Can Performance-Feedback Accuracy Be Improved? Effects of Rater Priming and Rating-Scale Format on Rating Accuracy

R. Blake Jelley and Richard D. Goffin
The University of Western Ontario

Performance appraisal information is often used for employee feedback and development. Research has found that assessments that are global (i.e., based on broad aspects of performance) and comparative (i.e., explicit interratee comparisons) may be most accurate in terms of Cronbach's (1955) differential accuracy, a type of accuracy that is directly relevant to the provision of feedback. Unfortunately, a global–comparative assessment may not give recipients the most useful diagnostic feedback. In this experiment, an innovative rater-priming manipulation was developed and tested on a sample of 109 participants. The priming manipulation had the effect of improving differential accuracy and providing diagnostic feedback. A 2nd independent variable involving 2 different Behavioral Observation Scale formats also was investigated. Explanations of findings, limitations of this experiment, directions for future research, and implications for performance appraisal practice are discussed.

This work focused on one of the most frequent uses of performance appraisal information, that is, employee feedback (Cleveland, Murphy, & Williams, 1989). Our goal was to develop and assess manipulations that would improve the accuracy of performance appraisal with respect to the identification of individual ratees' patterns of strengths and weaknesses.

To date, research on performance feedback has primarily investigated individuals' motivation to seek feedback, preference for positive over negative feedback, and the impact of feedback on future performance (Murphy & Cleveland, 1995). As summarized by Murphy and Cleveland (1995), "relatively little is known about how the design of the appraisal system might affect the frequency or quality [italics added] of feedback" (p. 91). It is assumed here that performance appraisal accuracy is an important, albeit insufficient, condition for feedback to positively affect future performance. In this experiment, methods to enhance the accuracy of performance-feedback information have been developed on the basis of results of recent research, and these new manipulations were tested to determine their effects on accuracy.

Given the need for research on the quality of performance appraisal feedback, alternative dependent variables for this type of research are considered. From this foundation, findings and implications of research relevant to the enhancement of feedback accuracy are discussed. Finally, the present experiment, which sought to contribute to the enhancement of feedback accuracy, is described.

Evaluation of Performance Appraisals

Distributional rater errors such as leniency, halo, and central tendency–range restriction have been used in the accumulation of a large body of research over many years but have recently been criticized as being poor criteria for performance appraisal research. Collectively, Cardy and Dobbins (1994), Murphy and Cleveland (1995), and others have presented compelling evidence of fundamental problems with the use of these common rater error measures in performance appraisal research. We are in agreement with these authors' main concerns; thus, as described below, rater accuracy measures constituted our dependent variables rather than rater error measures.

To conduct accuracy research, one must derive true score estimates that are acceptable standards of comparison (Sulsky & Balzer, 1988). That is, the true score estimates serve as the standards against which observed ratings are judged. Larger deviations between observed ratings and true score estimates are attributed to less accurate observed ratings. True score estimates are often based on the average ratings of people with target job expertise and enhanced opportunity to make informed ratings. Accuracy refers to both covariation with and distance between the scores provided by a rater for a given set of ratees and the corresponding true score estimates for that set of ratees (Sulsky & Balzer, 1988). The choice of which accuracy measure to use is important because the findings from any performance appraisal research are likely to be strongly influenced by this choice (Sulsky & Balzer, 1988). In a classic article, Cronbach (1955) demonstrated that four independent components of rating accuracy can be derived (Cardy & Dobbins, 1994; Cronbach, 1955; Sulsky & Balzer, 1988). We
concur with Cardy and Dobbins (1994) that of all accuracy measures "the Cronbach components should be accorded special status" (p. 39) because of their conceptual and mathematical independence. Accordingly, this experiment used Cronbach's four components of rating accuracy as dependent measures.

The first of Cronbach's (1955) components of rating accuracy, elevation (EL), is the differential grand mean. Inaccuracy with respect to this component reflects a rater's tendency to rate either too high or too low, averaged across all ratees and items (e.g., Murphy, Martin, & Garcia, 1982), relative to the true score estimates. Differential elevation (DE) is the differential main effect of ratees. This reflects the rater's accuracy in differentiating among ratees, averaging across all items and controlling for the rater's level of EL. DE is particularly relevant to administrative decisions because it evaluates a rater's accuracy in distinguishing between employees on the basis of their total job performance scores (Murphy & Cleveland, 1995). Stereotype accuracy (SA) is the differential main effect of rating scale items (or dimensions), averaging across rates and controlling for the rater's level of EL. This index indicates the accuracy with which a group of ratees' average performance on different items is differentiated by the rater. For example, this component reflects the rater's accuracy at determining whether the group of ratees are better at one aspect of their jobs or another. In analysis of variance (ANOVA) terminology, the final component, differential accuracy (DA), is the differential Ratee X Item interaction. This component reflects the variance not accounted for by the rater's level of EL or by the main effects of rates or items. Differential accuracy is typically interpreted as accuracy in diagnosing strengths and weaknesses of individual ratees (Cardy & Dobbins, 1994; Murphy & Cleveland, 1995).

Bearing in mind that the current work related to the use of performance appraisal for feedback purposes, we further considered the relevance of each of the accuracy components. Given that the provision of high-quality feedback has much to do with accurately diagnosing individual patterns of strengths and weaknesses, DA was given highest priority in the current research. The SA component may also be relevant to the use of performance appraisal for feedback purposes in the case of group feedback or training needs assessment. Thus, DA appeared to be the most important outcome measure for individual-level feedback accuracy research, whereas SA appeared to be of secondary importance. However, the use of only two of Cronbach's (1955) four components does not completely explain the squared deviations of ratings from true score estimates and does not allow determination of the extent to which improvements in any one type of accuracy might be obtained at the expense of other types of accuracy. Accordingly, the prudent choice was to use all four of Cronbach's components in this work while placing the greatest emphasis on DA and SA.

