Number of Assessment Center Dimensions as a Determinant of Assessor Accuracy

Barbara B. Gaugler
Rice University

George C. Thornton III
Colorado State University

Undergraduates (N = 131) were trained as assessors, who evaluated the performance of confederates in an assessment center simulation on 3, 6, or 9 dimensions. Number of dimensions significantly affected some assessment center judgments but not others. Ss who rated a small number of dimensions classified behaviors more accurately and made more accurate ratings than did Ss who rated a large number of dimensions. Number of dimensions did not affect the accuracy of assessors' observations nor the discriminant validity of their dimension ratings. Given these results and the findings of others (Hinrichs & Haanpera, 1976; Russell, 1985; Sackett & Hakeš, 1979; Schmitt, 1977), developers of assessment centers should limit the cognitive demands placed on assessors by, for example, minimizing the number of dimensions assessors are required to process.

The job of an assessor in an assessment center is exceedingly complex. Assessors must observe and record the performance of several candidates in job simulations and situational exercises, classify the recorded behaviors into dimensions of performance, and rate each candidate on each dimension. In the integration discussion, assessors evaluate each candidate, taking into account performance across all exercises, and make recommendations regarding each candidate's potential or developmental needs. In addition, assessors may be asked to consider (a) the importance of each dimension, (b) their amount of confidence regarding each dimension rating, (c) the interrelationships among dimensions, (d) the amount and type of training an assessor needs to overcome weaknesses with respect to particular dimensions, and (e) the probability that change can occur with regard to these dimensions (Byham, 1977).

Human beings have a limited capacity to process information at least in part because of limitations in the number of items they can retain in working memory (Lachman, Lachman, & Butterfield, 1979). A common means of compensating for this limitation is through recoding input information by means of reductive coding operations (Miller, 1956). For example, one tendency is to decrease informational complexity by seeing things, characteristics, and events as related, thereby recoding complex inputs into a few categories (Campbell, 1958). Reductionistic strategies are used more frequently when the person is faced with complex tasks (Payne, 1976, 1982). Given the complex nature of the assessor's job, we would expect assessors to simplify their task in this manner.

Much of the complexity in an assessor's job stems from evaluating a large number of dimensions. In reaching an overall evaluation, assessors may be asked to simultaneously consider ratings on as many as 25 dimensions (Sackett & Hakeš, 1979). There are several alleged reasons for this, the main one being to ensure adequate content coverage of the diverse requirements of the manager's job. Paradoxically, dimensions are also used to help assessors reduce the complexity of their information-processing tasks. The mechanisms by which this is presumed to happen include reduction in the storage and processing load, facilitation of grouping and communicating about similarities and differences in behaviors, and encouragement of assessors to draw inferences about imperceptible attributes from observed behaviors (Zedeck, 1986).

Although assessors are instructed to consider all information and all dimensions when evaluating assesses, several lines of research suggest that only a few dimensions are typically used (Hinrichs & Haanpera, 1976; Russell, 1985; Sackett & Hakeš, 1979; Schmitt, 1977). Regression of the overall assessment rating (OAR) on dimension ratings, factor analytic studies of the final dimension ratings, and construct validity studies of within-exercise dimension ratings all converge in the conclusion that assessors can use only three to six dimensions in structuring their evaluations. In addition, factor analytic research suggests that assessors' initial dimension ratings are dominated by a single general factor (Russell, 1985; Sackett & Hakeš, 1979), which is probably best described as halo. This halo may be an overall evaluation of whether a particular manager is good or bad and may be derived from assessors' schemata of what is a good manager.

In general, therefore, the evidence supports the reductive coding hypothesis more clearly than it does the simplification-through-dimensions view. Of course, the positions are not in direct conflict: One could get some benefit from dimensionality so long as the number of dimensions did not exceed the processor's limited capacity.

Requiring assessors to rate a large number of dimensions has
the additional disadvantage that it may adversely affect the discriminant validity of their ratings. Much recent attention has been directed toward understanding the construct validity of within-exercise dimension ratings (Sackett & Dreher, 1982; Silverman, Dalessio, Woods, & Johnson, 1986; Turnage & Muchinsky, 1982). Within-exercise dimension ratings have consistently been found to have low discriminant validity, leading some to conclude that assessors are evaluating performance in exercises rather than performance on dimensions (Sackett & Dreher, 1984).

