assessment center. The results suggest that the untranslated behavioral checklists improved the discriminant validity and reliability of dimension scores over a traditional graphic rating scale, but did not have a corresponding affect on the convergent validity of dimension scores. In addition, the untranslated behavioral checklist did not yield a significant relationship with performance. It is suggested that behavioral checklists have many benefits, and thus are very appropriate as a method of evaluating assessment center exercises. Possible areas of future research and additional practical considerations are discussed.

An Analysis of the Effects of Untranslated Behavioral Checklists on the Psychometric Properties of Assessment Centers

Since the 1950s, assessment centers have been used extensively for purposes such as selection, promotion, training, and career development (Schmidt, Ones, & Hunter, 1992) for a wide variety of positions from professional-level personnel to production line workers (Reilly, Henry, & Smither, 1990). They have been used in a host of organizational settings including manufacturing, government, military, and educational settings (Klimoski & Brickner, 1987). A recent meta-analysis of assessment center validity conducted by Gaugler, Rosenthal, Thornton, and Bentson (1987) supports the wide-spread use of assessment centers for predicting performance in a variety of jobs and organizations. Specifically, Gaugler et al. (1987) reported an average corrected validity coefficient of .37 for the fifty assessment centers examined. These meta-analytic results and others (Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984) have established the predictive validity of assessment centers.

Construct Validity of Assessment Center Exercises

However popular and effective a selection and diagnostic tool, assessment centers remain "the modern enigma in human resource practice" (Klimoski & Brickner, 1987). Very little is known about why assessment centers yield predictive validity. Assessment centers are designed to produce standardized measures of separate constructs thought to represent various job-related

abilities associated with successful performance (Byham, 1980). However, the preponderance of research suggests that assessment centers may not measure the constructs they purport to measure (Sackett & Dreher, 1982; Turnage & Muchinsky, 1982; Silverman, Dalessio, Woods, & Johnson, 1986; Robertson, Gratton, & Sharpley, 1987; Bycio, Alvares, and Hahn, 1987). Indeed, this problem with construct validity led Sackett and Dreher (1982) to conclude that there was "virtually no support for the view that the assessment center technique generated dimension scores that can be interpreted as representing complex constructs" (p. 409). Thus, the predictive validity of assessment centers cannot be attributed to the successful measurement of job-related abilities, but instead is thought to be associated with factors that are not well-understood (Klimoski & Brickner, 1987).

Many investigators (e.g., Sackett & Dreher, 1982; Silverman et al., 1986; Turnage & Muchinsky, 1982) have used the multitrait-multimethod matrix approach (cf. Campbell & Fiske, 1959) to examine the construct validity of assessment centers. To meet the requirements of construct validity according to this approach, ratings on the same dimension should be significantly correlated across exercises (convergent validity). In addition, these across-exercise correlations must be greater than the within-exercise correlations among different dimensions (discriminant validity). The multitrait-multimethod research investigating the internal construct validity of assessment centers, however, has consistently shown higher within-exercise correlations of different dimensions than across-exercise correlations of the same dimensions (Robertson et al., 1987; Sackett & Dreher, 1982; Silverman et al., 1986; Turnage & Muchinsky, 1982).

Neidig and Neidig (1984) contend that the high within-exercise correlations among different assessment center dimensions reported in the above research do not demonstrate measurement error, but instead represent a true exercise effect. According to Neidig and Neidig, the inclusion of multiple exercises in an assessment center is intended to assess behavior in a variety of job-related contexts, and "stable performance across exercises by all participants is not necessarily expected" (p. 184). Some assessees, for example, may be more effective in group exercises, whereas others may perform best on individual exercises. Thus, a lack of consistency in assessee behavior across exercises may reflect differences in individual effectiveness in various situations (Neidig & Neidig, 1984). Furthermore, Neidig & Neidig contend that the lack of behavioral consistency across exercises may also be related to the situational specificity of the exercises. Thus, the situational context determines the manifestation of a dimension in terms of specific behaviors. Leadership behaviors for one assessment center exercise, for instance, may be manifested very differently from those of another exercise because of the situational specificity of the exercises. Thus, the lack of convergence of dimension ratings across exercises may not be a flaw in assessment centers, but instead may represent an expected lack of consistency in assessee behavior due to the situational specificity of assessment center exercises (Neidig & Neidig, 1984).

Unlike Neidig and Neidig, Sackett and Dreher (1982, 1984) are concerned with the inability of assessment centers to measure intended constructs and seriously question the psychological meaning of assessment center dimension ratings that are virtually uncorrelated. Thus, Sackett and Dreher (1982, 1984) argue against assessment center designers "claiming to measure the intended constructs on content-validity grounds when the available empirical

evidence does not support the consistency of dimensional performance across exercises" (p. 187). Given the lack of construct validity evidence, Sackett and Dreher conclude that content-oriented exercise design is not sufficient to demonstrate the job-relatedness of assessment center exercises. Alternatively, they contend that additional validation evidence, either construct or criterion-related, is crucial.

In response to Sackett and Dreher's (1982) conclusions, Neidig and Neidig (1984) argued that the inability of assessment center exercises to meet the requirements of construct validity does not call into question the job-relatedness of assessment center methods. Given the overwhelming evidence of an exercise effect and the fact that assessment center exercises are essentially samples of job-related behavior, Neidig and Neidig (1984) and others (Byham, 1980; Haymaker & Grant, 1982; Jaffee & Sefcik, 1980; Schmitt & Noe, 1983) believe that the job-relatedness of assessment centers should be established on the grounds of content validity by treating individual exercises as work sample tests and using Subject Matter Experts to document the relationship between the content of the job, the assessment center dimensions, and the nature of the exercises (Byham, 1980; Haymaker & Grant, 1982). Sackett and Dreher (1984) concede that when assessment centers are used as a sample of present behavior, rather than a sign of future performance, a content validation strategy may be most appropriate for demonstrating the job-relatedness of assessment center exercises.

