Are Rating Dimensions in Assessment Centers Transparent for Participants? Consequences for Criterion and Construct Validity

Martin Kleinmann

The author investigates whether the extent to which participants recognize rating dimensions in assessment centers has an effect on performance. Results showed that people who more accurately identify dimensions perform better. Convergent validity of dimension ratings is greater when participants accurately perceive that the same dimension is being evaluated in 2 exercises. In the discussion, the author considers how the ability to adequately perceive rating dimensions (i.e., transparency) may influence criterion validity.

A number of studies have revealed the lack of construct validity in assessment centers. It has been shown consistently that ratings on the same dimensions in different exercises do not correlate as highly as do ratings on different dimensions in a single exercise (Bycio, Alvaes, & Hahn, 1987; Russell, 1987; Sackett & Dreher, 1982). Nonetheless, a meta-analysis of the relation between outcomes of assessment center exercises and personality constructs has revealed that clients who are successful in assessment centers generally possess intelligence, social skills, achievement motivation, authoritarianism, and high self-esteem (Scholz, 1992). These cognitive and social skills may be related to the ability to discern what dimensions are being assessed and to perform well on them.

Several hypotheses on the lack of construct validity have been discussed in the literature (Bungard, 1987; Bycio et al., 1987; Kleinmann, 1991; Klimoski & Brickner, 1987; Klimoski & Strickland, 1977; Neubauer, 1980). One hypothesis has addressed how the lack of transparency among rating dimensions for the participants might influence construct validity (Bungard, 1987; Kleinmann, 1991). Assessment centers lack transparency for participants in at least two ways: Participants may not know the rating dimensions, and they may not know which behaviors are relevant for each rating dimension. It is unlikely that participants will simply just wait to be evaluated passively; instead, they will try to gain the best possible evaluation. The consequence would be to engage in appropriate “impression management” (Tedeschi, 1981).

Orne (1962) has developed the construct of “demand characteristics” in this context. Sources of demand characteristics regarding desired behavior in an assessment center could be previous experiences with assessment centers, the company image, spatial framing conditions, the exercise material, behavior of observers and other participants, and so forth. These assumptions mean that not only actual abilities or skills but also an adequate perception of the situation (i.e., recognizing correct, evaluated dimensions) have a major impact on assessment center performance.

If all exercises were completely transparent, then every participant in an assessment center would know which dimensions were being assessed. In this case, problems raised by differences in the participants’ assumptions regarding which behavior they should exhibit could be ruled out. This would also remove problems arising because participants receive good ratings through having the luck or ability to recognize the necessary dimensions, although this does not necessarily covary with their actual ability on these dimensions. However, if, in contrast, the requirements were to differ intra- or interindividually in transparency, this would influence convergent validity and, perhaps, even criterion validity. If identical dimensions across various exercises were to differ intra-individually in transparency, there would also be intrapersonal differences in performance on these dimensions, as the participant’s behavior (and hence its rating) would vary on them. This would reduce convergent validity. For example, if participants were to recognize or fail to recognize identical dimensions in two different exercises, they would probably both behave and be rated more similarly if they would participants who had recognized only one of the identical dimensions. If the transparency of dimensions were to differ inter-individually across various exercises and there was a stable, cross-situational ability to recognize such implicit dimensions, participants would perform well in assessment centers in part because they recognized the dimensions and would possibly gain better ratings on the validity criteria because they also recognize the requirements of their jobs.

To test the hypothesis that rating dimensions in assessment centers lack transparency and that assessment center ratings depend on the recognition of these dimensions, Bungard has performed an initial exploratory study. Several weeks at the end of the assessment centers, participants in several assessment centers were asked about their assumptions regarding the measures that were used in a presentation exercise. It surprised how many different assumptions participants had as a possible basis for assessment. It was demonstrated that subjective ratings received by participants varied as a function of agreement between assumed requirements and actual evaluation measures.