When asked to rate a large number of nonorthogonal dimensions, assessors may have difficulty deciding which behaviors are indicative of which dimensions. Thus, a general stereotype of the assesse's performance in an exercise may influence individual dimension ratings. Performance in a particular exercise may become more salient than performance on a large number of overlapping dimensions. Therefore, one reason why assessment center ratings may represent performance in exercises rather than performance on dimensions is that assessors are asked to use a large number of dimensions that are probably not orthogonal in the real world.

In spite of the overwhelming evidence that assessors do not make discriminations among more than a few dimensions (Hinrichs & Haanpera, 1976; Russell, 1985; Sackett & Hakel, 1979; Schmitt, 1977), the number of dimensions assessors can effectively use has not been directly assessed. The purpose of our study was to examine the influence of the number of dimensions that assessors rate on the observation, classification, and evaluation of behaviors. Although the number of dimensions that assessors process can potentially influence the outcome of each of the critical phases in an assessment center (i.e., observation, classification, and evaluation), prior research has focused exclusively on the evaluation of assesse behavior. The accuracy with which assessors observe and classify behavior was evaluated in this study to determine at what point in the assessment center process the number of dimensions to be processed matters, if it matters at all.

In this investigation, subjects were given three, six, or nine dimensions to rate. We predicted that dimension ratings made by subjects who rate a small number of dimensions would be more accurate and have greater discriminant validity than dimension ratings made by subjects who rate a large number of dimensions. We also hypothesized that subjects who process a few dimensions will observe and classify behavior more accurately than subjects who process a large number of dimensions.

Method

Overview

Introductory psychology students served as assessors in an assessment center simulation to evaluate candidates for a residence hall student assistant position. After being trained as assessors, subjects observed the videotaped performances of confederates in various situational exercises. Assessors classified behaviors they recorded into three, six, or nine dimensions. Subjects then met in assessor teams, made reports, and made within- and across-exercise dimension ratings for each candidate.

Subjects

One hundred thirty-one students (57 men and 74 women) participated as assessors to fulfill the course research requirement. Their mean age was 18.3 years; ages ranged from 17 to 23 years. Eighty-two percent of the subjects were freshmen. 15% were sophomores, and 3% were juniors or seniors. The mean number of semesters subjects lived in residence halls was one semester: the range was from 1 to 7 semesters. The average, subjects had indirect contact with five student assistants; the number of student assistants known ranged from 1 to 24. Subjects were required to have lived in a residence hall for at least one semester to be eligible to participate in the study. They were assigned randomly to one of the three groups defined in terms of dimensionality (three, six, or nine), resulting in 40 to 46 subjects per group.

The use of college students as subjects in applied research has been heavily debated (Gordon, Slade, & Schmitt, 1986). College students are rarely used in assessment center research, yet students are frequently used as judges in interviewing research. Although the findings are mixed, students and professional interviewers have been found to make similar evaluations (Bernstein, Hake & Harlan, 1975; Dipboye, Fromkin, & Wiback, 1975). One criticism of using students as subjects is that they are often unfamiliar with the experimental task required of them. In this investigation, assessors had experience with incumbents in the target job. Furthermore, we are investigating basic psychological processes that are likely common to all people, even college students.

Procedure

Assessors participated in groups of nine. Each group consisted of three three-person teams. Subjects were trained in observing and classifying behaviors according to dimensions and in making dimension ratings. They then watched videotapes of the three confederates acting the role of student assistants in the three exercises. Each assessor observed a different assesse in the three exercises. Each assesse's performance in a particular exercise was observed by only one assessor from each team.

After watching each tape and recording observations, subjects classified behaviors into dimensions. Then they met in their respective assessor teams to review candidates one at a time. Team members shared their observations, exercise by exercise. For each candidate, assessors rated performance on each dimension independently, considering the candidate's performance on all exercises. Assessors did not make consensus, across-exercise dimension and overall assessment ratings for the candidates.

More than the required number of assessors were scheduled for each session to ensure that the assessor teams would be complete. All subjects recorded observations of candidates and classified behaviors into dimensions. Then excess subjects were randomly assigned to be observers during the integration session. Observers recorded observations and rated the presentation skills of the assessors. Observers did not share their observations of the candidates nor did they rate the candidates' performance. In summary, the total sample of subjects provided data on observation and classification, whereas only 111 subjects participated in the integration session and provided within- and across-exercise dimension ratings.

