An Examination of the Comparative Reliability, Validity, and Accuracy of Performance Ratings Made Using Computerized Adaptive Rating Scales

Walter C. Borman and Daren E. Buck
Personnel Decisions Research Institutes and University of South Florida

Stephen J. Motowidlo
University of Florida

Mary Ann Hanson
Personnel Decisions Research Institutes

Stephen Stark and Fritz Drasgow
University of Illinois at Urbana-Champaign

This laboratory research compared the reliability, validity, and accuracy of a computerized adaptive rating scale (CARS) format and 2 relatively common and representative rating formats. The CARS is a paired-comparison rating task that uses adaptive testing principles to present pairs of scaled behavioral statements to the rater to iteratively estimate a rater's effectiveness on 3 dimensions of contextual performance. Videotaped vignettes of 6 office workers were prepared, depicting prescribed levels of contextual performance, and 112 subjects rated these vignettes using the CARS format and one of the other competing format. Results showed 23%–37% lower standard errors of measurement for the CARS format. In addition, validity was significantly higher for the CARS format (d = .18), and Cronbach's accuracy coefficients showed significantly higher accuracy, with a median effect size of .08. The discussion focuses on possible reasons for the results.

Performance ratings remain the most commonly used job performance criterion in industrial and organizational (I/O) psychology (Landy & Farr, 1980; Murphy & Cleveland, 1995). For example, ratings are used as criteria in test validation research and as outcome measures in evaluations of the effectiveness of training treatments. In fact, it can be argued that ratings have been used to measure the effectiveness of virtually all I/O psychology interventions. Accordingly, there is considerable motivation to ensure that performance ratings are reliable and valid.

For many years, there was interest among I/O psychologists in rating formats and how manipulations of these formats might reduce rating errors and increase interrater reliability (e.g., Taylor & Wherry, 1951). Typically, the manipulations were fairly minor, such as the placement of the scale on the page (Madden & Bourdon, 1964) or the number of response categories on the scale (Bendig, 1954). However, in 1963, Smith and Kendall introduced a format that was fundamentally different from other rating instruments. Behavioral expectations scales, later referred to as behaviorally anchored rating scales (BARS), extended the notion of critical incidents (Flanagan, 1954) by placing behavioral statements at different places on the scale according to their effectiveness levels. Smith and Kendall reasoned that these behavioral anchors would help raters make more objective evaluations.

Additional related behavior-based formats were introduced. Latham and Wexley (1981) developed behavior observation scales, with favorably worded behavioral statements that required the rater to judge the frequency with which each behavior was exhibited by the ratee. Blazn and Ghiselli (1972) introduced the mixed standard scale (MSS). This format had an effective, middle, and ineffective behavioral statement representing each dimension. The rater's task was to indicate whether the ratee was more effective (+), less effective (-), or at the same effectiveness level (0) for each behavioral statement. An effectiveness scale score could then be derived for each dimension according to a scoring system that took into account both the number of +, –, and 0 ratings and the effectiveness level of the statement associated with each of these ratings.

Two other behavior-based formats might be mentioned. Borman's (1979) variant of BARS was the behavior summary scale (BSS). Rather than having as anchors specific behavioral examples describing individuals' performance, the BSS had more general
anchors at each of three or four effectiveness levels that summarized all of the behavioral anchors representing that effectiveness level and dimension. Finally, Kane (1986) introduced the performance distribution assessment method, which generates a modal performance-level score, a variance-in-performance score, and a negative range avoidance score (i.e., how well the ratee avoids ineffective performance).

Each of these behavior-based rating format ideas had appealing features. However, the following question arose: Does format make a difference relative to rating errors or the reliability and validity of the ratings generated by raters using the different formats? Not all of the relevant format comparison studies have been conducted, but the studies that have been completed generally have shown small differences between formats in terms of level of rater errors, reliability, validity, or accuracy. For example, reviews of format comparison studies (Landy & Farr, 1983; Schwab, Heneman, & DeCotis, 1975) concluded that the psychometric properties of the BARS format are probably not much better than the psychometric properties of graphic rating scales (GRS). Dickinson and Zellinger (1980) found no significant differences in discriminant validity between BARS and MSS. Borman (1979) found only small differences in halo, interrater reliability, and validity for BARS, the BSS, and GRS. Landy and Farr (1980) went so far as to estimate that the variance accounted for in psychometric quality by rating format was as little as 4%. In fact, they called for a "moratorium" on rating format research, citing the largely negative results.