Construct Validity and Assessor Cognitive Demands

Some researchers have not abandoned attempts to establish the construct validity of assessment centers in favor of relying on content-oriented construction to demonstrate the job-

relatedness. Instead, these researchers contend that the failure of assessment centers to produce convergent validity of dimension ratings is due to the cognitive complexity of the rating task (Gaugler & Thornton, 1989; Reilly et al., 1990; Silverman et al., 1986). According to Gaugler and Thornton (1989), the job of assessors is exceedingly complex and may overwhelm their limited information processing capabilities. In a typical assessment center, for instance, assessors must observe and record the performance of candidates on situational exercises, classify the observed behaviors into dimensions, and then rate each candidate on each dimension (Gaugler & Thornton, 1989).

Recently, much attention has been given to the influence of assessment center methodology, (Gaugler & Thornton, 1989; Silverman et al., 1986; Reilly et al., 1990) on the cognitive complexity of the rating task and the construct validity of dimension ratings. For example, the research conducted by Reilly et al. (1990) is of particular interest to the present study because Reilly et al. investigated the effects of a behavioral checklist rating scales on the convergent and discriminant validity of assessment center dimension ratings. To develop the checklists, Reilly et al. instructed assessors to completed a first set of assessments for two group exercises and identified specific behavioral responses to each of the exercises "that, when they occurred, caused them to judge an assessee as being higher or lower" (p. 74) in a particular dimension. Using the Smith and Kendall (1963) retranslation technique, the assessors then categorized the behaviors into dimensions within each exercise. The final behaviors comprising the checklist for each of the group exercise were those that met a criterion of 80% agreement among the assessors.

Reilly et al.'s (1990) findings suggest that the introduction of the behavioral checklist significantly improved convergent validity over prechecklist ratings (from .24 to .43). In addition, the convergent validity of their dimension ratings was slightly higher than the discriminant validity of assessment ratings (.43 versus .41). On the basis of these results, Reilly et al. concluded that the use of behavioral checklists alleviated the cognitive demands placed on raters by focusing their attention on specific sets of behaviors relevant to the dimensions assessed. By categorizing behaviors by dimension, the retranslation process further reduced cognitive processing by eliminating the need for assessors to classify behaviors into their relevant dimensions.

While Reilly et al.'s (1990) results are encouraging from the standpoint of improving the pattern of convergent and discriminant validity among assessment center dimensions, the retranslation procedure employed by Reilly et al. eliminated 111 critical behavioral responses identified by the assessors. The omission of these behaviors from the behavioral checklist scales calls into question not only the content validity of the rating procedure, but also its fairness in evaluating candidates, since no credit would be given for responses not meeting the 80% retranslation criterion. Silverman et al. (1987) remind us of the importance of evaluating assessment center methodology on overall dimension scores because selection and promotion decisions ultimately rests on these overall ratings. Thus, a method that would simultaneously improve the construct validity of dimension ratings while giving credit for all content-valid responses elicited by candidates would satisfy not only the requirements of construct validity, but also those of content-oriented test construction.

The present study involves the evaluation of a method that attempts to obtain similar gains in convergent validity and discriminant validity as those reported by Reilly et al. (1990) without sacrificing content validity and fairness in evaluating candidates. Specifically, the present study explored the effects of an "untranslated" behavioral checklist on the construct validity of dimension ratings. The untranslated behavioral checklist used in the present investigation attempted to included all behavioral responses elicited by the assessment center exercises, not just those meeting a retranslation criterion. In addition, the present study attempted to extend the work of Reilly et al. (1990) in two critical areas not explored. First, Reilly et al. investigated the convergence of dimension ratings for two group exercises, both of which involved an assembly problem. Thus, the convergence of dimension ratings could be expected given the similarities in the situational contexts of the exercises. The present study examined whether similar convergence of dimensions across exercises could be obtained across four job simulation exercises that varied greatly in content. Second, Reilly et al. (1990) were unable to present findings regarding certain psychometric properties of the behavioral checklist. The present study evaluated both the behavioral checklist's reliability and criterion-related validity by comparing the checklist to a conventional graphic rating scale format on these psychometric characteristics. Criterion validity is of special interest given the argument that the predictive validity of assessment centers may be related to subtle criterion contamination (Klimoski & Brickner, 1987). Reilly et al. (1990) suggest that criterion validation studies including both a behavioral checklist and conventional scale ratings may help to determine if the well-established relationship between overall dimension ratings and performance is actually due to criterion contamination in which the ratings capture subtle factors that are unrelated to

effective task performance (e.g., presentation skills), but may attract high performance ratings in an organizational setting.

Given the evidence of the reliability (Neidig & Neidig, 1984) and criterion validity (Gaugler et al., 1987; Hunter & Hunter, 1984; Schmitt et al., 1984) of assessment center ratings in the existing literature, similar results were expected regarding the effects of the "untranslated" behavioral checklist on these psychometric properties. In addition, it was hypothesized that the use of the "untranslated" behavioral checklist would produce similar patterns of convergent and discriminant validity reported by Reilly et al. Such findings were anticipated given that the "untranslated" behavioral checklist scales were expected to offer the same "cognitive-reduction" advantages as those constructed by Reilly et al. through the use of the retranslation procedure. Specifically, these advantages include (1) focusing assessors' attention directly on specific behavioral responses elicited by the exercises and (2) organizing these behaviors according to the operational definitions of the dimensions, thereby eliminating the need for such categorization by the assessors. Such results would not only support Reilly et al.'s findings and bolster the use of behavioral checklists for improving the construct validity of assessment center exercises, but would also eliminate the need for the retranslation process in developing behavioral checklist scales. Eliminating this process would simultaneously address the concerns associated with the content validity and fairness of the evaluation process.