Improving the Accuracy of Performance Appraisals

Researchers have looked primarily at two approaches to improve performance appraisal accuracy: rater training and scale development (Woehr & Huffcutt, 1994). Landy and Farr (1980) called for a moratorium on scale development because a large body of research had failed to suggest the superiority of any one method over others. However, at the time of the Landy and Farr review, research on scale development was done using rater error measures as the main criteria. Consequently, relatively little scale development research has been conducted in which rating accuracy measures have served as criteria, and this is one reason to lift the moratorium on this line of research (Cardy & Dobbins, 1994). The Landy and Farr (1980) review also marked the beginning of a shift in performance appraisal research toward raters' cognitive processes (DeNisi, 1996). Despite their call for a moratorium on scale development, Landy and Farr proposed in their model of the performance-rater process that the rating instrument is one of several factors (e.g., the purpose of rating, rater and ratee characteristics) that affect rater cognitive processes such as observation, storage, and retrieval. This suggestion foreshadowed later calls for an integration of rating format and cognitive research (e.g., Cardy & Dobbins, 1994; DeNisi, 1996; Murphy & Cleveland, 1995). "Ideally, formats should be configured so that the operations required of raters reflect natural cognitive processes leading to efficient and effective processing of performance information" (Borman, 1991, p. 289). However, none of the major rating-scale formats, such as the Behaviorally Anchored Rating Scale, Behavioral Observation Scale (BOS), or Mixed Standard Scale, were developed on the basis of cognitive models of the appraisal process, and with the exception of Feldman (1986), little work of this type has been conducted to date (Cardy & Dobbins, 1994). Therefore, in addition to the reason described above, another reason to study rating-scale formats is that they may affect the way in which raters mentally process performance information and could conceivably be designed to improve the quality of ratings by using what is known about human cognition (Borman, 1991; Cardy & Dobbins, 1994; DeNisi, 1996; Feldman, 1986; Murphy & Cleveland, 1995).

A third reason that research on rating-scale formats should not be abandoned is that the Landy and Farr (1980) review did not extend to relative or comparative rating approaches wherein employees are to be explicitly compared with one another for performance assessment (Cardy & Dobbins, 1994; Wagner & Goffin, 1997). Rather, the Landy and Farr review focused on absolute approaches, wherein each ratee is compared to some external standard (e.g., frequency of behavior) rather than to one another for performance assessment. Despite early pessimism regarding the viability of relative formats (e.g., Landy & Trumbo, 1980; see also Miner, 1988, for a list of criticisms), recent evidence suggests that such approaches to performance appraisal may present a promising alternative to the more commonly used absolute approaches in terms of criterion-related validity (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Nathan & Alexander, 1988) and rating accuracy (Wagner & Goffin, 1997).

Historically, most relative approaches suffered from serious flaws (e.g., rankings are restricted to an ordinal level of measurement and cannot be meaningfully compared across raters), and this has inhibited research in this area. Among the few innovative approaches to relative rating formats are Miner's (1988) technique and the Relative Percentile Method (RP; Goffin, Gellatly, et al., 1996; Wagner & Goffin, 1997). Miner (1988) has described the development and use of a "rated ranking technique" (p. 291) wherein raters initially rank their employees on one aspect (dimension) of their job performance. Each employee is then rated by means of an absolute procedure (i.e., a rater selects a label from outstanding to poor that best describes an individual's performance on a given dimension), and this rating has to be consistent...
with the initial rank order. Unfortunately, the rated ranking procedure did not demonstrate superiority over straight rankings in terms of reliability or construct validity (Wagner & Goffin, 1997).

A newer comparative approach is the RPM (Goffin, Gellatly, et al., 1996; Wagner & Goffin, 1997). Although the RPM is not yet widely used, it has been relied on to provide criterion data in several publications (e.g., Christiansen, Goffin, Johnston, & Rothstein, 1994; Gellatly, Paunonen, Meyer, Jackson, & Goffin, 1991; Goffin, Rothstein, & Johnston, 1996; Meyer, Paunonen, Gellatly, Goffin, & Jackson, 1989). As illustrated in Goffin, Gellatly, et al. (1996), when using the RPM to rate a given dimension of performance, a rater makes ratings for all of his or her ratees on the same form. A separate form is used for each performance dimension assessed. A percentile scale is used in the RPM, and raters are instructed that an average performer would receive a score of 50, meaning that approximately 50% of the population of people working in a given job in that organization would be less effective on the given performance dimension. Additional benchmarks are described for high and low levels of performance. Unlike ranking procedures, the RPM is not dependent on an ordinal level of measurement. The RPM imposes no particular distribution; raters are free to place ratees anywhere along the 0–100 continuum, except that no exact ties are allowed. The RPM has typically been applied to measure relatively broad dimensions of job performance rather than critical incidents. Goffin, Gellatly, et al. found that the RPM showed higher correlations with cognitive ability, personality, and vocational interest measures than did a popular absolute format, the BOS (Latham & Wexley, 1977), suggesting its greater construct validity. However, Goffin, Gellatly, et al. were not able to assess rating accuracy in their field study.

A laboratory investigation was conducted to directly assess rating accuracy of the comparative format underlying the RPM (Wagner & Goffin, 1997). With Cronbach's (1955) rating-accuracy components as their dependent variables, Wagner and Goffin (1997) evaluated separately the effects of rating method (relative vs. absolute) and item type (global vs. specific; described below). Wagner and Goffin found that the comparative rating format underlying the RPM was more accurate than an absolute performance appraisal method in terms of DA ($\tau^2 = .44$) and SA ($\tau^2 = .10$) but that there were no relative versus absolute differences in terms of EL and DE. Accordingly, Wagner and Goffin suggested that comparative performance appraisals may have the greatest advantage over absolute approaches for purposes such as training needs assessment and employee feedback.

