Validity of Behavioral Assessment for Predicting Military Recruiter Performance

Walter C. Borman
Personnel Decisions Research Institute, Minneapolis, Minnesota

This article describes results of an evaluation experiment to determine the validity of an assessment center designed to help select U.S. Army recruiters. Sixteen experienced recruiters assessed 57 soldiers entering the Army's recruiter school, and assessment ratings were correlated with subsequent performance in training. A composite of assessment ratings yielded corrected validities of near .50. In contrast, "first impression" evaluations, ratings based on a structured interview, and scores on a test carefully developed to predict success in military recruiting correlated near zero with performance in training. Possible reasons for these findings are discussed. Further, results of the study confirm that valid assessment does not require behavioral scientists as assessors, and analyses suggest that statistical composites of assessment ratings on individual exercises may be slightly more valid than "clinical" consensus judgments made after discussion of assessment performance. Suggestions are made to study assessment ratings employing a person perception framework. Such an approach should increase scientific understanding of the assessment process and enhance validity of the method.

Assessment centers have been especially successful in predicting success in management jobs (e.g., Byham, 1970; Cohen, Moses, & Byham, 1974; Dunnette, 1971; Finkle, 1976; Howard, 1974). Several studies demonstrate that assessment ratings are valid predictors of progress in management (e.g., Bray & Grant, 1966; Moses & Boehm, 1975). Further, the method has been lauded as an important improvement over pencil-and-paper testing, and in studies that have employed both tests and assessment, substantial incremental validity is gained using assessment ratings in addition to test scores (e.g., Bray & Grant, 1966; Mitchel, 1975; Wollowick & McNamara, 1969).

Although the majority of reports are favorable regarding the use of assessment to predict performance, one study (Hinrichs, 1969) showed that ratings of management potential made after a review of personnel files correlated .46 with assessment ratings, suggesting that assessment might to some extent duplicate a much simpler and less costly procedure. Also, Grant and Bray (1969) found that ratings on certain dimensions based on carefully structured interviews with managers correlated well with assessment judgments; this suggests that an interview might be used as effectively as assessment, although validity of the interview has of course been widely questioned (e.g., Schmitt, 1976). The point is that certain questions have been raised regarding the relative validity of assessment compared with less expensive procedures (see also Cascio & Silbey, 1979). Although the general levels of validity have been high for assessment,
more research is needed that employs alternate predictors within the same sample.

Another issue associated with assessment involves the kinds of jobs for which simulations are appropriate. Most assessment centers have simulated aspects of management jobs, but centers have been developed to estimate potential for success in sales (Bry & Campbell, 1968), leadership in the military (Veule & Campbell, 1974; Byham, Note 1; Helme, Willemin, & Grafton, Note 2), police work (Dunnette & Motowidlo, Note 3), the pharmacy profession (Silzer, Note 4), vocational rehabilitation counseling (Byham, Note 1), and the job of stockbroker (Hellervik, Hunt, & Silzer, Note 5), among others. Future research should continue to explore applications of assessment in different kinds of jobs.

Still another issue raised by Bray and Grant (1966) in their seminal work, and also addressed by Wollowick and McNamara (1969), is the relative contribution to validity of different types of assessment exercises. In these two studies distinct differences in levels of validity were found between exercises, raising the possibility that some types of simulations are inherently better than others at eliciting behaviors that prove to be valid indicators of future performance. More work is needed to explore this possibility.

The present research addresses these issues. An assessment center was developed to evaluate a person's potential for success as a U.S. Army recruiter (Borman, Note 6). Then an experiment was conducted to test the validity of the assessment program. A summary description of the exercises used in the assessment program is given in Table 1.

Method

Subjects and Procedure

Subjects were 57 soldiers (all but one were men) entering the U.S. Army recruiter school; as such, they had no previous recruiting experience. Assessors were 16 experienced and successful Army recruiters who had just graduated from a course for future station commanders.

The 16 recruiters had no previous experience as assessors and, for the most part, no behavioral science training. To prepare the group for the assessment program, I administered a 2-day training course focusing on (a) behavior observation principles, (b) attention to appropriate behavioral cues during the different assessment exercises, and (c) proper use of the assessment rating scales developed in previous job analysis research (Borman, Toquam, & Rosse, Note 7). Assessment for each group of six subjects lasted 1 day; the total evaluation program continued for 10 days.

Twelve assessors were assigned in pairs to subjects in the structured interview and five role-playing exercises. Four additional assessors, working in pairs, provided "first impression" ratings, evaluations of subjects made after only brief exposure to them during a 15-minute introductory session at the beginning of the assessment day. In this session subjects introduced themselves, received general instructions regarding the day's activities, and occasionally asked questions about the program. The author made physical attractiveness and likability ratings of each subject on the basis of interaction with them in the introductory meeting and during exercise instruction sessions with them before each exercise.