METHOD

Participants

Assessees (N = 178) were candidates for a Police promotional examination. The candidates included 164 males, 14 females, 132 whites, and 46 minorities. Assessors (N=41) were Captains and Majors representing various police departments across the country. The assessors included 34 males, 7 females, 29 whites, and 12 minorities. Each candidate was assessed by a team of two assessors assigned to one of four situational exercises comprising the assessment center. The number of teams for each exercise ranged from three to eight teams depending on the complexity of the exercise and the rating task. Each team of raters evaluated an average of forty-three (43) candidates using both a behavioral checklist and a graphic rating scale.

Procedure

Exercises and dimensions. The pairs of assessors observed and rated the candidates on one of four job simulation exercises. The exercises consisted of three situational videos and an in-basket, and were developed on the basis of job analysis information and direct input from local subject matter experts (SMEs). Each of the situational videos depicted job-related scenarios that unfolded across multiple scenes. The scenarios portrayed in the situational videos included: *apprehending a fleeing suspect involved in an armed robbery*, *counseling a subordinate with a suspected drug addiction*, and *directing crowd control activities at an abortion clinic protest*.

In responding to the situational videos, the candidates were required to assume the role of the target position, analyze the situation presented in each scene, and state the actions they would take in response to the situation. Time limits for each scene varied according to their complexity from two to four minutes. Candidates' responses to each scene of the situational videos were videotaped to be subsequently rated by assessors.

The in-basket exercise included a sample of the memos, forms, reports, and other paperwork typically found in the target position's in-basket. Additional job-related situations were also presented in the in-basket. Examples of these situations include: *evidence of declining performance of an officer*, *information suggesting the need for platoon training in report writing*, and *indications of possible sick leave abuse by some platoon members*.

The candidates were given two and a half hours to analyze all of the in-basket items and to prepare their responses to the items. The candidates were then given forty-five minutes to present their responses, which were also videotaped to be rated later by assessors.

The above job simulation exercises were designed to measure nine dimensions identified through job analysis procedures as representing job-related abilities required for effective performance in the target position. These dimensions were (1) *Interpersonal*; (2) *Development of Subordinates*; (3) *Leadership and Delegation*; (4) *Problem Analysis & Decision Making*; (5) *Organization & Coordination*; (6) *Investigation & Police Work*; (7) *Oral Communication*; (8) *Control and Follow-Up*; and (9) *Use of Police References and Quantitative Resources*. (See Appendix for the definitions of the dimensions.)

All of these nine dimensions were measured in the in-basket exercise and one of the situational video exercises. The two remaining video exercises measured only the first eight

dimensions and did not assess *Use of Police References and Quantitative Resources*. A tenth dimension identified through the job analysis, *Written Communication*, was measured by a written exercise contained within the in-basket. This dimension was not included in the present study, however, because it was not assessed in more than one exercise.

Behavioral Checklist construction. Behaviorally-specific responses used to develop the behavioral checklist scales for each of the four exercises were generated from two sources. First, prior to the administration of the assessment center local SMEs who assisted in the development of the exercises were polled for examples of poor, average, and excellent responses to the exercises. The SMEs were then asked to (1) categorize the responses into their relevant dimensions and (2) assign the responses a weight on a scale from -1 to 3, where "-1" is *a response that would have an adverse or negative affect on the situation*, "0" is *a response that would have no affect on the situation*, "1" is *a response that is the least preferable or acceptable in the situation*, "2" is *an average or standard response for the situation*, and "3" *is an excellent response in the situation*. The SMEs' assignment of dimension and ratings to each response was then used to assist the test development staff in developing the behavioral checklist scales for each exercise of the assessment center.

The candidates represented the second source of behavioral responses used in the construction of the behavioral checklists. Following the administration of the assessment center, test development staff members listened to a random selection of candidates' taped performances in order to collect additional responses to each of the exercises. Approximately five percent (N=40) of the candidates' taped performances were reviewed until novel responses were no longer identified.

Because test development procedures precluded using local SMEs after the administration of the assessment center, test development staff members were used to categorize candidate generated responses into dimensions. Following exercise-specific training (described below), the assessors were presented with both the SME- and candidate-generated responses as well as the value of each response assigned by the SMEs. They were then instructed to independently weight each response using the -1 to 3 scale described above. A round-robin technique was then used in which individual assessors stated their ratings. Assessors then engaged in discussion regarding discrepancies in their ratings until a consensus was reached. The final weight for each checklist response was the consensus weight assigned by the assessors. The combined use of these two sources of responses ensured that a near exhaustive list of content valid responses was included on the behavioral checklists.

Assessor training. All teams of assessors participated in a one day training session designed to standardize evaluation approaches and increase the accuracy of ratings. The training session was divided into two segments which included general and exercise-specific training. In general training, assessors received information regarding the target position and structure of the organization. In addition, assessors were instructed on methods of observation and notetaking, the use of rating forms, and the consensus process. Exercise-specific training consisted of reviewing the operational definitions of the dimensions to be assessed, in addition to reviewing each scene of the job simulation exercise and its corresponding behavioral checklist scales. At the end of exercise-specific training, assessors independently rated a hypothetical candidate. Discussion then followed in which the assessors received feedback regarding the accuracy of their ratings.

Rating Procedures. Pairs of assessors for each job simulation exercise observed and recorded the candidates' responses to each scene. Because the candidates' performances were videotaped, the assessors could review their responses as many times as necessary. Immediately after reviewing each candidate's presentation of responses to a scene, the assessors independently completed the behavioral checklist scales for the scene by marking all responses elicited by the candidate. Following the completion of the behavioral checklist, assessors completed the graphic rating scale by making an overall rating for each dimension assessed by the exercise.

Final dimension scores for the behavioral checklist were calculated by summing individual response scores (weighted -1 to 3) for each dimension. These behavioral checklist sums (BCS) for the dimensions were then combined across the two assessors. Dimension scores for the graphic rating scales were represented by the overall rating given each dimension. These ratings were also combined across the two assessors.