For the past 20 years, Landy and Farr's (1980) suggestion has been followed for the most part. And yet, it still seems compelling to explore rating format ideas that might result in more reliable and valid judgments about work performance. Perhaps the way to conceptualize this quest to obtain more reliable, valid, and accurate estimates of job performance using ratings is to broaden the notion of "format" to include consideration of measurement models and raters' cognitive processes. If rating formats are thought of as mechanisms for helping raters to search for ratee behavior in an efficient and organized way and to translate this observed behavior into evidence relevant for making accurate evaluative judgments about a ratee's performance, then experimenting with different strategies for obtaining valid performance information seems worthwhile. Ideally, a rating format will be aligned with raters' cognitive processes in the sense that the format leads raters naturally to make evaluative judgments about observed ratee behavior veridical with actual ratee performance-related behavior (Borman, 1991). Accordingly, we concur with an observation made by Guion and Gibson (1988) that it may be premature to give up altogether on rating format research. Small adjustments made to present formats are very unlikely to result in higher reliabilities and validities; however, it still seems important to experiment with formats fundamentally different from those currently used in hopes of developing a format that is more in alignment with raters' cognitive processes or that somehow calibrates raters' perceptions to help them make more precise judgments about observed performance.

More generally, there is considerable motivation to improve criterion measurement in I/O psychology. The vast majority of the research in personnel selection, for example, has been devoted to predictor development, with much less emphasis placed on development and construct validation of criterion measures (e.g., Campbell, McCloy, Oppler, & Sager, 1993). If scientific understanding of the links between predictor constructs and job performance is to be enhanced, more effort toward improving and refining criterion measurement will be required.

One possible idea in the direction of a different rating measurement method started with consideration of Thurstone's (1927) law of comparative judgment in the context of performance rating process. Thurstone developed a method for scaling stimuli, on a relative scale, on the basis of paired-comparison judgments. His method was based on the proposition that people's reactions to stimuli are subjective and vary from one individual to another. If it is further assumed that these reactions are normally distributed, Thurstone's method can be used to place stimuli on an interval scale. In particular, pairs of behavioral statements might be presented to the rater with instructions to pick the statement that is more descriptive of the ratee. If interval scale judgments of ratee performance levels can be achieved with this method, the paired-comparison judgments may provide ratings that are more precise than those generated by other rating formats that use a linear numerical scale, which arguably provides only ordinal-level measurement. Another idea that might make the paired-comparison format even more effective is to apply an item response theory (IRT) adaptive testing orientation to the method. For example, the rater could be presented with a series of behavioral statements paired such that responses to each successive pair provide a more precise estimate of ratee performance.

Accordingly, our notion at this point for the present study was to develop and evaluate a paired-comparison rating task that used adaptive testing principles to help raters estimate a ratee's performance level through an iterative paired-comparison rating process. The idea was to present initially two behavioral statements associated with a dimension, one reflecting somewhat below average performance and the other reflecting somewhat above average performance. Depending on which statement the rater indicated was more descriptive of the ratee, the rating algorithm (as yet undeveloped) would select two additional behavioral statements, one with a scaled effectiveness level somewhat above the effectiveness value of the statement picked first as the more descriptive and the other with a scaled effectiveness level somewhat below the effectiveness value of that initially chosen statement. The rater's selection of the more descriptive statement for the second paired comparison would then revise the initial estimated ratee effectiveness level, and, as before, the algorithm would select two more statements for which effectiveness values bracketed the revised estimated performance level. Thus, analogous to adaptive testing, a ratee's "true" effectiveness level was to be estimated in an IRT sense by this iterative paired-comparison rating task that presents in sequence item pairs that maximize the amount of information about performance derived from each choice of an item.

After these paired-comparison rating scales had been developed, we conducted format comparison research to evaluate the reliability, validity, and accuracy of the paired-comparison computerized adaptive rating scales (CARS) in comparison to two popular and representative rating formats, GRS and BARS. The study was laboratory-based, using videotaped raters with scripted target performance scores to assess comparative levels of reliability, validity, and accuracy for the three formats. The first hypothesis was that the CARS format would provide more reliable ratings than the other two formats. That is, interrater reliability would be higher for
the CARS than for either the GRS or the BARS. This is because the iterative and adaptive nature of the CARS format should lead to more reliability and precision in those ratings. The second hypothesis was that the CARS ratings would be more valid using Borman’s (1977) correlational validity index. The rationale for this hypothesis was that the greater precision attained with the CARS would result in a more correct rank ordering of rates on each dimension, especially among rates near the middle of the performance distribution. Finally, we hypothesized that the CARS would yield more accurate ratings on Cronbach’s elevation, differential elevation, stereotype accuracy, and differential accuracy indices. The rationale for this hypothesis was very similar to that for validity. The added reliability and precision for CARS ratings to differentiate between rates and the better performance-level information on each rate that should be derived from the iterative CARS estimation procedure should result in more accurate performance ratings. The Method section describes the format comparison study in more detail.

Method

In this section, we describe the performance domain targeted by the CARS, BARS, and GRS; discuss scale development procedures for the three formats; describe development of the videotaped performances and target scores for these rates; and present the sample and procedures used in the rating format comparison research.

The Contextual or Citizenship Performance Domain

Borman and Motowidlo (1993) described the contextual performance concept as the culmination of research on three related topics: (a) organizational citizenship behavior (Organ, 1988), (b) prosocial organizational behavior (Brief & Motowidlo, 1986), and (c) a model of soldier effectiveness (Borman, Motowidlo, Rose, & Hansen, 1983). Each of these topics contributes to the notion of enlarging the performance domain beyond those behaviors that are directly linked to tasks.