Assessment Center Dimensions

To manipulate the number of dimensions used in the respective groups, a pool of dimensions was required from which sets of three, six, and nine could be drawn. These numbers of dimensions were chosen on the basis of previous research that suggested three dimensions as sufficient to capture most of the variance in assessors' judgments (Sackett & Hakel, 1979), and nine as the upper limit of the number of dimensions used in prior discriminant validity studies.
Construction of a dimension pool was accomplished by using a job-analysis procedure that included importance ratings. Importance, of course, is a variable that must be controlled in order not to be confounded with set size. Supervisors of student assistants were interviewed concerning the important behaviors required to be a successful student assistant. These behaviors were used to construct 27 preliminary dimensions. Dimension definitions were then discussed with the supervisors, who rated the importance of each of the dimensions on a scale from 1 (not important at all) to 10 (extremely important). The 15 dimensions with the highest mean ratings were retained and then evaluated by 10 undergraduate psychology students who rated their importance on the scale just described. The mean importance ratings were 7.2, 7.1, and 7.6 for the small, medium, and large sets, respectively.

In addition to average importance, the average similarity among dimensions in a set must be controlled to avoid serious confounding with the set size variable. The nature of the dimensions typically assessed in assessment centers suggests that as the number of dimensions increases, their degree of overlap increases. Thus, to control similarity, it was necessary to equate overlap as much as possible. To this end, previous research, student ratings, and expert evaluation were used in the following manner:

Prior research on dimension ratings. In general, three factors have been identified in studies concerned with dimension overlap (Thornton & Byham, 1982); (a) administrative skills (e.g., planning and organizing); (b) interpersonal skills (e.g., oral communication and leadership); and (c) the amount of activity demonstrated by the assessor (e.g., aggressiveness). In our investigation, an equal number of dimensions was selected from each of these three factors for each of the three sets of dimensions.

Student perceptions of similarity among dimensions. A pair-comparison questionnaire was used to assess students' perceptions of the similarity among dimensions. Seventeen students were asked to rate the similarity of all possible combinations of dimension pairs on a 1 (not similar at all) to 10 (virtually identical) scale. The mean similarity among various combinations of three, six, and nine dimensions was computed. The final sets of dimensions selected had roughly the same mean similarity rating, as follows: 5.7, 5.7, and 5.5, for the three-, six-, and nine-dimensional groups, respectively.

Opinions of a panel of experts. A panel of three experts evaluated the three groups of dimensions and the minimal standards of performance (described later) to ensure that they were equally distinct. The experts were graduate students who had participated in an advanced graduate seminar on assessment centers and who were experienced both in assessment and in the design of exercises for assessment centers.

As a result of these measures, the amount of overlap among dimensions within a set was similar for each set of dimensions. The dimensions representing the interpersonal factor were oral communication, tolerance for personal stress, and sensitivity; those for the administrative factor were planning and organizing, leadership, and time management; and those for the activity level factor were self-confidence, initiative, and assertiveness.

Assessor Training

Assessor training focused on (a) helping assessors develop a common understanding of the dimensions being observed and evaluated, (b) teaching assessors what behaviors to observe in each exercise and how to observe them accurately and completely (this included practice and personalized feedback), (c) providing assessors with a common standard for rating behavior in and across exercises, and (d) teaching them how to participate in the integration session.

Subjects were trained for 1.5 hr. Although many researchers conduct more extensive training, the results of a meta-analysis on assessment center validity (Gaugler, Rosenthal, Thornton, & Benton, 1987) found that amount of training over a range of .5 to 15 days does not moderate assessment center validity. There is no evidence to suggest that a specific minimum amount of time is necessary to adequately train assessors in the basic observational and judgment processes.

Minimal Standards of Performance

To provide assessors with a common standard for rating behavior in exercises, we developed minimal standards of performance for each dimension observable in each exercise. For each dimension, the minimal standard consisted of all the behaviors a candidate must exhibit, or must not exhibit, to be rated as acceptable. An attempt was made to specify only those behaviors that were representative of one, and only one, dimension. However, the expert raters found several behaviors to be indicative of more than one dimension.

Exercises

We developed an interview simulation, leaderless group discussion, and conflict mediation exercise because these were relevant to the student assistant position and because most of the dimensions in this study could be observed in the three exercises (Thornton & Byham, 1982). All dimensions were observable in the three exercises, except that leadership and time management could not be evaluated in the group discussion because the confederates could not display enough relevant behaviors.