Results

Overview

Several analyses were performed to determine if the use of the untranslated behavioral checklist was associated with improvements in the psychometric properties of assessment center ratings. First, both the multitrait-multimethod matrix (cf. Campbell & Fiske, 1959) and factor analysis approaches were used to compare the construct validity of the behavioral checklist sums (BCS) to the graphic rating scale dimension scores. Second, both internal consistency (coefficient alpha) and interrater reliability indices were used to examine the reliability of the BCS in relation to the graphic rating scale dimension scores. Finally, a concurrent validation approach was used to identify any differences in the criterion validity of the total scores produced by the behavioral checklist and the graphic rating scale.

Analysis I: Multitrait-Multimethod Matrix

The convergent and discriminant validity of the BCS and graphic rating scale dimension scores was calculated according to the multitrait-multimethod matrix approach (cf. Campbell & Fiske, 1959). To examine the scales' convergent validity, the correlations among the same dimensions measured across each of the exercises (monotrait-heteromethod correlations) were calculated. These mean monotrait-heteromethod correlations for each dimension of the behavioral checklist and graphic rating scale are listed in Table 1. As Table 1 illustrates, the

---

Insert Table 1 about here

---

grand mean monotrait-heteromethod correlation of the BCS and the graphic rating scale dimension scores was .254 (range=.079 to .336; SD=.074) and .301 (range=.236 to .376; SD=.042), respectively.

The discriminant validity of the dimension scores produced by the behavioral checklist and graphic rating scale was assessed using two methods recommended by the multitrait-multimethod matrix approach (cf. Campbell & Fiske, 1959). First, correlations among different dimensions measured in the different exercises (heterotrait-heteromethod correlations) were calculated and compared to the monotrait-heteromethod correlations. As presented in Table 1, the grand mean heterotrait-heteromethod correlation was .236 (SD=.090) for the BCS and was .297 (SD=.074) for the graphic rating scale. Table 1 also illustrates that these grand mean heterotrait-heteromethod correlations were slightly lower than their monotrait-heteromethod counterparts, especially for the BCS. This suggests some evidence of discriminant validity for both the BCS and graphic rating scale dimension scores.

The second, more stringent method used to assess the discriminant validity of the BCS and the graphic rating scale dimension scores involved calculating the correlations among the different dimensions within each of the exercises (heterotrait-monomethod correlations). The mean heterotrait-monomethod correlations for each of the exercises appear in Table 1 for both the behavioral checklist and the graphic rating scale. As illustrated, the grand mean heterotrait-monomethod correlation was .393 (range=.273 to .482; SD=.103) for the BCS and .613 (range=.559 to .719; SD=.075) for the graphic rating scale dimension scores.

The results of the multitrait-multimethod approach indicate that the use of the behavioral checklist resulted in improvements in the discriminant validity of the dimension scores.

Applying the more rigorous criterion for discriminant validity, the results clearly show a smaller average within-exercise (heterotrait-monomethod) correlation for the BCS (.393), as compared to the graphic rating scale dimension scores (.613). The multitrait-multimethod results did not suggest, however, that the behavioral checklist was associated with gains in convergent validity. Specifically, the average across-exercise (monotrait-heteromethod) correlations for both the BCS and the graphic rating scale dimension scores are considerably lower than their within-exercise (heterotrait-monomethod) counterparts, suggesting poor convergent validity. Moreover, the across-exercise (monotrait-heteromethod) correlations associated with the BCS were much lower than those of the graphic rating scale dimension scores.

While the level of discriminant validity produced by the behavioral checklist is approximately the same as that reported by Reilly et al. (1990) (r=.38), the level of convergent validity does not approximate that found by Reilly et al. (r=.44), as predicted. Instead, the level of convergent validity in the present study is similar to the level reported in other assessment center research (e.g., Bycio, 1987; Robertson et al., 1987; Russell, 1987).

Analysis II: Factor Analysis

A second approach to investigating the construct validity of the BCS and graphic rating scale dimension scores followed from Sackett and Dreher (1982) and Silverman et al. (1986). Specifically, these researchers examined the underlying dimensionality of assessment center data by performing a principal-axis factor analysis using a VARIMAX rotation on the intercorrelation matrix of dimension scores across exercises.

In performing the principal axis procedure, both nine- and four-factor solutions were hypothesized as potentially meaningful based on the number of dimensions and exercises. The four-factor solution was more interpretable and is presented in Tables 2 and 3 for the behavioral checklist and graphic rating scales, respectively. In examining the factor structures for the two scales, it is evident that very distinct exercise factors are present with only a few dimensions loading on more than one factor. These factor analytic results are consistent with those found in other research (Sackett & Dreher, 1982; Silverman et al., 1986). In addition, they are consistent with the findings of the multitrait-multimethod procedure in that both suggest little evidence of the consistency of dimensional performance across exercises.

Insert Tables 2 and 3 about here

Analysis III: Internal Consistency and Interrater Reliability

The reliability of the BCS and graphic ratings scales dimension scores was assessed by the methods of internal consistency and interrater reliability. The internal consistency of the dimension scores for the two scales was computed using Cronbach's coefficient alpha. These correlation coefficients are presented in Table 4 for both the behavioral checklist and graphic rating scale. As Table 4 illustrates, the grand mean coefficient alpha is .903 (range=.834 to

---

Insert Table 4 about here

.945; SD=.036) for the BCS and is .823 (range=.783 to .859; SD=.021) for the graphic rating scale dimension scores.

Interrater reliability for the behavioral checklist and graphic rating scales was calculated by correlating the dimension scores for the two assessors. The average interrater reliability for each dimension on the behavioral checklist and graphic rating scale is presented in Table 5. As illustrated, the grand mean interrater reliability for the BCS is .976 (range=1.00 to .953; SD=.014), and the grand mean interrater reliability for the graphic rating scale dimension scores was .904 (range=.944 to .854; SD=.028).