Global Versus Specific Items: Accuracy and Feedback Utility

Herein, a specific rating format refers to one that uses important job behaviors as items, whereas a global rating format refers to a rating scale that uses relatively broad aspects, or dimensions, of performance as items. Another finding of Wagner and Goffin (1997) was that global items were more accurately rated than specific items with respect to DA and SA, whereas the converse was true for EL and DE. The authors therefore suggested that global items might be more accurate for employee and organizational development purposes, whereas specific items might be superior for between-person or administrative purposes. Moreover, the magnitude of the advantage for global over specific items on DA and SA was substantial, accounting for large proportions of variance ($\tau^2 = .93$ and .84, respectively). Unfortunately, global items tend not to be maximally diagnostic in terms of offering specific information about how to improve performance (Latham & Wexley, 1977).

Considering the preceding evidence collectively, the question emerges as to how to capitalize on the advantages that global approaches have with respect to DA and SA (as discussed earlier, these are the two types of accuracy most relevant to performance feedback) while still providing the greater diagnosticity associated with specific items? The paradox this presents is that to have maximally accurate (DA and SA) judgments, raters should use global items that lack the specific information seen as most useful for employee development (Latham & Wexley, 1977). One goal of the present experiment was to help resolve this paradox so as to facilitate more accurate judgments of ratee strengths and weaknesses on specific job behaviors that can be used for employee feedback. Additionally, we sought to take advantage of the potentially superior DA and SA of comparative formats, as reviewed earlier.

RPM Priming

As discussed earlier, knowledge of human cognition may be useful when designing rating scales and other interventions to enhance performance appraisal accuracy. DeNisi and Williams (1988) noted that although the major cognitive models of performance appraisal (e.g., DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; IJgen & Feldman, 1983) differed in terms of emphasis, they all shared common concerns about how raters acquire, store, recall, and combine performance information to make judgments. This experiment focused on raters' retrieval of information from memory to accurately recognize and rate critical incidents of ratee performance.

Long-term memory can be conceptualized as a network of related concepts (Lord & Maher, 1991). Within the long-term memory network, information is organized into bundles of more highly interconnected units, which help retention and retrieval (Feldman, 1986). Concepts and information in the memory network vary in terms of their levels of activation. Information with high activation can be retrieved more quickly and with greater accuracy than information with low activation (Lord & Maher, 1991). According to Feldman (1986), people's use of schematic (cognitive strategies based on domain knowledge) or categorical ("fuzzy sets" of like objects or concepts) information, both to help form networks of associations and to retrieve information, suggests that categorical representations may be more accessible for recall than detailed (specific) memories.

Conceptualizing memory as an associative network of related information has implications for the design of methods to help raters retrieve more specific information from memory. The retrieval of information is thought to occur through the parallel spread of activation from more highly activated to less highly activated concepts (or memory nodes). The spread of activation may occur because of conscious or subconscious processing, including cognitive priming (Lord & Maher, 1991). A prime is any stimulus that is presented before a target stimulus to affect a response to that target. Priming "causes an activation of both that concept and other concepts that are related to it" (Ashcraft, 1989,
p. 277). Therefore, one possible way to capitalize on the potentially greater accuracy of the global item type without sacrificing the diagnostic information associated with specific items is by priming raters with global ratings before they rate specific job behaviors. Moreover, by using a global and comparative format, the main effect of comparative over absolute ratings with respect to DA and SA (Wagner & Goffin, 1997) might further strengthen this manipulation. As it happens, the RPM format developed by Goffin, Gellatly, et al. (1996; discussed earlier) is just such a combination of global and comparative formats.

The process of having raters make RPM ratings and directing their attention to these ratings as benchmarks was designed to make these ratings more accessible, that is, more highly activated, in memory. It was assumed that global and comparative judgments about a ratee’s performance would be related to more specific memories of ratee behavior. Through spreading activation, these more specific memories would become more highly activated and more likely to be recognized correctly, resulting in better BOS ratings.

It could be argued that the use of relatively broad dimensions as primes might lead to internally generated associations and recall of false memories consistent with what would be expected of a category prototype or schema (DeNisi & Peters, 1996; DeNisi & Williams, 1988; Feldman, 1981, 1986; Igen & Feldman, 1983; Murphy & Balzer, 1986; Murphy et al., 1982). Nonetheless, as stated by Alba and Hasher (1983), “the stored record of any event is far more detailed than prototypical schema theories imply” (p. 225). According to schema theories, only a subset of ideas are selected for storage. Of the subset of schema-consistent information that is selected, few original details are represented in memory. Instead, schema theories assume that abstraction of the underlying semantic content or meaning occurs with respect to representation in memory. Incompleteness or distortion in recall is assumed to be largely a function of these selection and abstraction processes (see Alba & Hasher, 1983, for a more complete discussion). The process of retrieval from memory is thought to involve the generation of probable detail based on the few details of the memory episode selected for representation in memory, along with general knowledge and relevant schemas. However, reconstruction is a rare phenomenon occurring only in special circumstances (Alba & Hasher, 1983). The evidence reviewed by Alba and Hasher indicated that a great deal of original, memory-episode information remains available in memory, in contrast to the assumption that only schema-relevant information will be stored. Moreover, memory may not be as abstractive as schema theories would assume. Alba and Hasher reviewed evidence of memory for detail of both well-learned and relatively novel information. In contrast to the selection and abstraction assumptions of schema theories, a great deal of detailed information seems to be available in memory (i.e., represented in the memory system). However available information can vary in terms of accessibility (i.e., the ease with which it can be retrieved). Alba and Hasher noted that the retrieval situation and cues are important in determining which available information can be accessed.

Feldman (1986) drew on Alba and Hasher’s (1983) view of memory and posited that a more valid basis for impressions, in terms of the validity associated with the organizing schema and category system, would lead to more accurate behavioral recall. As discussed earlier, Wagner and Goffin (1997) reported DA and SA main effect advantages both for global items and for comparative items. The accuracy advantages reported by Wagner and Goffin could conceivably be attributable to global items and comparative items, each representing a relatively valid basis for impressions (vs. specific items and absolute items, respectively). Because the RPM typically uses global-comparative items, it may have an advantage, in terms of validity of impressions, over other global item types that do not simultaneously stimulate comparative impressions. Thus, the first independent variable in this study consisted of priming (vs. not priming) the raters by requiring them to rate performance using the RPM scale before making the more specific ratings, which are potentially more useful for feedback purposes. The nature of the specific ratings is described in the next section.