Correspondence concerning this article should be addressed to Martin Kleinmann, Institut für Psychologie, Christian-Albrechts-Universität, Olshausenstrasse 40/60, D-24098, Kiel, Federal Republic of Germany.
Nonetheless, it should be noted that these effects were not tested statistically. Furthermore, Bungard (1987) did not specify whether participants had already been informed about their performance at the time of the study or whether the interviewer knew the participants’ ratings or the study hypotheses.

Because Bungard (1987) tested the hypothesis that performance in an assessment center depended on recognizing the rating dimensions in only one exercise, a series of questions remain unanswered: (a) Does performance in an assessment center in other types of exercises also depend on recognizing rating dimensions? (b) Do persons receive high ratings in assessment centers because they recognize the rating dimensions, and does this mean that they are also rated higher on the validity criterion because they also know the requirements of their jobs? That is, is there a stable, cross-situational ability to recognize such implicit requirements? (c) What effects do intraindividual differences in recognizing the correct requirements in various situations have on convergent validity?

All of these issues could not be addressed in the present study. However, the following hypotheses were formulated:

Hypothesis 1. In an assessment center, participants develop different assumptions about which dimensions will be rated. These assumptions differ in the extent to which they correspond to actual requirements; that is, people differ in their cross-situational ability to recognize the relevant dimensions.

Hypothesis 2a. The more precisely people have recognized the target dimensions, the better their ratings in assessment centers.

Hypothesis 2b. People perform better on dimensions that they have recognized than on those that they have not recognized.

Hypothesis 3. If identical dimensions are either recognized or not recognized to an equal extent in two different exercises, the performance on the two exercises will be rated more similarly than when the dimension is recognized in only one of the exercises.

In summary, the present study was designed to test whether participants’ recognition of the dimensions rated in assessment centers influenced their ratings, whether participants differed in their ability to recognize the necessary dimensions and to behave accordingly, and whether intraindividual differences in the recognition of dimensions had an effect on convergent validity.

Method

Overview

A 1-day assessment center was constructed containing five exercises. Potentially ratable target dimensions for each exercise were worked out in a pretest. Some of these dimensions were used to rate participants in the subsequent assessment center. Observers received assessment center training. In the assessment centers, participants were rated on the selected dimensions as well as on other evaluation measures. In addition, participants’ assumptions about which dimensions would be rated were recorded. The correctness of participants’ assumptions regarding rated dimensions was classified and compared with actual observer ratings on these dimensions.

Subjects

Sixty college students took part in the pretest. Most of them were studying business management. The sample contained 23 women and 37 men who were between 23 and 31 years of age. There were six observers: two female and four male psychology students who were between 24 and 27 years of age.

The 56 participants in the main study were college students who were about to take their final examinations. Most of them were also studying business management. The sample contained 18 women and 38 men who were between 24 and 30 years of age. The male observers were either psychology students or graduate psychologists. Three observers were women and six were men; their ages were between 25 years and 30 years.

Prestudy

Participants’ assumptions about the behaviors relevant in the assessment center were assessed in a prestudy in which participants performed individual exercises. A 1-day assessment center was organized with five exercises (three group discussions either with or without role assignments and with different goals, a presentation exercise, and an in-basket exercise) based on designs reported by Jeserich (1981). Twelve persons worked on each exercise. Two subgroups, each containing 6 persons, were asked to perform the first group discussion in two separate rooms at the same time. Three persons observed each discussion. Then, the 12 participants in the two discussions were given small cards and asked to write down their hypotheses regarding which behavior would be rated by the observers. In a second step, they were asked to write down any other target dimensions that they thought it would be possible to observe during the exercise. The 12 participants then reassembled in one room. The cards containing participants’ responses to both questions were pinned up on a board. Cards that seemed to address similar rating dimensions were pinned together in clusters. From these clusters, the participants had to select the eight clusters that they considered to be the best observation measures. From these eight clusters, they had to choose four that they considered to be the most plausible ones to be observed in the exercise. Different subjects repeated this procedure for the two other group discussions.