Assignments of assessors to subjects and exercises were made such that (a) assessor pairs never worked together more than 1 day during the 10-day program to avoid possible rater effects on interrater agreement results; (b) assessors were assigned to the main group of 12 for 7 or 8 of the 10 days and to the group of 4 for the other days; and (c) each pair viewed each subject in one and only one of the six exercises during the day (e.g., Rater Pair 1 viewed Subject A in the personal interview, Subject B in the cold calls, etc.; Rater Pair 2 viewed Subject A in the cold calls, Subject B in the interviews, etc.). This latter feature is very important because it provided the opportunity to evaluate the relative contribution to validity of each exercise.

To the best of my knowledge, the present experiment provides the first legitimate evaluation of the validity of individual exercises. Specifically, with exercises late in an assessment program, assessors providing evaluations have typically seen a subject in earlier exercises or during other assessment center activities, and previous perceptions of a subject's effectiveness almost certainly affect ratings on the later exercises. In this study, when an assessor pair rated a subject in an exercise, it was the first time the pair viewed that subject's performance. Also, the scheduling of the day's activities made assessors' daily interactions with subjects they had not yet evaluated very unlikely.

For each exercise assessors were instructed to evaluate subjects on the 4-11 dimensions relevant for that exercise. They also provided an overall performance rating for each exercise.

Training Criteria and Research Design

Criteria from two phases of the recruiter training program were employed to evaluate the validity of assessment. A Phase 1 composite of scores on three objective tests comprised one criterion. The tests measured mastery of prospecting and selling techniques taught in Phase I. An intraclass correlation of .94, indexing the composite's reliability, suggests a homogeneous Phase 1 criterion measure.

In Phase 2, students practiced the telephoning and interviewing techniques they had learned in Phase 1. In small groups and with the help of a staff instructor, they practiced their skills in hypothetical recruiting situa-
Table 1
Summary of Assessment Program

1. **Structured interview.** Assessors ask a series of questions targeted at the subject’s level of achievement motivation, potential for being a “self-starter,” and commitment to the Army.

2. **Cold calls.** Subject has an opportunity to learn a little about three prospects and must phone each of them for the purpose of getting them to come into the office. Assessor role players have well-defined characters (prospects) to portray.

3. **Interviews.** Two of the three cold-call prospects agree to come in for an interview. The subject’s job is to follow up on what was learned in the cold-call conversations and to begin promoting Army enlistment to these people. A third “walk-in” prospect also appears for an interview with the subject.

4. **Interview with concerned parent.** Subject is asked to prepare for and conduct an interview with the father of one of the prospects he or she interviewed previously.

5. **5-minute speech about the Army.** Subject prepares a short talk about an Army career that he or she delivers to the rest of the group and to the assessors.

6. **In-basket.** Subject is given an in-basket filled with notes, phone messages, and letters on which he or she must take some action.

...
Table 2
Correlations Between Assessment Ratings

<table>
<thead>
<tr>
<th>Assessment measures</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. First impression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Physical attractiveness</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Likability</td>
<td>02</td>
<td>35**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Structured interview</td>
<td>14</td>
<td>25</td>
<td>13</td>
<td>26*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Cold calls</td>
<td>35**</td>
<td>12</td>
<td>05</td>
<td>32*</td>
<td>48**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Interviews</td>
<td>02</td>
<td>19</td>
<td>05</td>
<td>27*</td>
<td>49**</td>
<td>65**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Interview with parent</td>
<td>11</td>
<td>28*</td>
<td>19</td>
<td>27*</td>
<td>65**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. In-basket</td>
<td>07</td>
<td>20</td>
<td>14</td>
<td>26*</td>
<td>30*</td>
<td>65**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Speech</td>
<td>07</td>
<td>20</td>
<td>14</td>
<td>16</td>
<td>41**</td>
<td>49**</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Consensus assessment rating</td>
<td>12</td>
<td>23</td>
<td>23</td>
<td>44**</td>
<td>70**</td>
<td>76**</td>
<td>75**</td>
<td>32</td>
<td>60**</td>
</tr>
</tbody>
</table>

Note. N = 57. Decimal points are omitted.
* p < .05, two-tailed. ** p < .01, two-tailed.

.48 for relationships between cold calls, interviews, interview with parent, and speech assessment ratings). Ratings on the structured interview and in-basket, presumably tapping somewhat different personal characteristics, correlate less highly with other exercise ratings (Mdn r with the other four exercises = .26).