Rating-Scale Formats: Serial and Parallel

Notwithstanding Murphy and colleagues’ criticisms of the BOS (e.g., Murphy et al., 1982), in principle, the level of specificity present in the BOS has high potential for the provision of useful feedback (Latham & Wexley, 1977) and was therefore chosen for data collection purposes here. Traditionally, the BOS uses a serial-rating format; however, on the basis of the work of Goffin, Gellatly, et al. (1996) and Wagner and Goffin (1997), a new parallel BOS format that facilitates comparative judgments also was considered here. Comparative judgments may be encoded automatically during observation and could potentially serve as valid retrieval cues. Accordingly, the second independent variable in this experiment was the use of a serial versus parallel BOS for the rating of performance. Using the traditional serial BOS format, participants rated each ratee, one at a time, on all behavioral incidents, before moving on to rate the next ratee. The parallel BOS format required participants to rate all ratees simultaneously on each behavior before moving on to the next behavior.

Method

Participants

Undergraduate psychology students (N = 112; 69 women and 43 men) participated as raters in return for course credit or the experience of participating in applied psychological research. The participants ranged in age from 16 to 39 years (M = 20.5), and most were 1st-year university students (n = 90).

Targets

The target job, university lecturer, was chosen so as to be meaningful to the students serving as raters in this experiment. Four videotaped segments of lectures given by different male university instructors were used as the performance-rating targets. Each lecture segment was approximately 10 min in duration. This allowed for participants to view an appropriate sample of ratee job performance while also permitting multiple (four) ratees to be included in the 50-min observation session. The videotaped lectures were originally recorded for a study by Renaud (1996). The ratee vignettes were carefully chosen from a battery of 32 to be similar in interest of topic taught, ratee gender and race, and audiovisual quality while also showing variance in ratee performance. These tapes are of real instructors conducting actual classes rather than actors in contrived settings. The variance between the tapes, therefore, may not be as large as in scripted performances but may be more generalizable to actual work behavior.
Rates were identified by the letters A, B, C, and D to protect confidentiality.

**Job Behaviors Used in Performance Rating Scales**

Items for the two different versions of the BOS (serial and parallel) used the same behaviors drawn from the Teacher Behaviors Inventory (TBI, Diagnostic Version; see Murray, 1987, 1991). The job analysis used to generate the TBI items was conducted as a prerequisite to a program of research in higher education (Murray, 1997). Specifically, trained observers were sent to record the classroom behaviors of numerous instructors (Murray, 1997). Ratings of these items have been subjected to factor analysis, and considerable construct validity evidence exists for the TBI (Murray, 1991). For example, TBI scores have been shown to be related to various educational criterion measures such as student ratings, study hours, and exam performance. Moreover, Murray (1991) recognized the behavioral incidents contained in the TBI as applicable across a wide range of academic disciplines.

The TBI is multidimensional; however, the following three TBI dimensions were chosen a priori for use in this experiment because they largely comprised items that could reasonably be observed in the 10-min video segments: Clarity, Enthusiasm, and Speech. A further reason for choosing these dimensions was their centrality to the target job (Murray, 1991). A small number of items from the three chosen dimensions were excluded because of observability concerns with respect to the videotaped vignettes. The final set of TBI items used to calculate accuracy scores is given in the Appendix. An infrequency item, “Climbs on desks or other furniture,” was included for each target ratee (i.e., a given participant would rate this item once for each of the four ratees, amounting to four infrequency items) to detect possible nonpurposeful responding by the rater (Jackson, 1970). Elevated ratings on the infrequency items were taken as evidence that the participant might not have been rating in a careful and meaningful way because the respective behavior was known not to have occurred in our video segments.

**Manipulations and Design**

The participants, all of whom served as raters, were randomly assigned to one of four conditions in a 2 x 2 completely randomized factorial design. The two independent variables in this experiment were RPM priming (primed or not primed) and BOS format (serial or parallel). This was a between-subjects design with four dependent measures, that is, Cronbach’s (1955) four components of accuracy. Twenty-eight participants were assigned to each of the four cells.

**RPM priming.** The priming manipulation consisted of having participants make initial global-comparative ratings of performance for given dimensions using the RPM (Goffin, Gellatly, et al., 1996). After making these ratings for a given performance dimension (i.e., Clarity, Enthusiasm, or Speech), participants were then asked to make ratings on several specific job behaviors for each respective dimension. They were told to use their RPM ratings as benchmarks for making the specific behavior ratings, and they were encouraged to refer to their RPM ratings as often as they wished. Participants in the nonprimed condition simply completed the Clarity, Enthusiasm, and Speech specific ratings without the RPM priming.

**BOS formats: Serial and parallel.** The second independent variable was the type of rating form used for the ratings of specific job behaviors, serial or parallel. Both of these formats used identical frequency-of-behavior anchors; however, as explained earlier, the difference between serial and parallel processing refers to how the participants considered the targets. The serial format was designed to have participants rate each target on a number of behaviors (items) before rating the other targets. The parallel format was designed to have participants rate all targets simultaneously on each behavior (item) before moving on to rate the next behavior. Figure 1 shows two sample items presented in a serial format, and Figure 2 presents the same items in a parallel format.

**Development of True Score Estimates**

Ten graduate students and one industrial/organizational faculty member served as expert raters for the development of true scores. All experts were thoroughly familiar with the job of university lecturer, and none reported knowing any of the videotaped ratees. The procedure pioneered by Borman (1977) and used by Murphy et al. (1982) was used as a guideline to obtain true score estimates. Accordingly, expert raters were given the opportunity to view the videotaped targets as many times as desired to confidently make the ratings, were allowed to take notes, and were requested to make ratings immediately after watching the videotaped lectures rather than after a time delay. No expert reported watching the segments less than twice.