Borman and Motowidlo’s (1993) contextual performance taxonomy was based largely on these three areas of research. Their five factors included (a) persisting with enthusiasm and extra effort as necessary to successfully complete one’s own task activities; (b) volunteering to carry out task activities that are not formally part of one’s own job; (c) helping and cooperating with others; (d) following organizational rules and procedures; and (e) endorsing, supporting, and defending organizational objectives.

Recent research by Coleman and Borman (2000) delineated the entire organizational citizenship behavior, prosocial organizational behavior, model of soldier effectiveness, and contextual performance domain into 27 distinct concepts. Expert judges (i.e., I/O psychologists) categorized these 27 constructs according to content. Factor analyses, multidimensional scaling analyses, and cluster analyses of the categorization data resulted in a three-dimensional model of contextual performance: (a) interpersonal support, (b) organizational support, and (c) job-task conscientiousness.

The present three-dimensional taxonomy was used in the development of the content model for the new CARS format. An important feature of the contextual performance domain is that the dimensions are likely to be similar across different jobs and organizations. This means that the dimensions from this study may have good generalizability to other jobs.

Coleman and Borman’s (2000) taxonomy seemed to be a relatively parsimonious way to depict the contextual performance domain. However, we conducted another study to confirm and refine the three-dimensional model. This study took as a starting point approximately 5,000 performance examples from 22 studies conducted by researchers at Personnel Decisions Research Institutes. In particular, we first determined whether each performance example represented task performance or contextual performance. Then, we used Coleman and Borman’s taxonomy as a guide to sort the approximately 2,300 contextual performance examples into one of the three contextual performance factors. Themes based on the content of the performance examples included in each dimension were noted and used to modify Coleman and Borman’s taxonomy and as a starting point for creating the CARS items.

The final conceptual model used in the CARS and the other two formats retained a three-dimension structure. The refined model is presented in Table 1. Now that adjustments to the three-dimension conceptual model of contextual performance were complete, it was time to develop the rating formats.

Scale Development Procedure

As mentioned, we developed three different rating formats for the research: CARS, GRS, and BARS.

CARS scale development. Behavioral statements were generated to reflect themes contained in each dimension at varying levels of effective-

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal support</td>
<td>Helping others by offering suggestions, teaching them useful knowledge or skills, directly performing some of their tasks, and providing emotional support for their personal problems. Cooperating with others by accepting suggestions, informing them of events they should know about, and putting team objectives ahead of personal interests. Showing consideration, courtesy, and tact in relations with others as well as motivating and showing confidence in them.</td>
</tr>
<tr>
<td>Organizational support</td>
<td>Favorably representing the organization by defending and promoting it as well as expressing satisfaction and showing loyalty by staying with the organization despite temporary hardships. Supporting the organization’s mission and objectives, complying with organizational rules and procedures, and suggesting improvements.</td>
</tr>
<tr>
<td>Conscientious initiative</td>
<td>Persisting with extra effort despite difficult conditions. Taking the initiative to do all that is necessary to accomplish objectives even if they are not normally a part of one’s duties and finding additional productive work to perform when one’s duties are completed. Developing knowledge and skills by taking advantage of opportunities within the organization and outside the organization through the use of one’s own time and resources.</td>
</tr>
</tbody>
</table>
ness. One hundred twenty-four statements were generated in all. Examples of these behavioral statements are presented in Table 2. These statements were retranslated in two separate stages. Initially, 37 Air Force noncommissioned officers from various specialties were asked to categorize the statements into one of the three contextual performance dimensions and to rate the effectiveness level for each on a 4-point scale (1 = very ineffective, 2 = somewhat ineffective, 3 = effective, and 4 = very effective). This process resulted in more than 65% of the raters agreeing on the intended dimension for all but 5 of the 124 statements. In fact, the average agreement for placing behavioral statements into their intended dimension was 90%. In a similar manner, there was good agreement on the intended effectiveness level of the statements. Raters agreed within 0.50 scale points of the intended effectiveness level for 80% of the statements (Borman, Hanson, Kubisiak, & Buck, 2000).

Results from the first retranslation stage were used to make revisions to the behavioral statements. After the statements were revised, 26 staff members from Personnel Decisions Research Institutes reclassified each of the 124 statements and rated its effectiveness level. None of these staff members were involved in generating the behavioral statements. This retranslation stage improved on the already favorable results found in the first retranslation. Average agreement of placing behavioral statements into their intended dimension was 96%, with all but 10 of the statements placed in the intended dimension by 85% or more of the raters. Agreement on the intended effectiveness level improved as well. Raters agreed within 0.50 scale points of the intended effectiveness level on 96% of the statements.