For the presentation exercise, 12 persons once more prepared themselves in two separate rooms, with three observers in each room. Subsequently, the same procedure was used as was followed in the group discussion. The in-basket exercise was carried out by 12 other persons. The subsequent procedure for collecting dimensions was the same as that used in the group discussion. The four out of eight dimensions that seemed to be most probable in individual exercises were selected as the basis for rating participants in the actual assessment center.

Observer Training

Observers were familiarized with the exercises, the four dimensions selected per exercise, observation errors, and so forth in 3 days of observer training. The design of this training was based on Jeserich’s model (1981). I selected three observers at random for each assessment center while retaining the same gender distribution in each assessment center. Observers were informed that the purpose of the study was to obtain indications for an optimal design of exercises by surveying participants for their opinions.

Main Study

Assessment centers were carried out with 56 participants at a conference hotel. They were set in a job-applicant-training framework. A personnel consulting company acted as organizer. Participants were recruited by a student organization that regularly organized seminars with outside companies for interested college students. Announcements did not mention that the job applicant training was related to a scientific study. Interested persons could contact the student organization by tel-
ephone, and, after payment of a course fee, they received written confirmation of their participation giving the location and time of the training course. None of the participants had previously attended an assessment center.

Along with six participants and three observers, two assistants took part in each assessment center. Their task was to ensure that the organization ran smoothly and to survey participants during breaks.

After the participants had been welcomed by a representative of the student organization, the presenters informed the participants about the program. They were told that, after each single exercise, the observers would withdraw to rate the exercise. In addition, they were told that they would also be asked about their perception of the exercises after each single exercise, but that this survey was not part of the assessment center and would not be rated as it was being carried out on behalf of the personnel consultancy that hoped to gain information that would improve exercises.

After this introduction, the first exercise was presented and carried out. At the end of the first exercise, observers left the room to individually rate participants on the four dimensions. The assistants distributed questionnaires to participants, who were once more assured that this part of the assessment center would not be rated because its only purpose was to record participants' assumptions about which behavior was rated and their action strategies. The actual question used was: "One of my assumptions during this exercise was that it was very important to . . . . As a result, I tried to act in the following way: . . . ." This sentence was repeated several times in the questionnaire. The assistants collected the completed questionnaires, and the observers reentered the room and presented the next exercise.

This alternation between exercise and survey was repeated for all five exercises. After the last exercise, the observers withdrew to confer on one rating for each participant on a 5-point rating scale for each dimension and to rank order the management potential of each individual candidate. Observers also reported whether they considered the candidate to be eligible for a training post in a large company.

After a short break, participants were presented with a list of the eight potential requirements for the first exercise and were asked to read it. Then the participants' own questionnaires, containing their answers on assumed dimensions, were returned to them. Their task was now to assign the dimensions that they had perceived during the exercise to the dimensions on the list. The dimensions on the list were defined and presented just like those rated by the observers during the exercise. Each perceived dimension should have been classified to one dimension on the list. Subjects could classify several of their assumptions to one dimension on the list as long as this provided the closest correspondence. If participants' dimensions could not be classified unequivocally, this should have been noted. Then, the assistants collected the papers.

This procedure was repeated for all exercises. Perceived dimensions were assigned to potential dimensions only after the end of the assessment centers. In this way, participants would not obtain any additional cues regarding potential dimensions during the assessment centers that could be applied in the following exercises. As described above, four of the eight possible dimensions were rated in each of the five exercises. Hence, the number of correctly recognized dimensions could vary between 0 and 20. The lists containing the eight different dimensions per exercise were designed so that the sequence of correct, while rated, dimensions varied randomly on the lists.

Results

I first hypothesized that, in an assessment center, participants develop assumptions about the dimensions assessed that differ in the extent to which they correspond to actual requirements; that is, people differ in their cross-situational ability to recognize the relevant dimensions.