The rating format used by the expert raters in developing the true scores was thought to be a potential confound. That is, if the format used by the expert raters was more similar to that used by participants in some experimental conditions than the format used in other conditions, the possibility would arise that any observed differences in accuracy between conditions could be spurious. To reduce this possibility, experts were given small cue cards, each referring to one target and one job behavior. Experts were then free to rate the behaviors and targets in any order they desired. Thus, the rating format administered to the expert raters was neither parallel nor serial by design.

Expert raters made independent ratings using the same 5-point frequency scale as the participants. However, experts were also given the opportunity to view the videotaped targets as many times as desired to confidently make the ratings, were allowed to take notes, and were requested to make ratings immediately after watching the videotaped lectures rather than after a time delay. No expert reported watching the segments less than twice.

The rating format used by the expert raters in developing the true scores was thought to be a potential confound. That is, if the format used by the expert raters was more similar to that used by participants in some experimental conditions than the format used in other conditions, the possibility would arise that any observed differences in accuracy between conditions could be spurious. To reduce this possibility, experts were given small cue cards, each referring to one target and one job behavior. Experts were then free to rate the behaviors and targets in any order they desired. Thus, the rating format administered to the expert raters was neither parallel nor serial by design.

Expert raters made independent ratings using the same 5-point frequency scale as the participants. However, experts were also given the opportunity...
to indicate that they could not rate a particular behavior based on a given video segment. This allowed for identification of behaviors that were not sufficiently observable in the chosen 10-min vignettes. The mean of the experts’ ratings of each behavior for each target served as the true score estimate. One expert rating (one behavior with respect to one ratee) was omitted because it was an extreme outlier, suggesting that the rater had misread the item.

We considered the variability of each item for each target across the expert raters, as well as the number of experts who indicated an inability to provide a rating of the item, as evidence that certain items, in the context of these particular video segments, might not be suitable for obtaining high-quality true score estimates. Accordingly, 6 items were dropped from subsequent analyses. Infrequency items (described earlier) were also excluded from accuracy score computations. In total, 20 of the TBI items were used in the accuracy calculations (see the Appendix).

The average interexpert correlation across all retained items and targets was .77. This is well above the appropriate 90th percentile value reported in Viswesvaran, Ones, and Schmidt’s (1996) meta-analysis. We obtained an estimated reliability of the panel of 11 experts of .97, by means of the Spearman–Brown prophecy formula (Nunnally, 1978). The average interexpert standard deviation for all retained items was .63 on the basis of the 1-to-5 frequency rating scale.

**Procedure**

Participants were assembled in small groups, and each participant took part in two different sessions, that is, an observation session followed by a rating session. In the observation session, participants viewed the four videotaped vignettes described above. The four target rates were presented once only and in a constant order across conditions. As in past performance appraisal research (e.g., Murphy & Balzer, 1986; Wagner & Goffin, 1997), there was a delay between observation and rating that was designed to simulate the memory demands required of performance raters in applied settings. Accordingly, the rating session took place 48 hr after the observation session. During rating sessions, participants were subjected to the independent variables described above and were presented with labeled still photographs of each ratee as an aid in remembering which rate was referred to as A, B, C, or D.

Three experimenters, two males and one female, ran the rating sessions. Experimenters were provided with a standardized written script for administering the sessions. The primed and nonprimed conditions were run in separate rooms because the respective rating materials were noticeably different. Every effort was made to counterbalance experimenters across experimental conditions.

**Results**

**Preliminary Analyses**

Of the original 112 participants, 2 were omitted from the analyses because of apparent infrequent or nonpurposeful responding, as indicated by their ratings of the infrequency items described earlier. One other participant was excluded because of prior knowledge of at least one of the rates. Data from 109 participants (66 women, 43 men) were used in the analyses. Missing data constituted less than .01% of all items and were replaced by mean substitution.

Cronbach’s (1955) components of accuracy were computed, based on the 5-point frequency rating scale used for the BOS items, using formulas provided by Cardy and Dobbins (1994), Murphy et al. (1982), and Sulsky and Balzer (1988). These components represent deviations of participant ratings from the corresponding true scores; therefore, lower scores indicate greater accuracy. Descriptive statistics for each condition appear in Tables 1 and 2.

**Multivariate Analyses of Variance**

To assess the possibility of experimenter effects, a multivariate analysis of variance (MANOVA) was conducted, using the experimenter in the rating session as an independent variable and the four accuracy components as dependent variables. No significant difference as a function of experimenters was detected, Pillais’ trace = .06, F(8, 208) = 0.82, p > .05, η² = .03.

To test the main hypotheses, a completely randomized MANOVA with two independent variables (RPM priming vs. no priming and serial vs. parallel rating-scale format) and four dependent variables (EL, DE, SA, and DA) was conducted. The multivariate interaction was not significant, Pillais’ trace = .01,

<table>
<thead>
<tr>
<th>Accuracy component</th>
<th>No RPM priming</th>
<th>RPM priming</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Serial rating scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>.79</td>
<td>.11</td>
</tr>
<tr>
<td>DE</td>
<td>.36</td>
<td>.11</td>
</tr>
<tr>
<td>EL</td>
<td>.25</td>
<td>.11</td>
</tr>
<tr>
<td>SA</td>
<td>.59</td>
<td>.10</td>
</tr>
<tr>
<td>Parallel rating scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>.82</td>
<td>.12</td>
</tr>
<tr>
<td>DE</td>
<td>.47</td>
<td>.16</td>
</tr>
<tr>
<td>EL</td>
<td>.37</td>
<td>.27</td>
</tr>
<tr>
<td>SA</td>
<td>.73</td>
<td>.19</td>
</tr>
</tbody>
</table>

*Note.* Accuracy scores are based on a 5-point frequency rating scale; lower scores indicate greater accuracy. RPM = relative percentile method; DA = differential accuracy; DE = differential elevation; EL = elevation; SA = stereotype accuracy.
Table 2
Marginal Means and Standard Deviations of Accuracy Measures