The results of the reliability analyses suggest that while both the behavioral checklist and graphic rating scale yielded moderately high coefficient alphas and interrater reliability coefficients for the dimension scores, the behavioral checklist produced higher internal consistency and interrater agreement than did the graphic rating scale. It must be noted, however, that the higher internal consistency of the behavioral checklist, as compared to the graphic rating scale, may be related to the number of responses included on the checklist. Thus, internal consistency may not be appropriate for assessing the reliability of behavioral checklists.

In general, the results of the reliability analyses for both the behavioral checklist and graphic rating scale are consistent with earlier research reporting moderately high interrater reliability among assessment center dimension ratings (Neidig & Neidig, 1984).

Insert Table 5 about here

Analysis IV: Criterion Validity

Total scores for the assessment center produced by the behavioral checklist and graphic rating scale were computed by standardizing each dimension score and then weighting it according to its relative weight in the overall test plan. These standardized and weighted dimension scores were then summed to produce a total score for both the behavioral checklist and the graphic rating scale. The correlation between the behavioral checklist and graphic rating scale total scores was .898.

The total scores for the two scales were then correlated with candidates' service ratings for the previous two years. In completing the performance ratings, the candidates' supervisors rated the candidates on fifteen possible dimensions using a five-point scale in which "5" was *Outstanding* and "1" was *Unsatisfactory*. These dimension ratings were then averaged to compute an overall performance rating score.

As shown in Table 6, only the graphic rating scale total score had a significant relationship with candidates' 1992 (r = .163) and 1993 (r = .174) performance ratings. The relatively modest validity coefficients reported in Table 6 for both the behavioral checklist and

Insert Table 6 about here

graphic rating scale may be attributed to the lack of variance in the performance ratings. The mean overall rating for years 1992 and 1993 were 4.701 (SD=.357) and 4.72 (SD=.382), respectively. This explanation seems most plausible, given that an analysis of the reliability of

the 1992 and 1993 performance ratings yielded a coefficient alpha of .739 and .818, respectively.

As previously stated, Klimoski and Brickner (1987) have suggested that criterion contamination may be responsible for the established relationship between assessment center scores and performance ratings. This claim would be investigated by a regression analysis in which graphic rating scale total scores would be regressed on the candidates' overall performance rating scores, holding the behavioral checklist total score constant. Such an analysis could not be performed in the present study due to the relatively weak relationship between the performance ratings and the behavioral checklist and graphic rating scales, and the strong intercorrelation of the total scores for the two scales. Under these conditions little effect for either scale, holding the other constant, would be expected.

Discussion

The main goal of this study was to investigate the effects of an "untranslated" behavioral checklist on certain psychometric properties of an assessment center, specifically the construct validity of assessment center dimension ratings. It was predicted that the behavioral checklist alone, without the use of the retranslation procedure employed by Reilly et al. (1990), would reduce the cognitive complexity of the rating task and produce the same pattern of convergent and discriminant validity reported by Reilly et al. Such findings would eliminate the need for the retranslation process in the construction of behavioral checklists and would address issues related to the content validity and fairness of the evaluation process.

The results suggest that the untranslated behavioral checklists improved the discriminant validity and reliability of assessment center dimension scores over traditional graphic rating scales, but did not have a corresponding effect on the convergent validity of dimension scores, as was expected. In addition, the untranslated behavioral checklist, as compared to the graphic rating scale, did not yield a significant relationship with performance.

While the untranslated behavioral checklist failed to produce a similar level of convergent validity as that reported by Reilly et al., the level of discriminant validity obtained was approximately that of Reilly et al. (.39 vs. .41). Moreover, the level of discriminant validity in the present study is better than that reported in earlier assessment center research. Reilly et al. provided a summary of the convergent and discriminant validity findings of this research.

This summary is presented in Table 7. An inspection of Table 7 confirms that only Sackett and Dreher (1982) and Reilly et al. (1990) have obtained levels of discriminant validity below .45.

Insert Table 7 about here

The level of discriminant validity achieved with the untranslated behavioral checklist may be explained by the cognitive-reduction benefits associated with these scales. As explained earlier, even without employing the retranslation approach, the behavioral checklists identify specific behavioral responses and organize them into their relevant dimensions based on the dimension's operational definition. Research has shown that assessors employ their own reductionist strategies to contend with the cognitive complexity of the evaluation task by using only a few dimensions (Gaugler & Thornton, 1989). Eliminating the need to categorize responses may have reduced the cognitive demands imposed on the assessors, and in turn decreased the amount of convergence typically found among dimension ratings within exercises.

Two explanations are offered for the failure of the untranslated behavioral checklist to similarly affect the convergent validity of dimension ratings. First, the omission of the retranslation procedure from the behavioral checklist construction process may have resulted in poorer convergent validity. Specifically, Reilly et al. (1990) postulated that the retranslation procedure may benefit assessors by providing them with a clearer understanding of the dimension definitions, and thus enabling them to more effectively identify and categorize behaviors. Reilly et al., however, rejected this as a possible explanation for their findings,

stating that prior research has not demonstrated that the extent of assessor training moderates assessment center validity (Gaugler et al., 1987). This explanation is also rejected as a means of describing the results of the present study. Thus, eliminating the retranslation process, as proposed, would not be expected to have a negative impact on the convergent validity of dimension ratings.