To allow for variability in dimension ratings among candidates, one poor, one moderate, and one very good candidate were portrayed by the role players. The poor assese made poorly on most dimensions, the moderate assese showed variability in performance, and the very good assese performed above average on most dimensions.

Results

Observation Accuracy

Two trained research assistants matched subjects' recordings with the behaviors identified by the three experts. Each subject's data were scored by one of the research assistants. The inter-rater agreement in scoring observation and classification accuracy was computed prior to scoring subjects' data. There was total agreement among the judges on 96% of the observations in the interrater agreement study.

The accuracy of subjects' recorded observations was measured two ways (McIntyre & Benton, 1984), both involving the concept of a good observation. A good observation is a behavioral statement that specifically describes what a person says or does. The total number of good observations displayed on the videotape relevant to each dimension was identified by the expert judges who reviewed the videotapes as often as desired. The first accuracy measure was the ratio of the number of good observations to the total number of good observations displayed on the videotapes (Good/True). The second was the ratio of the number of good observations to the total number of observations recorded by a subject (Good/Total).

The methods of assessing observation accuracy were used to obviate the problem inherent in free recall as a measure of observation, namely, that sensitivity of observation is confounded with the response mode of recording observations (K. Murphy, personal communication, March 12, 1986). In a free recall for-
Table 1
Means and Standard Deviations of Accuracy Measures

<table>
<thead>
<tr>
<th>Dependent measure</th>
<th>Three</th>
<th>Six</th>
<th>Nine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation accuracy a, b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good/True</td>
<td>.527</td>
<td>.586</td>
<td>.550</td>
</tr>
<tr>
<td>SD</td>
<td>.118</td>
<td>.125</td>
<td>.150</td>
</tr>
<tr>
<td>Good/Total</td>
<td>.698</td>
<td>.803</td>
<td>.771</td>
</tr>
<tr>
<td>SD</td>
<td>.190</td>
<td>.116</td>
<td>.248</td>
</tr>
<tr>
<td>Classification accuracy a, b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.339</td>
<td>.211</td>
<td>.230</td>
</tr>
<tr>
<td>SD</td>
<td>.130</td>
<td>.080</td>
<td>.090</td>
</tr>
<tr>
<td>Rating accuracy c, d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elevation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.215</td>
<td>.334</td>
<td>.354</td>
</tr>
<tr>
<td>SD</td>
<td>.169</td>
<td>.205</td>
<td>.216</td>
</tr>
<tr>
<td>Differential elevation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.349</td>
<td>.308</td>
<td>.347</td>
</tr>
<tr>
<td>SD</td>
<td>.219</td>
<td>.205</td>
<td>.195</td>
</tr>
<tr>
<td>Stereotype accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.204</td>
<td>.300</td>
<td>.358</td>
</tr>
<tr>
<td>SD</td>
<td>.103</td>
<td>.079</td>
<td>.095</td>
</tr>
<tr>
<td>Differential accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.360</td>
<td>.471</td>
<td>.516</td>
</tr>
<tr>
<td>SD</td>
<td>.130</td>
<td>.112</td>
<td>.086</td>
</tr>
</tbody>
</table>

a Higher numbers indicate greater observation and classification accuracy. b The number of subjects in the three-, six-, and nine-dimension groups were 46, 45, and 40, respectively. c Lower numbers indicate higher rating accuracy. d Subjects who did not participate in the integration session (i.e., observers; N = 20) did not make across-exercise dimension ratings and therefore were excluded from the rating accuracy analyses. For the three-, six-, and nine-dimension groups, Ns = 37, 40, and 34, respectively.

The results of an analysis of variance (ANOVA) on classification accuracy indicated that the number of dimensions significantly affected classification accuracy, F(2, 128) = 20.25, p < .001, ω² = .23. Scheffé tests of significant differences between group means indicated that assessors who classified behaviors into three dimensions (M = .339) did so more accurately than did those required to use six (M = .211) or nine (M = .230) dimensions. In addition, subjects in the three-dimension condition recorded fewer good observations than subjects in the other groups, although this finding was not statistically significant, F(2, 128) = 2.57, p > .05. The average number of good observations made by subjects who rated three, six, or nine dimensions were 34.91, 39.00, and 36.45, respectively. Therefore, we are confident that subjects in the three-dimension group were more accurate than subjects in the other groups, not just better recorders of good observations.