<table>
<thead>
<tr>
<th>Condition</th>
<th>DA</th>
<th>DE</th>
<th>EL</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPM—priming</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>.81</td>
<td>.12</td>
<td>.41</td>
<td>.31</td>
</tr>
<tr>
<td>Yes</td>
<td>.76</td>
<td>.12</td>
<td>.48</td>
<td>.31</td>
</tr>
<tr>
<td>Rating-scale format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serial</td>
<td>.77</td>
<td>.11</td>
<td>.40</td>
<td>.26</td>
</tr>
<tr>
<td>Parallel</td>
<td>.79</td>
<td>.13</td>
<td>.49</td>
<td>.36</td>
</tr>
<tr>
<td>Accuracy grand means</td>
<td>.78</td>
<td>.12</td>
<td>.45</td>
<td>.31</td>
</tr>
</tbody>
</table>

Note. Accuracy scores are based on a 5-point frequency rating scale, and lower scores indicate greater accuracy. Means in the same column and condition with different subscripts differ significantly (Univariate F-test, p < .05). The difference between serial and parallel means with respect to SA is significant at p < .001. DA = differential accuracy; DE = differential elevation; EL = elevation; SA = stereotype accuracy; RPM = relative percentile method.

Univariate Analyses

To further investigate the manner in which the treatment factors affected the individual dependent variables, separate ANOVAs were conducted using EL, DE, SA, and DA. The significant MANOVAs for the main effects of the independent variables suggested that each univariate effect can be evaluated with less chance of Type I error inflation (Stevens, 1996). The low magnitude of the pooled within-cell correlations among the four dependent variables (r = .15) suggested that these univariate tests were not redundant.

The univariate results for the main effect of RPM priming revealed significant effects on both the DA, F(1, 105) = 5.56, p < .05, and DE, F(1, 105) = 4.21, p < .05, accuracy measures. Specifically, participants who received RPM priming made significantly more accurate ratings with respect to DA, but they were significantly less accurate with respect to DE.

The univariate results for the effect of rating-scale format (serial or parallel) on the accuracy components indicated that serial and parallel formats resulted in significantly different accuracy scores for EL, F(1, 105) = 5.98, p < .05, and SA, F(1, 105) = 7.08, p < .05, whereas no significant difference was found with respect to DA. In the case of all three significant differences, the traditional serial format was more accurately rated than was the parallel format.

Discussion

The primary goal of this experiment was to assess methods of improving performance ratings, methods that had the potential to provide performance appraisal feedback that was both specific and accurate. In this regard, all four of Cronbach’s (1955) accuracy components (EL, DE, SA and DA) were examined, with greatest attention given to DA, which has been interpreted as most relevant for assessing the accuracy with which a rater correctly identifies individual patterns of performance strengths and weaknesses (Cardy & Dobbins, 1994; Murphy & Cleveland, 1995). As discussed earlier, SA was second most important because accuracy with respect to this component may be relevant to the identification of the strengths and weaknesses of work groups. EL and DE were not as directly relevant to the use of performance appraisal for the provision of feedback but were included largely to assess the extent to which gains in some forms of accuracy might be obtained at the expense of other forms of accuracy.

One method that we derived for the enhancement of feedback accuracy consisted of priming the raters by requiring them to initially rate performance using a global-comparative rating method termed the RPM (Goffin, Gellatly, et al., 1996), which, on the basis of past research (Wagner & Goffin, 1997) and theory (Alba & Hasher, 1983; Feldman, 1986; Lord & Maher, 1991), seemed likely to enhance DA and SA. The RPM priming approach was compared with a no-priming condition. A second method that had potential for improving the accuracy of appraisals for feedback purposes was a parallel BOS format we derived that facilitated comparative judgments of specific job behaviors. Previous research (Goffin, Gellatly, et al., 1996; Nathan & Alexander, 1988; Wagner & Goffin, 1997) provided grounds for optimism that a comparative approach would be advantageous. The new parallel BOS format was compared with a traditional serial BOS format. As discussed below, partial support for these manipulations was found; however, some negative consequences for certain accuracy components also were detected. Generally, the significant effects in this experiment were associated with modest effect sizes (4%-6% of the variance), except for the effect of the serial versus parallel manipulation in terms of SA, which was larger (19% of the variance).

RPM Priming

From the perspective of improving the accuracy of performance appraisal feedback, this research was, indeed, successful in that the
use of RPM priming improved DA. Presumably, initial ratings for BOS items also served as primes for subsequent items, yet the use of global-comparative priming before rating any specific item is still enhanced DA. This finding is consistent with the view that a more valid basis for impressions can improve memories for specific behaviors and improve performance appraisals (Feldman, 1986). It is also consistent with the notion that activating global and comparative impressions through cognitive priming may cause a spread of activation in memory related, but less activated, memories of more specific behaviors that become more accessible after priming. However, this evidence is not by itself proof of these mechanisms because measures of the hypothesized process variables (e.g., response latencies) were not collected. Future researchers may wish to replicate the priming effect and simultaneously collect relevant measures of cognitive operations.

Although the use of RPM priming had intended effects on the DA component of accuracy, it resulted in less accurate ratings with respect to DE. Thus, RPM priming may be beneficial for improving the accuracy with which raters identify individual strengths and weaknesses when making specific job behavior ratings but may attenuate accuracy for between-person comparisons. The detrimental effect of RPM priming on DE is more difficult to explain solely from the perspective of network activation. If the RPM priming serves to activate related concepts (e.g., critical-incident memories) and thereby make them more salient, we would not necessarily expect worsened DE.

As one possible way to understand this accuracy trade-off, assume that when rating specific behavioral items, raters consider both the ratee’s standing on the underlying broad factor that the item reflects and the ratee’s standing on the specific behavior embodied in the item. It appears plausible that without the priming manipulation, raters were more likely to heavily weight the ratees’ standing on broad dimensions of performance when assigning scores to the specific behaviors. This explains why DE accuracy was better in the no-priming condition. However, the RPM priming manipulation might have improved DA by offering participants an opportunity to express their more global judgments apart from the specific behavioral ratings they made. Thereby, primed raters might not feel compelled to express underlying global judgments regarding the overall, or dimensional, standing of the ratees in each and every specific item. Rather, these results suggest that the primed participants appeared to be more likely to focus on the unique variance of each item. Alternatively, nonprimed raters may be less accurate with regard to DA because they have not had the opportunity to clearly express their global judgments regarding the performance of rates. Consequently, their ratings of specific behaviors are dominated by more global impressions of the ratees, and they fail to adequately consider the specifics of each behavioral item.