A more plausible explanation for the poor convergent validity in the present study relates directly to the methodology used in Reilly et al.'s investigation. As mentioned earlier, these authors investigated the convergent validity of two group exercises both of which involved an assembly problem. Because responses to exercises are situationally determined, consistency in dimensional performance could be expected for these two exercises because of their like contexts, and thus would explain the convergence of dimension ratings across the two exercises. This explanation is supported by two lines of evidence. First, the factor analytic results of the present study, which evaluated four contextually different exercises using a behavioral checklist, clearly showed the presence of distinct exercise factors, not separate dimensions. Similar factor analytic results have been obtained in other assessment center research that includes multiple exercises (Sackett & Dreher, 1982; Silverman et al., 1987). Second, the findings of the present study revealed very different levels of convergent validity than those of Reilly et al. (.25 vs. .43), even when a behavioral checklist was also employed. Moreover, no other research investigating the convergent validity of multiple exercise reports convergent validity of this magnitude (see Table 7). The exercises used in these studies included group exercises, in-baskets, role plays, and interviews; in no instance were only two exercise similar in format and content used.

Future research should determine if the gains in convergent validity achieved by Reilly et al. were due primarily to the selection of exercises similar in situational contexts. If Reilly et al.'s results cannot be replicated using multiple exercises, then the available evidence supports what others have identified as an "exercise effect" (Neidig & Neidig, 1984) in which variance in dimensional performance across exercises is expected due the different situations in which the candidate is placed. The abundance of evidence demonstrating such an effect has led to a recommendation that attempts to measure constructs be abandoned in favor of measuring specific behavioral responses to work samples designed to simulate important work activities identified through job analysis (Neidig & Neidig, 1984; Sackett & Dreher, 1984).

In this vein, behavioral checklists would be very appropriate as a method of exercise evaluation. As the results of the present study indicate, behavioral checklists improved the internal consistency and interrater reliability of dimension scores, and thus could be expected to yield similar results for scores on overall exercises. As mentioned earlier, internal consistency may not be as appropriate as interrater reliability for the evaluation of the behavioral checklist's reliability because of its tendency to increase proportionate to the number of items assessed.

The high interrater reliability ($r = .976$) obtained in the present study may be attributed to the near objective level of the rating task when employing a behavioral checklist (Reilly et al., 1990). Not only would the objective nature of the behavioral checklist increase the amount of agreement among assessors, but it also has the added benefit of reducing the reliance on consensus discussion which can be time-intensive. Instead, assessors can independently rate

candidates using the behavioral checklist and limit their discussion to areas in which there is disagreement.

In addition to increasing the reliability of assessment center ratings, the behavioral checklists also has other advantages. Foremost among these benefits is its ability to ensure content validity of the evaluation process. Sackett (1987) points out that in developing assessment centers, evaluations of content validity are typically made on the basis of the stimulus materials alone with little attention being given to the scoring process. Developed with the assistance of subject matter experts, behavioral checklists ensure the identification of content valid behavioral responses to the exercises. In identifying and retaining all responses elicited by a particular exercise, even those with a low or zero weight, ensures the most complete (content valid) scales. As previously mentioned, ensuring the content validity of the rating process is very important from both the perspectives of face validity and fairness in evaluating candidates.

Reilly et al. (1990) have suggested that behavioral checklists offer still other benefits that are related to the evaluation process itself. Specifically, not only is the rating process less cognitively demanding, but it is also simplified by eliminating the need for raters to categorize behaviors and discuss their ratings. Likewise, assessor training can also be simplified when employing behavioral checklists by focusing on the recognition and recall of specific behaviors. Finally, the feedback process can be enhanced by the use of behavioral checklists in that assessees can be provided with more specific feedback regarding their performance on an exercise. It must be mentioned, however, that in reducing the evaluation process to an objective

level, consideration must be given to the attitudes of the assessors regarding the rating process, especially if they have been selected for their expertise.

In sum, the results of this study suggest that "untranslated" behavioral checklists failed to produce the levels of convergent validity reported by Reilly et al. (1990). However, this may be due to a prevalent exercise effect found when evaluating multiple exercises that are contextually different. However, the untranslated behavioral checklist was associated with gains in discriminant validity and reliability. In addition, its use may increase the content validity of the assessment process itself. Because of the unreliability and skewness in the distribution of the performance ratings in the present study, no conclusions could be made regarding the criterion validity of the behavioral checklist. Future research should examine the differential validity of various evaluation methods to determine their affect on criterion validity. In addition, such research would allow for an investigation of the claim that criterion contamination is responsible for the well-established relationship between assessment center ratings and performance.

Practical Implications

While this study suggests that there are many advantages associated with the use of untranslated behavioral checklists in the evaluation of assessment centers, such as improvements in certain psychometric characteristics and a clear establishment of the content-validity of the evaluation process, we caution potential users regarding certain practical limitations associated with behavior checklists, based on our experience over the last eight years.

As previously mentioned, consideration must be given to assessors' attitudes since the evaluation process is reduced to a near objective level. Assessors are often selected for their qualifications, and behavioral checklist scales do not fully utilize this expertise. To determine what impact the behavioral checklist scales had on assessors' attitudes, a 10-item scale assessing preferences for using the graphic rating scale or the behavioral checklists was developed and administered to the assessors participating in the present study. An analysis of the survey's results showed no significant differences in rater preferences for the two scale types ($t(27) = .65$). However, a prevalent theme in the assessors' comments was that they often felt hindered by not having any discretion when making ratings on the behavioral checklists.

In addition to considering the effects behavioral checklists may have on assessors' attitudes, the potential user should also be concerned with the time and cost involved in developing these scales. Employing the behavioral checklist will protract the scale development process. For example, to ensure the completeness (content validity) of responses on the behavioral checklist, it is often necessary to increase the number of SMEs involved in scale construction. It may also be necessary, as in our case, to sample candidates' performances in an effort to identify a near exhaustive list of content valid responses.

Behavioral checklists can also make the rating process itself more time- and labor-intensive. Depending on the complexity of the exercise, the number of responses on the behavioral checklist can be great. As the volume of responses increases, so too does the difficulty of the rating task. As a result, additional assessors are often needed to compensate for the time required to rate a single candidate.