<table>
<thead>
<tr>
<th>Correctly recognized dimensions</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–4</td>
<td>6</td>
<td>10.7</td>
</tr>
<tr>
<td>5–8</td>
<td>35</td>
<td>62.5</td>
</tr>
<tr>
<td>9–12</td>
<td>12</td>
<td>21.4</td>
</tr>
<tr>
<td>13–16</td>
<td>3</td>
<td>5.4</td>
</tr>
<tr>
<td>17–20</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Note. Mode = 6, median = 7, range = 15.

As Table 1 shows, individual participants varied greatly. It was also notable that no participants recognized all rating dimensions. It was clear that individuals differed in the accuracy with which they recognized dimensions and that accuracy was low in general.

The variation of recognized dimensions in individual exercises was also considerable (see Table 2). In the presentation exercise, for example, a text on scenario techniques (Reibnitz, 1983) had to be worked out, presented, and examined for its usefulness. Only 9% of the participants recognized the releva...
vance of the dimension "own power of judgment," whereas 68% recognized "didactic ability."

In all exercises, there was no single dimension that all participants recognized as being relevant; that is, no dimension was transparent for all participants.

Each assessment center contained five exercises. For each exercise, participants had to recognize four dimensions that were being rated by observers. In different exercises, dimensions were sometimes the same and sometimes different. If there is such a thing as an ability to recognize correct rating dimensions, it should have been possible to identify participants who were able to recognize appropriate dimensions across several exercises. Accordingly, there should have been participants who recognized a particular large number or a particularly small number of dimensions in the various exercises. This could be tested statistically by estimating internal consistency among the exercises (i.e., test reliability). Items were individual exercises that could take a value between 0 and 4 (number of recognized dimensions). The test had a length of five items (number of exercises). If there was no recognition of correct dimensions, internal consistency should have only fluctuated randomly around zero.

I estimated internal consistency among the exercises with Hoyt's (1941) method, which uses the analysis of variance. In the present case, the F test was highly significant, $F(55, 220) = 2.35, p < .01$; that is, a statistically significant proportion of the variance could be attributed to differences in the ability to recognize requirements. Using this method (for more details, see Hoyt, 1941), I found the internal consistency correlation to be .59. It might have been higher if the test had contained a larger number of items. Hence, recognition of dimensions did not vary randomly from exercise to exercise: Participants differed in their ability to recognize dimensions in individual exercises. I deal with the effects of this finding on assessment center criterion validity in the Discussion section.

In my second hypothesis, I anticipated that the more precisely people recognized the target dimensions, the better their ratings in assessment centers. Participants varied in the extent to which they recognized correct, action-guiding target dimensions. As the probability of recognizing the right dimensions could depend on the number of assumptions reported, I first tested whether the number of correctly recognized dimensions was greater among participants who had reported a particularly large number of assumptions. There was no notable correlation between the number of assumptions and their accuracy ($r = .06, n s$).

The number of dimensions recognized correctly by the participants was then related to the measures recorded. I used three measures to assess performance in assessment centers: (a) a sum of ratings on dimensions, (b) an estimation of management potential, and (c) a decision about whether or not participants would be given a job. All three measures correlated with the number of correctly recognized dimensions ($n = 56$) as follows: total rating on assessment center dimensions, .30 ($p < .05$); management potential, .44 ($p < .01$); and employment decision, .25 ($p < .05$). Participants' ratings on all three measures varied as a function of the number of dimensions that they had recognized. Dimension recognition was thus linked to the assessment of the participant.

In Hypothesis 2, I also predicted that people will perform better on dimensions that they have recognized than on those that they have not recognized. Participants received better ratings on dimensions that they had recognized ($M = 3.5$) than on dimensions that they had failed to recognize ($M = 2.8$). $t(55) = 6.6, p < .01$. Hence, intrapersonal variations in ratings given to individuals on various dimensions also depended on the person's recognition of the relevant dimensions.

In Hypothesis 3, I predicted that if identical dimensions were either recognized or not recognized to an equal extent in two different exercises, performance on the two exercises would be rated more similarly than when the dimension is recognized in only one of the exercises.

