A Second Look at the Relationship Between Rating and Behavioral Accuracy in Performance Appraisal

Juan I. Sanchez and Phillip De La Torre
Florida International University

The accuracy of ratings concerning ratees' strengths and weaknesses was hypothesized to correlate with the accuracy of memories concerning ratees' behaviors, although delay of the rating task was expected to attenuate this relationship. In contrast, because it was presumed that raters would rely on their on-line impressions to formulate holistic evaluations, no association between accuracy of overall ratee evaluations and behavioral accuracy was hypothesized. The results of a laboratory study (N = 262) supported these predictions, suggesting that accurate behavioral memories might help in formulating accurate evaluations of ratees' strengths and weaknesses.

The assumption that enhanced opportunities to observe and record behavior lead to accurate ratings pervades many areas of performance evaluation. For example, in scoring work-sample tests, evaluators observe behavior at different steps of a job task, and then they rate performance at each step (Hedge & Teachout, 1992). In a similar fashion, assessment center evaluators are instructed to observe and record behaviors that are later relied on to generate ratings (Thornton & Byham, 1982). During the observation period, assessors are encouraged to record behaviors and to delay the formation of judgments, which are thought to be most accurate when fully based on a wide sample of ratee behaviors. Behavioral observation training (Thornton & Zorich, 1980) is another example of intervention intended to preserve behavioral memories that has rating accuracy as its ultimate goal.

In short, raters are directed to rely on their memories of behaviors because rating accuracy is thought to be a function of the accuracy of behavioral memories. This presumed relationship between rating and behavioral accuracy has weak empirical basis, and it was the focus of the present study. Indeed, social cognition research has evidenced that there is often no correspondence between judgments made about a person and memory for specific facts about that person (Lichtenstein & Srull, 1987; Srull & Wyer, 1989). In the context of performance appraisal, researchers have also questioned the relationship between behavioral accuracy and rating accuracy (Murphy, 1991; Murphy & Balzer, 1986; Murphy & Cleveland, 1991, p. 148).

In perhaps the only explicit study of this issue in the performance appraisal domain, Murphy, García, Kerka, Martín, and Balzer (1982) found support for a positive relationship between rating accuracy and the accuracy of frequency ratings of critical behaviors. Usage of a common method (i.e., ratings) to index both behavioral and rating accuracy might have inflated their intercorrelation in Murphy et al.'s study. Inflation might have occurred because (a) raters might have indicated high frequencies for behaviors that were consistent with the impressions they held for a ratee and low frequencies for behaviors that were inconsistent with such impressions (Sulska & Day, 1992) and (b) lenient raters might have indicated high frequencies, whereas stringent raters might have indicated low frequencies. Because they include both present and foil items and a yes–no format, behavioral accuracy tests using signal-detection theory (SDT) are less vulnerable to impression-induced bias and to individual differences in leniency or stringency than are frequency ratings (Borman & Hallam, 1991; Lord, 1985; Snodgrass & Corwin, 1988).

Regardless of potential inflation in past estimates of the relationship between behavioral accuracy and rating accuracy, asking which factors determine the emergence of this relationship may be a more useful question than asking whether there is a relationship between these two types of accuracy. Indeed, knowledge of the moderators of this relationship may pinpoint those rating tasks that demand accurate behavioral memories. One of these moderators may be delays between observation and the
rating task. Delays may affect the availability of behavioral memories and, therefore, attenuate their relationship with ratings produced under such high memory demands (Murphy, Philbin, & Adams, 1989; Nathan & Lord, 1983; Williams, DeNisi, Meglino, & Cafferty, 1986).

Another potential factor affecting the relationship between behavioral accuracy and rating accuracy is the purpose of observation, which is sometimes referred to as processing objective. When observation is conducted with no specific objective in mind or only a general objective, such as to comprehend the information presented, a global impression of the person may not be formed at the time of information acquisition (Lichtenstein & Srull, 1987; Srull & Wyer, 1989; Williams, Cafferty, & DeNisi, 1990; Woehr & Feldman, 1993). If later asked to make an evaluative judgment, raters will need to retrieve their behavioral memories, with the consequent increase in the relationship between behavioral accuracy and rating accuracy.