In sum, careful consideration should be given to the choice of evaluation methods for assessment centers. There are many benefits to using behavioral checklists. as suggested. However, behavioral checklists may not be appropriate in all situations, given the practical concerns we have described. Specifically, behavioral checklists may not be appropriate for smaller organizations that do not have the staff and other resources required to support their development, administration, and scoring. We suggest that behavioral checklists may be most appropriate for evaluating exercises that are designed to measure very specific behaviors, thus limiting the number of responses. In addition, we recommend that behavioral checklists are used to evaluate assessment centers for positions with smaller numbers of candidates. Otherwise, there will be dimishing returns on the effort extended to obtain better quality data, due to the time and cost involved in using behavioral checklist.

References

Bycio, P., Alvares, K.M., & Hanh, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. Journal of Applied Psychology, 74, 478-494.

Byham, W.C. (1980, February). Starting an assessment center the right way. Personnel Administrator, 27-32.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 36, 81-105.

Dreher, G.F., & Sackett, P.R. (1981). Some problems with applying content validity evidence to assessment center procedures. Academy of Manangement Review, 6, 551-560.

Gaugler, B.B., & 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. III, & Bentson, C. (1987). Meta-analysis of assessment center validity [Monograph]. Journal of Applied Psychology, 72, 493-511.

Haymaker, J.C., & Grant, D.L. (1982). Development of a model for content validation of assessment centers. Journal of Assessment Center Technology, 1(1), 15-17.

Hunter J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-68.

Jaffee, C.L., & Sefick, J.T. (1980, February). What is an assessment center? Personnel Administrator, 40-43.

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

Neidig R.D., & Neidig P.J. (1984). Multiple assessment center exercises and job relatedness. Journal of Applied Psycholgy, 69, 182-186.

Reilly, R.R., Henry, S., & 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.

Robertson, I., Gratton, L., & Sharpley, (1987). Psychometric properties and the design of managerial assessment centers: Dimensions into exercises won't go. Journal of Occupational Psychology, 60, 187-195.

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

Sackett P.R., & Dreher G.F. (1984). Situational specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. Journal of Applied Psychology, 67, 401-410.

Schmidt, F.L., Ones, D.S., & Hunter, J.E. (1992). Personnel selection. Annual Review of Psychology, 43, 627-670.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Schmitt, N., & Noe, R.A. (1983). Demonstration of content validity: Assessment center example. Journal of Assessment Center Technology, 6(2), 5-11.

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

Smith, P.C. & Kendall, L.M. (1963) Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Turnage JJ, Muchinsky PM. (1982). Transituational variability in human performance with assessment centers. Organizational Behavior and Human Performance, 30, 174-200.

TABLE 1

Dimension and Exercise Correlations for the Behavioral Checklist and the Graphic Rating Scale

	Behavioral Checklist Mean *r*	SD	Graphic Rating Scale Mean *r*	SD
Dimension (Monotrait-Heteromethod Correlations)				
Interpersonal	.262	.064	.270	.079
Development of Subordinates	.295	.071	.315	.086
Leadership & Delegation	.219	.107	.314	.058
Problem Analysis & Decision Making	.288	.071	.292	.073
Organization, Coordination, & Resource Allocation	.336	.147	.317	.042
Investigation & Police Work	.282	.118	.300	.090
Oral Communication	.238	.065	.236	.044
Control & Follow-Up	.284	.057	.376	.106
Use of Police References & Quantitative Resources	.079	.098	.359	.030
Grand Mean and SD	.254	.074	.309	.042
Heterotrait-Heteromethod Correlations				
Grand Mean and SD	.236	.090	.297	.074
Exercise (Heterotrait-Monomethod Correlations)				
Apprehending Suspects Situational Video	.478	.237	.559	.115
Problem Subordinate Situational Video	.273	.111	.562	.106
Protest Situational Video	.341	.107	.615	.063
In-Basket	.482	.140	.719	.082
Grand Mean and SD	.393	.103	.613	.075

TABLE 2

Rotated Factor Pattern for the Behavioral Checklist Sums (BCS)

Dimensions	Exercises	I	II	III	IV
Interpersonal	Apprehending Suspects	.69			
Development of Subordinates	Apprehending Suspects	.42			
Leadership & Delegation	Apprehending Suspects	.79			
Problem Analysis & Decision Making	Apprehending Suspects	.89			
Organization, Coordination, & Resource Allocation	Apprehending Suspects	.86			
Investigation & Police Work	Apprehending Suspects	82			
Oral Communication	Apprehending Suspects	.61			
Control & Follow-Up	Apprehending Suspects	.73			
Use of Police References & Quantitative Resources	Apprehending Suspects			.48	
Interpersonal	Problem Subordinate	.39		.36	
Development of Subordinates	Problem Subordinate			.48	
Leadership & Delegation	Problem Subordinate			.51	
Problem Analysis & Decision Making	Problem Subordinate			.55	
Investigation & Police Work	Problem Subordinate			.47	
Oral Communication	Problem Subordinate				
Control & Follow-Up	Problem Subordinate			.52	
Use of Police References & Quantitative Resources	Problem Subordinate			.45	
Interpersonal	Protest				.59
Development of Subordinates	Protest			.51	
Leadership & Delegation	Protest				.51
Problem Analysis & Decision Making	Protest				.74
Organization, Coordination, & Resource Allocation	Protest		.37	.43	.40
Investigation & Police Work	Protest				.76
Oral Communication	Protest				.60
Control & Follow-Up	Protest				.61
Interpersonal	In-Basket		.55		
Development of Subordinates	In-Basket		.86		
Leadership & Delegation	In-Basket		.72		
Problem Analysis & Decision Making	In-Basket		.79		
Organization, Coordination, & Resource Allocation	In-Basket		.73		
Investigation & Police Work	In-Basket		.66		
Oral Communication	In-Basket		.46		
Control & Follow-Up	In-Basket		.78		
Use of Police References & Quantitative Resources	In-Basket		.70		

Note. Only factor loadings greater than or equal to .35 are presented.