Between-rater agreement was good, with a mean standard deviation of 0.35 over all statements (Borman et al., 2000). On the basis of the results for the retranslation process, it was concluded that the behavioral statements reflected coherent contextual performance themes at well-agreed-on effectiveness levels. The effectiveness values were then transformed from the 1–4 scale to a 1–7 scale by using a straightforward linear transformation (e.g., 2 = 3.3, 3 = 5) to be more compatible with typical rating scales used in organizations.

Recall that the goal for CARS was to present these behavioral statements in a paired-comparison format and to use IRT and adaptive testing principles to help raters make more precise estimates of rates’ performance. A problem at this point was that there was no available IRT algorithm to estimate rates’ performance based on paired-comparison judgments. The paired-comparison case for IRT had never been fully developed. Stark and Drasgow (1998) took on this challenge. They reviewed several paired-comparison models and chose Zinnes and Griggs’s (1974) probabilistic unfolding model as the basis for algorithm development. Zinnes and Griggs’s model includes and extends research by Coombs (1950) and Thurstone (1927) on paired-comparison judgments. The model assumes that (a) rates have some standard of comparison, or ideal point, and (b) they will be more likely to choose the statement from a pair that more closely resembles that ideal point on the basis of the statement’s smallest absolute deviation from the ideal point. The model also assumes that errors around the raters’ ideal point and in the perception of each paired-comparison stimulus are normally distributed.

Stark and Drasgow (1998) developed an algorithm based on Zinnes and Griggs’s (1974) model to estimate item parameters and used this model to compute item characteristic curves. Estimation of item parameters for this model requires extremely large sample sizes. In an attempt to more efficiently develop stable and appropriate item parameters for the present research, subject-matter expert ratings of the effectiveness levels of behavioral statements were used in place of model-based item parameter estimates. Item characteristic curves were generated on the basis of these item parameters and were examined to determine the appropriateness of Zinnes and Griggs’s model for this application. The item characteristic curves compared favorably with researchers’ expectations.

In addition, conventional test and computer adaptive test (CAT) simulations were conducted to determine whether an adaptive test based on Zinnes and Griggs’s (1974) model could provide better performance ratings than a conventional test. Results showed that CAT standard errors of measurement were more uniform across latent trait values than the conventional-test standard errors. Consequently, the CAT resulted in more precise estimates of latent trait values with fewer items than the conventional test. Stark and Drasgow (1998) concluded that the adaptive test should “provide more accurate and precise estimates of job performance for most individuals than alternative rating scales formats” (p. 20). For more detail regarding development of the algorithm, see Stark and Drasgow (1998).

A computer program was developed to administer the paired-comparison CARS using the Visual Basic programming language and incorporating the algorithm developed by Stark and Drasgow (1998). The program first presents a dimension’s definitions and then, for each dimension, presents up to 15 pairs of statements, selected on the basis of estimates of the rate’s performance generated using the algorithm and choices for previous pairs. The program stops presenting pairs of items sooner when it runs out of pairs, which happens, especially, when the rate’s performance is either very high or very low. The intermediate performance estimates and the rate’s final performance score are determined by a Bayesian estimation process based on the effectiveness levels of all of the items selected as the more descriptive of the rate. Again, more detail on the algorithm appears in Stark and Drasgow (1998).

GRS. GRS for the three contextual performance dimensions were developed using a 7-point (1 = very ineffective, 7 = very effective) scale. Each contextual performance dimension was labeled and defined according to the three-dimensional content model.

BARS. The behavior scaling methodology developed by Smith and Kendall (1963) was used to construct three BARS for the contextual performance dimensions. Each contextual performance dimension had the same label and definition as the GRS. In addition, 5–6 behavioral examples from the pool of 124 were placed on each 7-point scale at the points reflecting their transformed mean retranslation effectiveness values.

**Procedure to Develop Videotaped Performances**

**Overview of procedure to develop videotaped performances.** First, scripts were developed that depicted office workers performing behaviors representative of each contextual performance dimension. Each script included personal support behaviors, organizational support behaviors, and conscientious initiative behaviors. The contextual performance behaviors

<table>
<thead>
<tr>
<th>Behavioral statement</th>
<th>Dimension</th>
<th>Effectiveness level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refuses to take the time to help others when they ask for assistance with work-related problems.</td>
<td>Personal support</td>
<td>Very ineffective</td>
</tr>
<tr>
<td>Expresses own personal satisfaction in being a member of the organization when asked by outsiders.</td>
<td>Organizational support</td>
<td>Effective</td>
</tr>
<tr>
<td>Consistently works on time or ahead of time, even when deadlines seem impossibly short.</td>
<td>Conscientious initiative</td>
<td>Very effective</td>
</tr>
</tbody>
</table>
included in each script were chosen to reflect realistic preset levels of effectiveness. The levels of effectiveness were taken from a "true score" matrix derived from estimated means, standard deviations, and intercorrelations among the three dimensions. The mean and standard deviation estimates for each contextual performance dimension were found in archival data, and the intercorrelations between the dimensions were taken from a recent meta-analysis (Viswesvaran, Schmidt, & Ones. 2001). These scripts were used to produce videotaped performances of approximately 5 min in length. Experts rated the effectiveness of the contextual performance behaviors exhibited by a particular actor on each videotape. The reliability of the expert ratings was examined, and consensus ratings were used as the true or target contextual performance scores for actors performing in this office job.