The literature on context effects in survey research (e.g., Schwarz, 1999; Schwarz & Sudman, 1992), especially the work on question-order effects (influence of a preceding question) and conversational norms, may be relevant to our RPM priming results. According to Strack (1992), a preceding item may have both activation and information functions for the interpretation of a subsequent item. The activation function is similar to what has already been presented in this article regarding the conceptualization of memory as a network of related concepts. The preceding item serves as a prime, which makes some concepts more accessible in memory, and available concepts may influence responses to a subsequent question.

However, respondents may not use information that is highly activated in memory (Schwarz, Strack, & Mai, 1991). The information function of a preceding question is that it serves as a basis from which respondents infer the questioner’s intended meaning and referent with regard to the subsequent question (Strack, 1992). Preceding questions also make it clear to the respondent what the questioner already knows (Strack & Martin, 1987) when the respondent has answered those questions. Some scholars (e.g., Schwarz et al., 1991; Strack, 1992) have suggested that implicit conversational norms governing cooperative conversational exchange (i.e., Gricean principles) have importance for survey methodology. “One of the principles that govern the conduct of conversation in everyday life (Grice, 1975) requests speakers to make their contribution as informative as is required for the purpose of the conversation, but not more informative than is required” (Schwarz et al., 1991, p. 5). This norm was highlighted by Schwarz and colleagues and is particularly relevant for our purposes. Specifically, respondents should be informative and avoid providing information that is redundant with what the questioner already knows (Schwarz et al., 1991; Strack, 1992). This is what psychologists refer to as the “given–new contract” (Schwarz et al., 1991, p. 5). This perspective is very useful for explaining why RPM-primed participants might have placed relatively greater emphasis on unique variance while nonprimed participants tended to focus on common variance or more general impressions. RPM-primed participants knew that they had communicated their overall judgment about the performance dimension in question. Therefore, communicating these more global judgments through the BOS ratings would have been redundant with information that had already been provided and would violate the given-new contract.

Similarly, the apparent tendency for nonprimed participants to focus more on common variance can also be interpreted from the perspective of cooperative communication. Performance appraisals are used for various purposes, including administrative decisions about pay raises and promotions (Cleveland et al., 1989). It is not surprising, therefore, that when asked to evaluate the performance of another person, especially when there are no disincentives for being honest, raters may place a premium on communicating their more global impressions because these may be seen as the most consequential. Thus, when BOS are the only means of communication available, raters may use them to communicate more global impressions. Empirical results leading to criticisms that memory-based BOS ratings are largely a function of general impressions and may not be measures of behaviors (Murphy et al., 1982) do not necessarily indicate that raters cannot be accurate at the level of critical incidents. Rather, it is possible that raters are, to some extent, able to recognize and differentiate the occurrence of various job behaviors but may be unwilling to act on this ability when they wish to communicate more general impressions. Rating-scale format research may be viewed from cognitive and communication perspectives that focus on how the rater processes information, what the rater is trying to communicate, and how formats can be optimally configured, with these things in mind, to obtain the highest quality information. The literature on context effects in survey research and normative conversational principles provides one possible explanation for our finding that there was a trade-off between DA and DE for the RPM priming manipulation.
Future research should more directly test its adequacy and compare it with alternative explanations.

Attempts to improve the accuracy of ratings for one purpose (i.e., the enhancement of DA for feedback) through the priming manipulation had negative consequences for an accuracy component relevant to other purposes (i.e., DE, which is relevant to administrative decisions). This provides support for the assertion that in practice, it may not be wise to use a single appraisal system for all purposes (Murphy & Cleveland, 1995). Thus, a priming manipulation such as the present one would be well placed as part of an appraisal system being used for feedback purposes (i.e., intrapersonal comparisons), whereas a nonprimed rating procedure may be more beneficial for administrative decisions such as promotion or pay raises, where interindividual accuracy is at a premium. Alternatively, the RPM ratings that were used only as a priming manipulation in this study may themselves be used for organizational purposes involving intrapersonal comparisons. The accuracy of these ratings was not currently assessed; however, previous research attests to the overall worthiness of this approach (Goffin, Gellatly, et al., 1996).

Rating-Scale Format

Although our priming manipulation served its intended purpose, the new parallel BOS rating-scale format developed here did not prove successful. The traditional serial BOS and our parallel BOS did not significantly differ with respect to DA; however, the serial BOS outperformed the parallel BOS on all other accuracy components (EL, DE, and SA). One methodological explanation for these findings is that our attempt to make the rating-scale format that the expert raters used in the development of the true scores neutral was not successful. Recall that the BOS rating format supplied to the expert raters was neither serial nor parallel; rather, expectations were given separate cue cards, each referring to one behavioral incident for a single ratee. Nonetheless, informal follow-up questions revealed that most experts used a strategy more consistent with the serial BOS format than with the parallel format. No expert reported using a strategy that could be considered exclusively parallel. Therefore, the finding that participants using the serial format were more accurate may at least partially be attributed to the expert raters’ general tendency to organize their own rating scales in this way.

The experts’ frequent use of a serial strategy is consistent with prior performance appraisal research. DeNisi, Robbins, and Cafferty (1989) and DeNisi and Peters (1996) found that when given a choice as to how to structure performance diaries, raters preferred to use a person-blocked (serial) structure and tended to use this structure to organize information in memory. Moreover, using a variety of dependent measures in both lab and field settings, DeNisi and colleagues (DeNisi & Peters, 1996; DeNisi et al., 1989; Williams, Cafferty, & DeNisi, 1990) reported that inducing the organization of information in memory according to ratee generally resulted in higher quality ratings. However, DeNisi (1996) cautioned that this does not necessarily imply that such organization will always be optimal.