In terms of contributions of ratings on individual exercises to the overall consensus rating, the three interactive role-playing exercises yield the highest correlations with consensus judgments. The structured interview and in-basket ratings were accorded less weight in the consensus ratings. Finally, the first impression, physical attractiveness, and likability ratings correlate positively with assessment ratings on exercises that involved face-to-face assessor-subject contact, but the correlations are generally low (Mdn r = .14; rs = .12, .23, and .23, respectively, with the consensus ratings).

Validity

Table 3 presents validity coefficients between the various assessment scores and the three criteria. Regarding the Phase 1 composite, ratings on all five role-playing exercises yield validities of .32 (p < .05) or greater. Correlations are not as high for the Phase 2 performance criterion, with only one of the validities significant at the .05 level; for the time-to-complete criterion, validities for the simulations range from .11 to .42 (one of the five validities was significant at the .05 level, and one was significant at the .01 level). One would expect approximately 1 significant correlation (p < .05) of the 15 to occur by chance, and 8 were found.

In contrast, the first impression, physical attractiveness, and likability ratings showed nonsignificant validities. Likewise, the structured interview ratings and pencil-and-paper test scores yielded generally lower validity coefficients.

Interestingly, when the assessment ratings are combined "mechanically" (unit weighting ratings on each dimension relevant for an exercise and pooling these ratings across the six exercises) rather than "clinically," as was the case for the consensus ratings, validities reach .48 (p < .01), .35 (p < .05), and −.33 (p < .05), respectively, against the three criteria.

In addition, there is reason to believe that the assessment validity estimates from this research (against training criteria) may be conservative in general. The seven persons who were evaluated during the assessment program but who were not included in the validity analyses because they dropped out before training tended to have either very high or very low assessment ratings. This means that the range of assessment ratings is restricted. Furthermore, four subjects with relatively low assessment scores were eliminated from training because of inadequate Phase 1 performance. This further restricted the range of assessment scores for validity analyses against Phase 2 criteria.

Accordingly the formula for restriction of range (Gulliksen, 1950) was applied, and the corrected validity coefficients for the unit-
Table 3
Correlations Between Assessment Measures and Training Criteria

<table>
<thead>
<tr>
<th>Variable</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Composite performance</td>
<td>Composite performance</td>
<td>Time to complete</td>
</tr>
<tr>
<td>Assessment measures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First impression</td>
<td>23</td>
<td>-18</td>
<td>02</td>
</tr>
<tr>
<td>Physical attractiveness</td>
<td>-06</td>
<td>-20</td>
<td>06</td>
</tr>
<tr>
<td>Likability</td>
<td>06</td>
<td>04</td>
<td>14</td>
</tr>
<tr>
<td>Structured interview</td>
<td>07</td>
<td>26</td>
<td>00</td>
</tr>
<tr>
<td>Assessment ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cold calls</td>
<td>32*</td>
<td>-03</td>
<td>-26</td>
</tr>
<tr>
<td>Interviews</td>
<td>34*</td>
<td>15</td>
<td>-28</td>
</tr>
<tr>
<td>Interview with parent</td>
<td>41**</td>
<td>24</td>
<td>-29*</td>
</tr>
<tr>
<td>In-basket</td>
<td>33*</td>
<td>36*</td>
<td>-11</td>
</tr>
<tr>
<td>Speech</td>
<td>32**</td>
<td>27</td>
<td>-42**</td>
</tr>
<tr>
<td>Consensus rating</td>
<td>38**</td>
<td>27</td>
<td>-24</td>
</tr>
<tr>
<td>Navy test keys</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human relations</td>
<td>-06</td>
<td>15</td>
<td>-27</td>
</tr>
<tr>
<td>Selling</td>
<td>02</td>
<td>17</td>
<td>-12</td>
</tr>
<tr>
<td>Organizing</td>
<td>-01</td>
<td>07</td>
<td>-15</td>
</tr>
<tr>
<td>Overall performance</td>
<td>-09</td>
<td>15</td>
<td>-09</td>
</tr>
<tr>
<td>Army test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental key</td>
<td>04</td>
<td>05</td>
<td>-21</td>
</tr>
<tr>
<td>Unit-weighted composite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratings on relevant dimensions</td>
<td>48**</td>
<td>35*</td>
<td>-33*</td>
</tr>
<tr>
<td>(all exercises)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. For Phase 1, n = 50; For Phase 2, n = 46.
Decimal points are omitted.
* p < .05 (two-tailed). ** p < .01 (two-tailed).