Consider three groups of participants: Groups 1, 2, and 3. As an example, assume that Group 1 has not recognized an identical dimension in either of two exercises whereas Group 2 has recognized this dimension in one of the two exercises. Assume that Group 3, on the other hand, has recognized this dimension in both exercises. It could thus be anticipated that Groups 1 and 3 would each receive more similar ratings than would Group 2. Convergent construct validity would therefore be correspondingly higher.

To test this hypothesis, I compiled all pairs of ratings of participants on each pair of two identical dimensions in different exercises. I separated these pairs of ratings into three groups. Group 1 contained pairs of ratings in which neither of the two identical dimensions had been recognized. In Group 2, only one of the dimensions had been recognized, and in Group 3, both dimensions had been recognized. Because participants differed in the frequency with which they recognized dimensions, a random pair of ratings was drawn for each participant from each of the three groups. An intraclass correlation coefficient was computed on these three sets of ratings. However, beforehand, I tested whether the variance of the ratings differed in the three sets, as this could possibly have led to artifacts. There were no significant differences in the variance of the pairs of ratings.

The coefficients for the three groups show that correlations between ratings on identical traits (dimensions) in various exercises were higher when participants either recognized or did not recognize these traits as rating dimensions to the same extent across the different exercises. For example, by Fisher's Z transformation, intraclass correlations for ratings of identical dimensions in two different exercises were .62 for Group 1 ($n = 56$), .32 for Group 2 ($n = 52$), and .57 for Group 3 ($n = 42$; Group 1 vs. Group 2, $p < .01$; Group 2 vs. Group 3, $p < .05$). Correlations between ratings of identical dimensions were lower for participants who recognized these traits in only some of the various exercises. This meant that participant ratings were more homogeneous for identical dimensions across different exercises if participants did not vary in the extent to which they recognized the dimensions in the individual exercises. Therefore, recognition of dimensions had a moderating effect on the size of convergent validity.

Discussion
Participants in assessment centers have to cope simultaneously with two tasks: (a) They have to respond to the content
of the given exercise, and (b) they have to consider whether or not their behavior will increase their chances of success. The potential lack of transparency in the dimensions of assessment center exercises increases the probability that participants will select incorrect action strategies. Differences in the transparency of dimensions across individuals are likely to influence construct and criterion validity.

It is surprising how much assessment center dimensions actually lack transparency for the participants. The degree of transparency also varies strongly between single dimensions. The consequence of these differences in transparency—that participants who recognize appropriate dimensions receive better ratings—is not surprising. Experimental psychology assumes that such effects contribute to an increase in error variance (Bungard, 1987). This would also apply to assessment centers if recognition of relevant dimensions were to vary at random between participants. However, this cannot be confirmed; instead, it seems likely that participants vary in the extent to which they are able to recognize which dimensions are relevant. This agrees with the finding that participants who are successful in assessment centers possess a high level of social skills (Scholz, 1992).

What are the implications for criterion validity? Sackett and Dreher (1982) have shown that the agreement between predictor and criterion variance in assessment centers cannot be obtained—as postulated—by assessing traits. However, there are very few other, empirically tested ideas on why valid predictions of career success and the like can be derived from assessment centers (see Klimoski & Brickner, 1987). It is still unclear whether some of the common variance between predictor and criterion is determined by the common recognition of the relevant dimensions in predictor measurement and criterion measurement. Such a hypothesis seems plausible, as real-life promotion criteria are probably also characterized by a high lack of transparency (see Bungard, 1987). If persons are assumed to differ in the extent to which they recognize which dimensions are relevant, and this is important for both assessment center ratings and criterion, it is conceivable that this factor provides an explanation of the common variance in validity studies.

However, an explanation of this issue requires further research aimed at (a) achieving a detailed theoretical understanding of the ability to recognize relevant dimensions, and (b) testing the susceptibility of criterion measures to this ability.

Another important finding of this study is that the level of convergent construct validity is influenced by the recognition of dimensions. Ratings given to individual participants in assessment centers are more similar when participants either recognize or do not recognize identical dimensions in different exercises and, correspondingly, behave in a more similar manner. This finding brings into question the frequent call to drop the use of trait-oriented rating systems in assessment centers (e.g., Robertson, Gratton, & Sharphey, 1987).