TABLE 3

Rotated Factor Pattern for the Graphic Rating Scale Dimension Scores

Dimensions	Exercises	I	II	III	IV
Interpersonal	Apprehending Suspects		.36		.46
Development of Subordinates	Apprehending Suspects				.56
Leadership & Delegation	Apprehending Suspects				.79
Problem Analysis & Decision Making	Apprehending Suspects				.81
Organization, Coordination, & Resource Allocation	Apprehending Suspects				.80
Investigation & Police Work	Apprehending Suspects				.76
Oral Communication	Apprehending Suspects				.75
Control & Follow-Up	Apprehending Suspects			.35	.72
Use of Police References & Quantitative Resources	Apprehending Suspects				
Interpersonal	Problem Subordinate			.56	
Development of Subordinates	Problem Subordinate			.75	
Leadership & Delegation	Problem Subordinate			.81	
Problem Analysis & Decision Making	Problem Subordinate			.80	
Investigation & Police Work	Problem Subordinate			.63	
Oral Communication	Problem Subordinate			.72	
Control & Follow-Up	Problem Subordinate			.74	
Use of Police References & Quantitative Resources	Problem Subordinate			.75	
Interpersonal	Protest		.71		
Development of Subordinates	Protest		.75		
Leadership & Delegation	Protest		.82		
Problem Analysis & Decision Making	Protest		.80		
Organization, Coordination, & Resource Allocation	Protest		.76		
Investigation & Police Work	Protest		.87		
Oral Communication	Protest		.75		
Control & Follow-Up	Protest		.80		
Interpersonal	In-Basket	.69			
Development of Subordinates	In-Basket	.86			
Leadership & Delegation	In-Basket	.89			
Problem Analysis & Decision Making	In-Basket	.87			
Organization, Coordination, & Resource Allocation	In-Basket	.86			
Investigation & Police Work	In-Basket	.82			
Oral Communication	In-Basket	.77			
Control & Follow-Up	In-Basket	.85			
Use of Police References & Quantitative Resources	In-Basket	.87			

Note. Only factor loadings greater than or equal to .35 are presented.

TABLE 4

Coefficient Alpha by Dimension and Total Score for the Behavioral Checklist and Graphic Rating Scale

	Behavioral Checklist Coefficient Alpha	Graphic Rating Scale Coefficient Alpha
Dimension		
Interpersonal	.883	.818
Development of Subordinates	.938	.839
Leadership & Delegation	.900	.839
Problem Analysis & Decision Making	.927	.820
Organization, Coordination, & Resource Allocation	.909	.800
Investigation & Police Work	.945	.828
Oral Communication	.858	.783
Control & Follow-Up	.933	.859
Use of Police References & Quantitative Resources	.834	.823
Grand Mean	.903	.823

TABLE 5

Interrater Reliability by Dimension for the Behavioral Checklist and Graphic Rating Scale

	Behavioral Checklist		Graphic Rating Scale	
	Mean *r*	SD	Mean *r*	SD
Dimension				
Interpersonal	1.000	.000	.943	.048
Development of Subordinates	.977	.026	.934	.074
Leadership & Delegation	.978	.045	.930	.089
Problem Analysis & Decision Making	.962	.012	.902	.064
Organization, Coordination, & Resource Allocation	.988	.007	.893	.094
Investigation & Police Work	.964	.030	.897	.067
Oral Communication	.953	.025	.873	.048
Control & Follow-Up	.973	.061	.913	.083
Use of Police References & Quantitative Resources	.992	.021	.853	.108
Grand Mean	.976	.014	.904	.028

TABLE 6

Correlations Between Performance Ratings and Behavioral Checklist and Graphic Rating Scale Total Scores

	Behavioral Checklist	Graphic Rating Scale
1992 Performance Ratings	.1425	.1625*
1993 Performance Ratings	.1376	.1757*

Note: *Significant at $p < .05$ level. N = 173 for 1992 Performance Ratings and N = 149 for the 1993 Performance Ratings.

TABLE 7

Convergent and Discriminant Validity Results of Assessment Center Research Summarized by Reilly et al. (1990)

Source	Average Convergent Validity	Average Discriminant Validity
Reilly et al. (1990)	.43	.41
Sackett & Dreher (1982)		
Company A	.07	.64
Company B	.11	.40
Company C	.51	.65
Turnage & Muchinsky (1982)		
Sample A	.45	.53
Sample B	.44	.52
Silverman et al. (1986)		
Sample A	.54	.65
Sample B	.37	.68
Russell (1987)	.25	.53
Bycio et al. (1987)	.36	.75
Robertson et al. (1987)		
Organization 1	.28	.64
Organization 2	.26	.66
Organization 3	.23	.60
Organization 4	.11	.49

Appendix
Definitions of Assessment Center Dimensions

1. *Interpersonal* - the ability to use human relations skills in interacting with subordinates, superiors, citizens, and other personnel within the department and outside agencies
2. *Development of Subordinates* - the ability to develop subordinates by establishing guidelines, observing behavior, and providing feedback, counseling, or disciplinary actions
3. *Leadership and Delegation* - the ability to direct activities of subordinates in order to achieve departmental goals
4. *Problem Analysis & Decision Making* - the ability to identify potential and existing problems and to make high quality, timely decisions
5. *Organization & Coordination* - the ability to organize and coordinate resources on scene and administratively
6. *Investigation & Police Work* - the ability to ask questions that obtain information to further an investigation and to perceive critical information and determine when to use different techniques
7. *Oral Communication* - the ability to communicate ideas, orders, and assignments orally to a wide variety of people
8. *Control and Follow-Up* - the ability to follow up on goals, assignments, unsolved and ongoing problems and projects
9. *Use of Police References and Quantitative Resources* - the ability to use police resources as guides in decision making and application