A weighted composite in Table 3 against the three criteria are, respectively, .53, .50, and -.49. These coefficients are considerably higher than the uncorrected coefficients, particularly for the Phase 2 criteria. Validities for the test variables changed very little when the restriction of range formula was applied.

Discussion and Conclusions

Results of this evaluation study indicate that the assessment program is reasonably successful in predicting recruiter training performance. After correction for restriction of range, composite assessment ratings have a correlation near .50 with success in training. These levels of validity are certainly higher than those found in previous validation work with Army recruiters (Fischl, Note 8), although training criteria, not job performance criteria, were employed here.

Regarding individual assessment exercises, ratings for the speech, interview with parent, in-basket, and interviews exercises provide the highest validities, with ratings on the cold calls simulation yielding reasonably good validity coefficients on two of the three criteria.

Also recall that there were a number of different competing methods of predicting success, although the relatively small sample size in the study renders formal statistical comparisons very difficult. Nonetheless, results suggest that initial impressions of an assessee have limited validity and, importantly, little impact on assessment exercise
ratings. Low correlations between exercise ratings and first impression, physical attractiveness, and likability ratings suggest that assessors' effectiveness judgments reflect much more than personal appearance and other superficial indicators of future performance.

The structured interview also proved to be less valid than the simulation exercises, and the tests yielded low validities as well. It is possible that the interview and tests will correlate more highly with job performance over time in recruiting.

Clearly, levels of validity for predictors depend on the nature of criteria. Assessment ratings likely tapped intellectual and performance-related individual differences associated with, respectively, Phase 1 and Phase 2 training criteria. In comparison, the interview presumably identified subjects' interests, past experiences, and accomplishments that may bear not so much on ability as on success in recruiting over the longer haul. Likewise, test scores measuring personal characteristics and vocational interests compatible with military recruiting may correlate more highly with longer term success on this job. Of course, assessment scores have also done well in predicting long-term effectiveness in careers (e.g., Bray, Campbell, & Grant, 1974).

An additional point is that results of the study confirm that experienced job incumbents with little or no formal behavioral science education can successfully be trained to provide valid assessments. Recruiters, like managers in previous studies, made reliable and valid judgments in the center.

Finally, in the present study composites of the assessment judgments are more valid in general than ratings on individual exercises, a finding similar to that of Bray and Grant (1966). Also, a mechanical composite proved to be slightly but consistently more valid than the clinically derived consensus ratings, supporting a previous finding by Wollowick and McNamara (1969). Of course the relatively small sample of subjects renders a powerful statistical comparison impossible here.

One reason to expect a mechanical composite to be especially effective in this research is that the assessor ratings on different exercises reflected completely independent observations of a subject's behavior, whereas assessors in previous studies have typically rated individual subjects in more than one exercise. Relatively low correlations between assessment ratings on different exercises may well be a result of this strict adherence to independence in the multiple observations of a subject's behavior.

It does not necessarily follow that assessment centers should, on the basis of this research, require maximum independence in behavior observations. It may be, for example, that assessors provide more valid ratings after viewing a subject in several different exercises. Person perceptions become more differentiated with increased rater-subject interaction (e.g., Koltuv, 1962), and possibly assessors' ratings made after several chances to view a subject's performance are more valid than those made earlier. The relative merit of stressing independence in observations across exercises or of ensuring multiple observations of subjects by the same assessor(s) should be evaluated in future research.

More generally, a promising approach to studying assessment ratings might be to apply a person perception framework. Similar to approaches taken in selection interview research (e.g., Hakel & Dunnette, 1970; Webster, 1964) and advocated for research on performance appraisal (e.g., Dunnette & Borman, 1979; Landy & Farr, 1980), more progress in scientific understanding of assessment is likely to emerge from examination of the person perception process in making assessment ratings. Clearly, perceptions of a subject's performance and potential develop during the assessment program, and many factors related to person perception bear on this development process. Assessors have first impressions of subjects, they may look for and attend to different kinds of behaviors, they may evaluate the same behaviors differently, and they are undoubtedly influenced by their own background and personal characteristics in sizing up individuals. In fact, the majority of factors identified by Wherry (Note 9), Borman (1978), and especially Landy and Farr (1980) as influencing performance appraisal ratings are likely to influence assessment ratings as well. Just as these writers argue for studying the performance rating process, I
recommend that attention be paid to the process associated with forming and making assessment ratings.

Reference Notes


References


Kolts, B. Some characteristics of intrajudge trait intercorrelations. Psychological Monographs, 1962, 76(33, Whole No. 552).


Webster, E. C. Decision making in the employment interview. Montreal, Quebec, Canada: Eagle, 1964.


Received February 13, 1981