Discriminant Validity

The nature of the assesses evaluated in this study precluded a classic analysis of discriminant validity. They differed from assesses in other studies in one fundamental way: The behavior of individual assessors in field studies was probably more heterogeneous across dimensions than the behavior of the confederates in our investigation. To ensure variability among candidates, we developed good, moderate, and poor candidates. However, a systematic correlation among dimensions was built into the true differences in performance (e.g., the good candidate was high on all dimensions and performed well in all exercises).

If the data were analyzed in the traditional way, any true variation caused by the independent variables would be masked by the consistency in performance of each candidate on the different dimensions, within and across exercises. To overcome this problem, the raw data matrix of within-exercise dimension ratings was “double-centered” (Gorsuch, 1983). For each rating made by each assessor, the following two means were subtracted: (a) the average rating made by that assessor across all exercises, dimensions, and candidates; and (b) the mean rating of the candidate in question, across all exercises, dimensions, and raters. By this process of double-centering the matrix, the leniency of individual raters and effects of the consistency in performance of individual candidates were reduced. In other words, deviation scores were created by removing assessor and candidate mean differences so that the true effects of the independent variables on the discriminant validity of assessors’ ratings could be examined. These deviation scores were used in both sets of discriminant validity analyses described later.

Correlations among all dimensions in all exercises were computed to form a multitrait-multimethod matrix. 1 Ratings on time management and leadership were excluded from the multi-

---

1 Correlations among all ratings are available from Barbara B. Gaugler.
Table 2
Mean Dimension and Exercise Correlations for Assessors Who Rated Three, Six, or Nine Dimensions

<table>
<thead>
<tr>
<th>No. dimensions</th>
<th>Correlation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Three (N = 37)</td>
<td>Six (N = 40)</td>
<td>Nine (N = 34)</td>
<td></td>
</tr>
<tr>
<td>Dimension (Monotrait–Heterotrait)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral communication</td>
<td>.83</td>
<td>.77</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>Self-confidence</td>
<td>.81</td>
<td>.80</td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>Planning &amp; organizing</td>
<td>.86</td>
<td>.81</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>Tolerance for personal stress</td>
<td>.84</td>
<td>.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initiative</td>
<td>.80</td>
<td>.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Assertiveness</td>
<td></td>
<td></td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td>Grand mean (convergent validity)</td>
<td>.83</td>
<td>.80</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>Heterotrait–Heteromethod</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand mean</td>
<td>.83</td>
<td>.77</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Exercise (Heterotrait–Monotrait)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group discussion</td>
<td>.88</td>
<td>.83</td>
<td>.80</td>
<td></td>
</tr>
<tr>
<td>Counseling session</td>
<td>.87</td>
<td>.85</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Roommate conflict</td>
<td>.92</td>
<td>.86</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Grand mean</td>
<td>.89</td>
<td>.85</td>
<td>.83</td>
<td></td>
</tr>
</tbody>
</table>

Note. Leadership and time management were excluded from this analysis because these dimensions were not rated in the group discussion.

The number of dimensions rated does not appear to have significantly influenced the discriminant validity of assessors' ratings. However, note that both the heterotrait–monotrait and heterotrait–heteromethod correlations are somewhat higher for assessors who rate a small number of dimensions than for those assessors who rate a large number of dimensions (heterotrait–monotrait: .89 vs. .83; heterotrait–heteromethod: .83 vs. .69). These findings suggest that assessors who rate a larger number of dimensions differentiate more among dimensions, both within and across methods, than assessors who rate a smaller number of dimensions.

Examination of the dimension and exercise correlations in Table 2 suggests that a large general factor is present in the ratings, as indicated by the uniformly high correlations. Other researchers (e.g., Sackett & Hakel, 1979) have also found that assessors' initial dimension ratings are dominated by a single factor and may best be described as halo.

ANOVA. The discriminant validity of subjects' within-exercise dimension ratings was calculated with an analysis of variance procedure (Kavanagh, MacKinney, & Wolins, 1971). The following four sources of variance were analyzed: (a) Assessee variance indicates the overall amount of agreement (convergent validity) on assesses over exercises and dimensions; (b) Assessee X Dimension interaction variance indicates the discriminant validity of an assessor's dimension ratings; (c) Assessee X Exercise interaction variance indicates the amount of situation (exercise) specificity of dimension ratings; and (d) error variance.