Generating intended true scores for contextual performance behaviors. Intended true scores were generated to facilitate the development of realistic videotaped rates. Means, standard deviations, and intercorrelations were needed for the three performance dimensions—personal support, organizational support, and conscientious initiative—to estimate realistic true score performance profiles across the three dimensions. Each score estimate the means and standard deviations, prior research (e.g., published literature, technical reports) on job performance ratings was reviewed, and for dimensions that matched reasonably closely with one of the three contextual dimensions, the means and standard deviations of the ratings were noted. The final mean and standard deviation estimates were simply the average of the empirically derived data from the studies reviewed. The means for the three dimensions were 4.71, 4.48, and 4.46, respectively, with the standard deviations ranging from 1.09 to 1.18.

Intercorrelations among the contextual performance dimensions were obtained from a meta-analysis conducted by Viswesvaran et al. (2001). This meta-analysis included nine performance dimensions, three of which mapped closely to the three contextual performance dimensions included in the CARS. Viswesvaran et al. provided meta-analytic estimates of the correlations between these dimensions, after halo and random error were removed; the correlations were .57, .49, and .60, respectively, between the first and second, first and third, and second and third dimensions described above.

These means, standard deviations, and intercorrelations were used to develop performance profiles for each of six target actors. Each profile consisted of three scores; each score represented a specific effectiveness level for one of the three contextual performance dimensions (personal support, organizational support, or conscientious initiative). The goal was to create six profiles of varying effectiveness levels that, taken together, mirrored the psychometric properties (i.e., means, standard deviations, and intercorrelations) of the contextual performance dimensions found in the literature. The final six performance profiles reproduced closely the desired covariance matrix.

Developing scripts and videotapes for each target person. These six profiles were used to produce scripts for each of the target persons. Each script was taken to ensure that the scripts did not contain the exact behaviors that were described on the BARS or in the CARS behavioral statements. This caution was necessary so that raters could not simply "track" the behaviors from the videotapes onto the BARS or CARS but, instead, would be required to evaluate the target person's behavior.

Each script was subjected to two iterations of initial subject-matter expert ratings and revisions. The experts for the initial script development step were two advanced I/O psychology graduate students. These ratings were used to guide script development to get the videotaped performances as close as possible to the intended profiles. After the two iterations of revisions, expert ratings closely matched the intended effectiveness levels. This process yielded the final scripts to be used by the actors in the videotaped performances.

Nine male actors were used to develop the videotapes. Six actors were used to represent each of the six target persons; three additional actors played coworkers and supervisors interacting with the target actors. All actors were instructed to conform to the scripted levels of performance so that the displayed behaviors would reflect the intended true scores as closely as possible. The actors were given several days to learn their scripts. The edited videotapes were used to revise the original scripts so that the final scripts reflected the actual dialogue on the videotapes.

Determining final target scores for the target persons. To determine the final target scores, three expert judges (two PhD I/O psychologists and an advanced I/O graduate student) familiarized themselves with the revised videotape scripts for each target person, viewed a target person's videotaped performance, and rated that target person's performance on each dimension before proceeding to the next videotape. The judges followed a systematic process in rating each target person. First, judges read each script to identify all contextual performance behaviors and categorized these behaviors into one of the three contextual performance dimensions. Next, they viewed the videotape for that script to identify visual nuances in contextual performance behaviors that may not have been reflected in the script's text. Judges used the 124 CARS contextual performance behavior statements and their corresponding effectiveness values as a guide to rate independently the effectiveness of the target person's performance on each dimension. The judges then discussed their ratings and reached a consensus in each case.

Interrater agreement between the initial expert ratings (i.e., before discussion) of each contextual performance dimension was assessed using Shrout and Fleiss's (1979) intraclass correlations (Case 3). The resulting intraclass correlations were .97, .96, and .97, respectively, for the three dimensions. On the basis of similar previous research, we expected that there would be high agreement between the expert judges (Borman, 1977, 1979). This was, indeed, the case; as such, the consensus expert ratings were used as the target score estimates against which accuracy and validity were evaluated. These final performance profiles were very similar to the intended performance scores.

Sample and Procedure

There were 114 participants in the format comparison study. Data from 2 of the participants were not retrievable because of faulty computer diskettes, reducing the sample size to 112. Thirty-three participants were employees from three local companies, and 79 participants were university students. All participants had work experience. Of these participants, 64% had supervisory experience, and 56% had experience conducting performance appraisals.