Nonetheless, it has to be noted that the recognition of dimensions alone is not sufficient to explain why there is little convergent validity in assessment centers. Even when these moderating effects are taken into account, the convergent correlation coefficients remain below a theoretical optimum. This could be because it is not possible to assess individual behaviors independently with rating dimensions in assessment centers. For example, if a test assesses two constructs, test material is presented in succession, making it possible to obtain a good rating on Requirement A while having a low rating on Requirement B. This is not the case in assessment centers: Several dimensions are tested simultaneously. Abilities on individual dimensions cannot be measured independently from other dimensions. If an assessment center exercise contains a rating dimension labeled ability to argue, it is bound to be confounded with the evaluation of another dimension labeled ability to assert oneself (Neubauer, 1980). People who cannot argue well are less able to assert themselves. For convergent construct validity, this means that the correlation between identical dimensions also depends on their confounding with other dimensions in the same exercise. This confounding is a possible explanation for the reduction in the size of convergent correlation coefficients.

It still must be asked whether or not it would be more appropriate to tell participants about the dimensions and about which behaviors will be evaluated. This can be countered by asking whether assessment centers would then still retain their validity. Further research is needed to clarify how much of the variance in criterion validity is determined by the ability to recognize the relevant dimensions.

References


ARE ASSESSMENT CENTERS TRANSPARENT?


Received October 17, 1991
Revision received May 10, 1993
Accepted May 10, 1993

Instructions to Authors

Manuscripts submitted for publication in *The Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, manuscripts must be clearly written in concise and unambiguous language and must be logically organized. The goal of APA primary journals is to publish information that is "new, true, important, and comprehensible" (A. D. Beck, 1976, *The Scientific Journal: Editorial Policies and Practices*, p. 30, St. Louis, MO: Mosby).

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). Manuscripts not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract containing a maximum of 960 characters and spaces (which is approximately 120 words) typed on a separate sheet of paper. Typing instructions (e.g., all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are subject to editing for sexist language.

The journal will publish both regular articles, or *Feature Articles*, and Research Reports. Authors can refer to recent issues of the journal for approximate length of regular articles. (Total manuscript pages divided by 4 is an estimate of total printed pages.) Research Reports feature shorter manuscripts, such as important replications or studies that involve a relatively narrow but unique contribution. For Research Reports, length limits are exact and must be strictly followed. The author may request Research Report status at the time of submission or the editor may suggest that a regular-length submission be pared down to Research Report length. In preparing Research Reports, authors should set the character and space limits at 60 characters per line and must not exceed 364 lines of text (approximately 14 pages of text proper). These limits do not include the title page, abstract, references, tables, or figures.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 5.24). As this journal is a primary journal that publishes original material only, APA policy prohibits as well publication of any manuscript that has already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data upon which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 6.25). APA expects authors submitting to this journal to adhere to these standards. Specifically, authors of manuscripts submitted to APA journals are expected to have their data available throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242 (or see "Ethical Principles," December, 1992, *American Psychologist*, Vol. 47, pp. 1597-1611). The journal will accept submissions in masked (blind) review format only. Each copy of a manuscript should include a separate title page with author names and affiliations, and these should not appear anywhere else on the manuscript. Furthermore, author identification notes should be typed on the title page (see *Manual*). Authors should make every effort to see that the manuscript itself contains no clues to their identities. Manuscripts not in masked format will not be reviewed.

Authors must submit five (5) copies of the manuscript. The copies should be clear, readable, and on paper of good quality. A dots matrix or unusual typeface is acceptable only if it is clear and legible. In addition to addresses and phone numbers, authors should supply electronic mail addresses and fax numbers, if available, for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to Philip Bobko, In-coming Editor, *Journal of Applied Psychology* Editorial Office, P.O. Box 130, Skillman, NJ 08538.