The appropriate values for comparisons across dimension groups are the intraclass correlation coefficients (ICCs) presented in Table 3. Comparisons of the ICC values among assessors who rated different numbers of dimensions suggest conclusions similar to those already discussed. Regardless of the number of dimensions rated, assessors' ratings showed high convergent validity. However, the assesses' sources of variation suggest somewhat higher convergent validity for assessors who rated fewer dimensions. Specifically, ratings made by assessors who rated nine dimensions had less convergent validity than ratings made by assessors who rated six (r = .80) or three (r = .83) dimensions of performance.

These results suggest that assessors' ratings had poor discriminant validity regardless of the number of dimensions rated. In general, neither of the criteria for discriminant validity suggested by Campbell and Fiske (1959) received much support. In our study, monotrait–heteromethod correlations were equal to or only slightly larger than correlations among different dimensions measured by different exercises (heterotrait–heteromethod correlations).

The second, more rigorous criterion for discriminant validity is that the monotrait–heteromethod correlations must be greater than the correlations among different dimensions measured by the same exercise (heterotrait–monotrait correlations). As indicated in Table 2, grand mean heterotrait–monotrait correlations were greater than their monotrait–heteromethod counterparts.

Table 3
Intraclass Correlations for Assessors Who Rated Three, Six, or Nine Dimensions

<table>
<thead>
<tr>
<th>Source</th>
<th>No. dimensions</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Three (N = 37)</td>
<td>Six (N = 40)</td>
<td>Nine (N = 34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessee (A) (convergent validity)</td>
<td>.83</td>
<td>.78</td>
<td>.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A X Dimension (discriminant validity)</td>
<td>.00</td>
<td>.03</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A X Exercise (exercise effect)</td>
<td>.06</td>
<td>.07</td>
<td>.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of dimensions (r = .06 and .07). The number of dimensions rated did not seem to affect the discriminant validity of assessors' ratings, as shown by the low values for the Assessee \times Dimension effect.

In summary, the results of both analyses suggest that the number of dimensions rated appeared to have influenced the convergent validity and exercise variance of assessors' ratings. Assessors who rated a small number of dimensions made ratings with greater convergent validity but less of an exercise effect than assessors who rated a large number of dimensions. Number of dimensions rated did not seem to significantly affect the uniformly low discriminant validity of assessors' ratings.

**Rating Accuracy**

The accuracy of each subject's dimension ratings was determined by comparing subjects' across-exercise dimension ratings with comparable ratings made by the experts. The three experts recorded and classified observations independently and made within- and across-exercise dimension ratings. They then arrived at consensus ratings for each across-exercise dimension rating for each candidate. These ratings were used as the true scores with which subjects' ratings were compared, as is often done in performance appraisal research (e.g., Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

Four rating accuracy measures were used: elevation, differential elevation, stereotype accuracy, and differential accuracy (Cronbach, 1955). **Elevation** is the average rating of an assessor over all asseessees and dimensions (overall accuracy); **differential elevation** is the average rating an assessor gives to each assessees across all dimensions (person accuracy). **Stereotype accuracy** is the mean rating an assessor gives to each dimension, across all asseessees (dimension accuracy); **differential accuracy** represents the assessor's sensitivity to assessees differences in patterns of performance (Person \times Dimension accuracy).

A MANOVA applied to the four rating accuracy indices showed that the number of dimensions significantly affected rating accuracy, $F(6, 214) = 10.56, p < .001, \omega^2 = .10$. Univariate $F$ tests revealed that the number of dimensions significantly affected three of the four measures of rating accuracy, as follows: elevation, $F(2, 108) = 4.57, p < .05, \omega^2 = .06$; stereotype accuracy, $F(2, 108) = 25.39, p < .001, \omega^2 = .31$; and differential accuracy, $F(2, 108) = 18.57, p < .001, \omega^2 = .24$. Scheffé tests demonstrated that for each of the three measures, assessors who rated a small number of dimensions made more accurate ratings than assessors who rated a large number of dimensions.

**Discussion**

The number of dimensions that subjects were asked to process significantly affected some assessment center judgments but not others. Assessors who dealt with fewer dimensions made more accurate behavior classifications and more accurate ratings with greater convergent validity and less method bias than assessors who dealt with a larger number of dimensions. However, the number of dimensions evaluated did not affect the discriminant validity of dimension ratings.