Participants were assigned to rate contextual performance using the CARS and either the BARS or the GRS. Fifty-seven participants used the CARS and the BARS, whereas 55 participants used the CARS and GRS formats. This design was used to examine possible reasons for any differences found in the psychometric properties between the various rating formats. That is, when the CARS and the GRS are compared, differences in psychometric properties might be attributed to the CARS behavioral scale format or the GRS numerically anchored scale format. Differences between the CARS and the BARS might be attributed to the adaptive component of the CARS. Finally, to minimize the impact of carryover effects on study results, we counterbalanced the groups so that about half used the CARS to make their first set of ratings (n = 28 for the BARS group and n = 29 for the GRS group) and about half used the CARS for their second set of ratings (n = 29 and 26, respectively).

Participants were given a basic orientation to the experiment and instructions on how to use the rating scales. They did not receive any formal rater error training. Participants viewed each videotaped vignette and rated the target person's performance immediately following the videotape. To increase raters' motivation, it was announced that participants with the most accurate ratings would be eligible for a monetary award ($50).
Results

Reliability

Agreement among raters was assessed using Shrout and Fleiss’s (1979) intraclass correlations (Case 2, 1-rater). For each rating format, intraclass correlations were computed for each dimension. Dimension intraclass correlations were transformed to z scores, averaged, and transformed back into correlations to obtain an overall reliability coefficient for each rating format. Table 3 presents the reliability coefficients obtained for each of the two paper-and-pencil rating formats and the corresponding CARS ratings. Recall that the sample was split into two groups: those who used the GRS and CARS to make their performance ratings and those who used the BARS and CARS to make their performance ratings. Accordingly, independent analyses were performed for each group.

Because there is no available test to assess the significance of differences between intraclass correlations, we compared the formats in terms of the improvement in precision. In particular, reliabilities and standard deviations for dimension ratings were used to calculate the standard error of measurement for each dimension for each rating format. Then, the percent reduction in the standard error of measurement using the CARS as opposed to each paper-and-pencil format was obtained by dividing the paper-and-pencil’s standard error of measurement into the corresponding CARS standard error of measurement and subtracting this amount from 1.0. Table 4 presents this information. From a standard-error-of-measurement perspective, the CARS provided substantially more precise ratings. CARS ratings had between 23% and 37% lower standard errors than the corresponding ratings on the other two formats.

Validity

Rating validity was assessed using Borman’s (1977) validity measure; it was computed for each rater. It simply measures the accuracy of raters’ rank orderings of raters for each dimension by correlating, across raters, their ratings with the raters’ corresponding final target scores. An overall validity score was computed by averaging the dimension-level validity scores for each rater and then averaging these overall scores across raters.

The resulting mean validities were .84 and .49, respectively, for the CARS–BARS comparison (pDifference < .001) and .82 and .72, respectively, for the CARS–GRS comparison (pDifference < .005). Effect sizes were estimated using Rosenthal’s (1987) t-to-r transformation, to determine the strength of the relationship between the rating formats and validity. Resulting correlations squared (r^2) were .28 for the CARS–BARS comparison, .09 for the CARS–GRS comparison, and .18 for the GRS versus both BARS and GRS. Interestingly, the overall effect (r^2 = .18) was about 4½ times larger than Landy and Farr’s (1980) lower bound estimate for format effect.

Accuracy

Rating accuracy was measured by comparing each rater’s ratings with the final target score estimates. Scores on evaluation, differential elevation, stereotype accuracy, and differential accuracy were computed for each rater by using Cronbach’s (1955) formulas. For each of these terms, smaller scores indicate greater accuracy. The means and standard deviations for each accuracy component are listed in Table 5 by rating format. Dependent r tests were used to compare each CARS accuracy component with its corresponding paper-and-pencil accuracy component.

CARS ratings were, indeed, significantly more accurate than ratings made using BARS on all four accuracy components, and CARS ratings were significantly more accurate than GRS ratings on three of the four components (all except differential accuracy). When CARS ratings were compared with the paper-and-pencil ratings (BARS and GRS combined), CARS ratings were more accurate on all four components.

Effect size estimates were calculated for these accuracy measures. Correlation estimates were computed and squared for each accuracy component for each group: CARS versus GRS, CARS versus BARS, and CARS versus paper-and-pencil. Table 6 presents these estimates. The effect sizes were not large (Mdn = .08), but, again, this result provides a somewhat more optimistic estimate of the influence of format on the quality of ratings than that provided by Landy and Farr (1980).

Discussion

In previous research, rating formats seemed to have had minimal effects on rater errors or rating accuracy (Bernardin, 1977; Landy & Farr, 1980; Taylor, Parker, & Ford, 1959). The CARS format, however, resulted in somewhat more reliable ratings, especially when the standard error of measurement was used as an index, and substantially greater validity and accuracy than GRS or BARS. Of course, these findings come from a lab study with videotaped rates. However, other lab studies using similar methodology (e.g., Borman, 1979) have shown no differences between rating formats in terms of accuracy. Thus, it seems there is some feature of the CARS format that creates more reliable and valid differentiation.