It is not clear to what extent the observed accuracy advantage for participants using the serial-rating-scale format was a function of the possible methodological confound discussed previously versus a true advantage for this format over the parallel format. The same literature on memory organization that explains the experts’ greater use of serial processing, and hence the methodological confound, is also consistent with the possible explanation that the serial format is rated more accurately because it reflects a natural organization of information in memory used by raters when conducting performance appraisals.

Another possible explanation is that participants assigned to parallel-rating-format conditions had to consider how the raters compared with one another in addition to considering how each fared in terms of the frequency-of-occurrence rating scale inherent in the BOS format. In contrast, participants using the serial format did not have the explicit interperson comparisons to contend with. Perhaps the parallel format in combination with a highly specific item format (i.e., BOS) places too many simultaneous cognitive demands on raters. In the face of such demands, raters may tend to vacillate between placing a premium on intrapersonal accuracy or interitem accuracy, depending on the particular item. The choice between these two alternative priorities might lead the rater to follow the path of least resistance (i.e., whichever priority engenders the least amount of effort), but this may change depending on the item. Such item-to-item vacillation in rating priorities could contribute to the generally lower accuracy of the parallel BOS format. Comparative approaches may prove difficult for raters to use at the level of specific job behavior ratings.

Generalizability

Realism is a common concern in performance appraisal research that uses accuracy as the dependent measure because such research requires the development of true scores and true scores, by and large, are only developed in laboratory environments. It has been suggested that it may be possible to generate true scores suitable for use in the field (Murphy & Cleveland, 1995), but this would ordinarily require the use of averaged ratings to obtain true score estimates, and such a procedure may not be optimal (Sulsky & Balzer, 1988). Although attempts were made to enhance the generalizability of the present research, by simulating memory demands and using a target job that is meaningful to the raters, several elements of applied settings (see Bernardin & Villanova, 1986; Feldman, 1981) were not simulated. Key features of applied settings that were not replicated in this experiment include the various sociopolitical aspects of applied performance appraisal. Murphy and Cleveland (1995) distinguished between performance judgments, which are private evaluations of ratees’ performance, and performance ratings, which are public statements about ratees’ performance. The latter are more heavily influenced by the rating context, including the social and political environment. The present research thus generalizes more readily to performance judgments than to ratings. The impact of performance appraisal context on the currently studied manipulations and the influence of more accurate feedback on subsequent ratee performance need to be evaluated in future research.

---

1 We would like to thank an anonymous reviewer for suggesting that our findings on structured ratings could be discussed in terms of the organization of information in memory.
Conclusions

1. Priming raters using an RPM (Goffin, Gellatly et al., 1996) format gave rise to a significant improvement in the type of rating accuracy that is most relevant to the provision of employee feedback.

2. The enhancement of accuracy for employee-feedback purposes (i.e., intraindividual comparisons) appears to be obtained at the expense of reducing accuracy relevant to interindividual comparisons. Consequently, previous suggestions that a single performance appraisal rating system not be used for all organizational purposes (e.g., Cleveland et al., 1989; Murphy & Cleveland, 1995) were supported.

3. The literature on context effects in survey research (see Schwarz, 1999; Schwarz & Sudman, 1992) is helpful in understanding why RPM-primed participants may be more accurate on DA but less accurate on DE than are nonprimed participants. Conversational norms may be helpful for understanding rating behavior and for configuring rating-scale formats to elicit from raters the most relevant, and highest quality, information for a given purpose.

4. A new parallel BOS rating format appears to be inferior to the traditional serial BOS format. This could be a result of this format being incompatible with raters’ organization of information in memory (DeNisi & Peters, 1996; DeNisi et al., 1989) or of simultaneously placing too many cognitive demands on raters.

References


the annual meeting of the American Educational Research Association, Chicago, IL.


Appendix

Teacher Behavior Inventory Items Used in Accuracy-Score Calculations

<table>
<thead>
<tr>
<th>Clarity: methods used to explain or clarify concepts and principles.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gives several examples of each concept.</td>
</tr>
<tr>
<td>2. Uses concrete, everyday examples to explain concepts and principles.</td>
</tr>
<tr>
<td>3. Repeats difficult ideas several times.</td>
</tr>
<tr>
<td>4. Stresses most important points by pausing, speaking slowly, raising voice, etc.</td>
</tr>
<tr>
<td>5. Points out practical applications of concepts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enthusiasm: use of nonverbal behavior to solicit student attention and interest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Speaks in a “dramatic” or expressive way.</td>
</tr>
<tr>
<td>8. Moves about while lecturing.</td>
</tr>
<tr>
<td>9. Gestures with hands or arms.</td>
</tr>
<tr>
<td>10. Walks up aisles beside students.</td>
</tr>
<tr>
<td>11. Tells jokes or humorous anecdotes.</td>
</tr>
<tr>
<td>12. Reads lecture verbatim from prepared notes or text.</td>
</tr>
<tr>
<td>13. Smiles or laughs while teaching.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speech: characteristics of voice relevant to classroom teaching.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Stutters, mumbles, or slurs words.</td>
</tr>
<tr>
<td>16. Speaks at appropriate volume.</td>
</tr>
<tr>
<td>17. Speaks clearly.</td>
</tr>
<tr>
<td>18. Speaks at appropriate pace.</td>
</tr>
<tr>
<td>19. Says &quot;um&quot; or &quot;ah.&quot;</td>
</tr>
<tr>
<td>20. Voice lacks proper modulation (speaks in monotone).</td>
</tr>
</tbody>
</table>

**Note.** These items were adapted from a 60-item inventory developed by H. G. Murray, Department of Psychology, University of Western Ontario, London, Ontario, Canada. It is not copyrighted and may be reproduced for any valid research or instructional development purpose. Please see Murray (1987) for the complete inventory. Reproduced for publication by permission of the author.