**Observation Accuracy**

Contrary to prediction, number of dimensions did not significantly influence observation accuracy. Subjects probably did not take into consideration the dimensions they would subsequently rate while making observations. Making and recording behavioral observations is a very difficult task: Information is presented to the observer at a very fast rate. Assessors must observe the behavior, decide whether or not it is worth recording, and then record the behavior observed. Probably only experienced assessors consciously take specific dimensions into account when recording observations. Thus, it is not surprising that the number of dimensions rated did not affect the accuracy with which inexperienced assessors made observations.

**Classification Accuracy**

Assessors who dealt with a small number of dimensions classified behaviors more accurately than assessors who processed a large number of dimensions. These results suggest that the assessors could better meet the cognitive demands put on them when dealing with only a small number of dimensions.

**Discriminant Validity**

Absence of any appreciable discriminant validity supports previous research on the construct validity of assessors' ratings, which showed that within-exercise dimension ratings have adequate convergent validity but inadequate discriminant validity (Sackett & Dreher, 1984; Silverman et al., 1986; Turnage & Muchinsky, 1982). However, the systematic correlation among dimensions built into the videotapes probably hindered assessors' ability to discriminate among dimensions. Collectively, these findings provide little support for the contention that dimension ratings can be interpreted as representing complex constructs.

**Rating Accuracy**

Assessors who rated a small number of dimensions made ratings with greater accuracy than assessors who rated a large number of dimensions. On the one hand, dimensions may help assessors reduce the amount of information that must be stored and processed and may facilitate the recall and interpretation of information (Zedeck, 1986). However, asking assessors to deal with a large number of dimensions may negate these effects. The present results support the contentions that (a) assessors have a limited capacity to process information (Lachman et al., 1979) and (b) differences in the number of categories available to raters influence their judgments (Landy & Farr, 1983). These results parallel those in the process-tracing literature, in which task complexity (e.g., the number of alternatives available to a decision maker) usually results in the decision maker's use of strategies such as elimination-by-aspects because they reduce information-processing demands (Payne, 1976, 1982). When performing complex tasks, people use different heuristics that keep the information-processing demands of the situation within the bounds of their limited capacity (Newell & Simon, 1972; Payne, 1980). Our findings also support the
research on decision making and performance appraisal, which has found that the greater the complexity of the judgment task, the more prone the method will be to cognitive biases (Einhorn & Hogarth, 1981; Payne, 1982). More generally, the results of this study support factor analytic research that demonstrates that assessors only use a few dimensions when making overall assessments of candidates (Hinrichs & Haanpera, 1976; Russell, 1985; Sackett & Hakel, 1979; Schmitt, 1977).

Future Research

This study evaluated the accuracy and discriminant validity of prediscussion dimension ratings. Because it is the postdiscussion dimension ratings and overall assessment ratings that are used to make employment decisions, future research should look at the effects of number of dimensions on these ratings. The consensus discussion process may negate the negative effects of asking assessors to deal with a large number of dimensions.

Other aspects of task complexity, such as the specificity and observability of dimensions, the number of exercises used, and the number of different types of decisions and recommendations assessors are asked to make, should also be evaluated. For example, assessment centers conducted for developmental purposes typically place a larger number of cognitive demands on assessors than do assessment centers conducted for hiring purposes.

Finally, although number of dimensions rated did not affect the discriminant validity of dimension ratings, discriminant validity would be more meaningfully assessed by using across-exercise dimension ratings. Whereas prior research on the discriminant validity of dimensions has focused on within-exercise dimension ratings, it is the across-exercise dimension ratings that are used to make personnel decisions. Across-exercise dimension ratings from an assessment center could be compared with ratings on the same dimensions obtained from selection interviews, pencil and paper tests, peer evaluations of asseesee's on-the-job performance, or self-evaluations.

Conclusions

This investigation has important implications for theory, research, and practice. The findings support the contention that assessors have a limited capacity to process information and that the greater the complexity of the judgment task, the more prone it will be to cognitive biases. We concur with Bycio, Alvaeres, and Hahn (1987) that developers of assessment centers should limit the cognitive demands placed on assessors, for example, by minimizing the number of dimensions assessors are required to process, especially in a selection or promotion situation in which only a yes/no decision is required. Not only would it reduce the information-processing demands on assessors and yield more accurate dimension ratings, but it would save the organization time and money. The influence of the number of dimensions assessors are asked to rate and of other aspects of task complexity on postdiscussion ratings and judgments should be investigated.

References