Table 3

*Intraclass Correlations (Case 2, 1-rater) of Ratings*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>GRS</th>
<th>CARS</th>
<th>BARS</th>
<th>CARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal support</td>
<td>.74</td>
<td>.82</td>
<td>.78</td>
<td>.84</td>
</tr>
<tr>
<td>Organizational</td>
<td>.72</td>
<td>.77</td>
<td>.74</td>
<td>.73</td>
</tr>
<tr>
<td>Conscientious</td>
<td>.72</td>
<td>.72</td>
<td>.71</td>
<td>.76</td>
</tr>
<tr>
<td>Initiative</td>
<td>.73</td>
<td>.78</td>
<td>.74</td>
<td>.78</td>
</tr>
</tbody>
</table>

Note. GRS = graphic rating scales; CARS = computerized adaptive rating scales; BARS = behaviorally anchored rating scales.

Table 4

*Percentage of Reduction in Standard Error of Measurement for CARS Versus BARS or GRS*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>CARS–GRS</th>
<th>CARS–BARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal support</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>Organizational</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>Conscientious</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>Initiative</td>
<td>33</td>
<td>26</td>
</tr>
</tbody>
</table>

Note. CARS = computerized adaptive rating scales; BARS = behaviorally anchored rating scales; GRS = graphic rating scales.
between rates. We now explore possible reasons for the promising results for CARS.

First, it might be argued that the CARS simply include more behavioral information on the scale, and this information helps raters to more precisely pinpoint rates' performance. This explanation suggests, for example, that if a BARS format included many more behavioral anchors, the reliability, validity, and accuracy of the BARS might be improved. And yet, the BSS format (Borman, 1979), in a sense, did just that by including in the behavioral summary anchors all of the behavioral information from a large number of behavioral statements. Results of that lab study showed no better results for the BSS than a BARS or GRS format.

In a related manner, it is possible that simply having more items (i.e., behavioral statement pairs) for CARS provides the boost in reliability and validity for this format. However, other multiple-item scales such as the summed scale (Campbell, Dunnette, Arvey, & Hellervik, 1973) have not yielded higher interrater reliability. For example, in Borman's (1979) study, no better interrater reliability or validity was found for a summed scale than for three other rating formats.

A second feature of CARS that differentiates it from BARS and that may provide a reason for these results is that CARS present more behavioral statements and are more targeted directly toward the rater, whereas BARS present the same rating anchors to each rater, regardless of the ratee being evaluated. The CARS present up to 15 pairs of behavioral statements for each rating dimension, and these statements are selected on the basis of a continuously updated estimate of the ratee's standing on the dimension being rated. Thus, an alternative explanation of the effectiveness of the CARS is that the adaptive nature of the format results in more precise ratings.

A third explanation for this study's results may involve the paired-comparison format itself. The cognitive operation of comparing observed behavior with two behavioral statements and the forced choice that is required may be a comparatively effective way to generate evaluative judgments. What might be useful in considering this possible explanation is to revisit Thurstone's (1927) law of comparative judgment. Recall that Thurstone's research was one of the cornerstones of Zimne and Griggs's (1974) model, on which the CARS are based. Thus, the CARS can be viewed as converting raters' judgments of ratee effectiveness to an interval scale on the basis of the same principles. Essentially, the choice of one stimulus over the other provides information concerning the placement of each stimulus (i.e., the ratee) on the underlying dimension at the interval scale level.

Because they do not use paired-comparison judgments, BARS and GRS are unlikely to provide this type of interval-scaling information. The scaling provided by CARS can be seen as more accurately reflecting the underlying measurement scale that raters are using to make their judgments. This scaling, in turn, may be what leads to more reliable, valid, and accurate ratings. Of course, the BARS and GRS do not prevent raters from providing interval-level information, but this information will be at the interval level only if people are able to provide ratings that directly reflect the true underlying scale of the attribute being rated. CARS do not assume that people are able to make ratings on a true interval scale. Rather, they use a scaling model to transform rater responses and place them on the appropriate interval scale.

Still another possible explanation for the results is that the CARS paired-comparison format focuses totally on behavior, with no consideration of numbers relating to effectiveness levels. In fact, raters are never even shown the scale points, so they must rely solely on behavior comparisons to make their judgments. This is not to say that raters using CARS cannot recognize that they are evaluating performance at a particular effectiveness level (high, medium, low, etc.). However, rather than having raters focus on a
more visual high–low global assessment required especially by the GRS, but also to some extent by the BARS, raters are asked to perform the forced-choice task.

If the no-numbers feature of CARS is important for explaining the format's psychometric superiority, we might expect the halo effect to be lower for CARS in that raters have reduced feedback on a ratee's evaluated performance levels on other dimensions when making paired-comparison judgments relative to a target dimension. In other words, with reduced direct knowledge of the ratee's evaluated performance levels on other dimensions, raters may be more likely to focus on a dimension's behavioral statements independent of their performance judgments made on other dimensions.

Accordingly, we computed a correlational index of the halo effect for each rater in the study on each of the two formats that he or she used by averaging the correlations between the ratings on the three dimensions for each format (N = 6 for each correlation: i.e., the number of videotaped performances). Mean correlations for each format were then computed, and the results for CARS, BARS, and GRS were .85, .65, and .84, respectively. The CARS–BARS difference was significant (p < .01), whereas the CARS–GRS difference was not. Thus, evidence was negative from this comparison regarding the no-numbers explanation.

Also relevant to the no-numbers explanation for the psychometric superiority of CARS, the MSS (Blanz & Ghiselli, 1972), like CARS, also asks raters to compare ratee behavior with behavioral statements that are not attached to a numerical rating scale. This format requires judgments about how observed ratee behavior compares with above average, average, and below average behavioral statements in terms of effectiveness. A format comparison study of the MSS and BARS (Dickinson & Zellinger, 1980) indicated about the same level of discriminant validity for the two types of scales. Thus, indirectly at least, this study failed to support the no-numbers explanation.

A possible reason why CARS was found to be psychometrically superior to BARS, whereas the MSS format was not, may lie in the type of behavior-comparison information derived from the two formats. With the MSS, the behavioral statements are at only three levels of effectiveness—above average, average, and below average; the CARS statements span the entire range of effectiveness. Thus, raters using the MSS do not have the opportunity to make "better than" or "worse than" judgments across as wide a variety of effectiveness levels as is the case with CARS. In addition, the paired-comparison format of CARS may provide more precise data than the MSS format, for which raters are asked to make a judgment about a single behavioral statement.

Further clues about why the CARS results are so favorable may be derived from reviewing the specific reliability, standard error of measurement, and accuracy component findings. The interrater reliability results are only slightly better for the CARS than for the other two formats, whereas the standard error findings strongly favor the CARS. Apparently, the agreement between raters when using CARS comes not so much from their agreeing on the ratee's rank order of effectiveness as from raters' convergence on the effectiveness level of each ratee on each dimension. This interpretation receives support from the result that elevation accuracy for CARS has a larger effect size than does differential accuracy. The level of the CARS ratings may be more accurately estimated than the rank order of ratees. Of course, the positive validity and differential elevation results for CARS run somewhat counter to this interpretation. On balance, the most likely explanations for the favorable results regarding CARS are (a) the CARS format presents behavioral statements more and more targeted directly toward the ratee (i.e., the adaptive feature of CARS) and (b) the iterative paired-comparison judgment task places ratee performance on an interval scale more so than the judgment tasks associated with other formats.

Our purpose in this article was focused primarily on basic research to explore if a very different rating format based on IRT principles might provide more reliable, valid, and accurate ratings of job performance. However, the question that obviously arises, from an applied perspective, is how practical is this format for use in organizations' performance appraisal systems? Also, how does CARS compare with the other formats regarding user acceptability? First, the content domain covered by CARS is currently restricted to contextual performance. The task performance or technical proficiency elements of job performance are not yet included. Second, the item bank development was quite labor-intensive. Preparing a new item bank for each study or application would be substantially more time-consuming than for any other rating format, including the behavior-based formats. Of course, for the contextual performance domain, the current CARS should require little or no scale development work. For the task performance domain, it may be difficult, but perhaps not impossible, to develop behavioral items that are relevant to a wide variety of jobs. Finally, the rating task itself takes longer than when each dimension is represented by a single scale. We estimate that raters in this lab study took two to three times longer to use CARS than BARS or GRS.

Regarding user acceptability, a survey was administered to evaluate participants' perceptions of the CARS and the other format assigned to them (i.e., GRS or BARS). Results of this survey were quite favorable for CARS. Participants preferred the CARS (64% to 28% for GRS, 8% no preference; 47% to 38% for BARS, 15% no preference). Also, users thought CARS were more accurate (64% to 21% for GRS, 54% to 26% for BARS), helped focus on relevant performance information (66% to 19% for GRS, 50% to 24% for BARS), and helped differentiate between good and poor performers (54% to 38% for GRS, 47% to 35% for BARS). The only comparatively poor reaction to CARS was with the item "are easy to understand" (50% in the top two categories—"a very great extent" or "to a great extent"—for CARS, 77% for GRS, and 60% for BARS).

An additional limitation of this research is that the format comparison was done in a laboratory setting. It will be important to compare the CARS and other rating formats in a field study, especially with respect to interrater reliability and construct validity.

Despite certain limitations, the study demonstrates that it may, in fact, be premature to abandon all research on rating formats or that we may want to conceptualize format more broadly to consider measurement models and raters' cognitive processes. Variance accounted for by format within this research was estimated to be 18% with rating validity as an index and 8% when rating accuracy was evaluated. The goal of designing a rating measurement method to more effectively align with raters' cognitive processes or to calibrate raters' perceptions of ratee job perfor-
mance to help them make more precise judgments veridical with actual ratee performance still seems worth the effort.

In particular, the CARS method appears to be a viable alternative to formats proposed previously. The favorable reliability, validity, and accuracy findings suggest that it may be an effective method for at least some applications.

References


Received April 28, 2000
Revision received December 1, 2000
Accepted December 1, 2000