In summary, memory demands and processing objectives may determine the extent to which accurate behavioral memories are related to rating accuracy. However, delays between observation and performance ratings are imposed by the limited time and resources that organizations are willing to invest in performance appraisal. Similarly, because supervisors probably switch back and forth between attempts to comprehend and attempts to evaluate employee behavior, asking real-world raters to follow a particular type of processing objective may be unrealistic.

In essence, the implications of research on the effects of delay and processing objective on the relationship between behavioral accuracy and rating accuracy may be bounded by the administrative realities constraining performance appraisal in organizations. In contrast, research on the aspects of judgment (e.g., holistic judgments concerning the whole person vs. dimensional judgments concerning the person's strengths and weaknesses) whose accuracy depends on behavioral memories may pinpoint those rating tasks that require accurate behavioral memories (Borman & Hallam, 1991; Sulsky & Day, 1992). Hastie and Park (1986) provided the rationale needed to uncover such rating tasks. They noted that online impressions are formed at the time the information is encoded and include those that are spontaneously made in person perception. On-line evaluative impressions would be encoded separately from the particular behaviors learned about the target (DeNisi & Williams, 1988), and they will be accessed later when a global evaluation of the rates is demanded (Lichtenstein & Srull, 1987; Wilson, Northcraft, & Neale, 1989). Given the relative independence between on-line impressions and memory for specific behaviors, no relationship between judgment and behavioral accuracy should be expected.

Conversely, when raters face specific judgments about ratees' strengths and weaknesses, on-line impressions are unlikely to provide sufficient information on ratees' particulars. Raters would need to retrieve behavioral memories and use them to reach such detailed judgments (Lichtenstein & Srull, 1987). Thus, a relationship between this type of dimensional judgment and memory should be expected.

Interestingly, the ideas developed by Hastie and Park (1986) concerning the association between behavioral and rating accuracy have implications for Cronbach's accuracy components. These components have been used as dependent variables in numerous studies of rating accuracy (Murphy & Cleveland, 1991, p. 226) because they represent the preferred strategy for decomposing the various sources of variance associated with the discrepancy between individual ratings and a set of proxy true scores. Elevation is the accuracy component associated with the average rating across all ratees and performance dimensions. Differential elevation is related to the average rating for each ratee across all performance dimensions (thus representing the level of accuracy of the raters' rank ordering of the ratees). Stereotype accuracy is associated with the average rating for each performance dimension across ratees (hence representing the accuracy of the rater's assessment of the group of ratees in the various performance dimensions). Lastly, differential accuracy reflects the accuracy of the rater's assessment of ratee differences along the various dimensions.

Two components of Cronbach's measures are concerned with the precision of holistic judgments of the group of ratees (i.e., elevation) and the rank order of ratees (i.e., differential elevation) across dimensions. Accuracy of these holistic judgments demands a precise impression of the rates' global level of performance rather than selective attention to behavioral information. According to Hastie and Park (1986), holistic evaluations are probably driven by on-line impressions partly independent from memory for behaviors, and therefore, we did not expect a relationship between behavioral accuracy and the elevation and differential-elevation components. In contrast, the stereotype-accuracy and differential-accuracy components are concerned with relative strengths and weaknesses of the group of ratees (i.e., stereotype accuracy) and individual ratees (i.e., differential accuracy). The information needed to reach these dimensional judgments would not be available in on-line impressions, and therefore, raters will be forced to retrieve behavioral memories and use them for their judgments. Hence, we proposed the following:

_Hypothesis 1:_ The accuracy of judgments concerning relative strengths and weaknesses captured by stereotype accuracy and differential accuracy will be positively related to behavioral accuracy.
Reliance on evaluative impressions may increase with delays between the observation and the rating period (DeNisi, Cafferty, & Meglino, 1984; Nathan & Lord, 1983; Phillips, 1984). The association between behavioral and rating accuracy may therefore diminish under the severe memory demands often found in performance ratings obtained in field settings. Under high memory demands, raters are likely to resort to their on-line impressions, which, although initially influenced by behavioral information, would no longer be related to the fading behavioral memories that raters have when the rating task is delayed (Hastie & Park, 1986; Wyer & Srull, 1986).

Hence, we proposed the following:

Hypothesis 2: The relationship between behavioral accuracy and the differential-accuracy and stereotype-accuracy components of rating accuracy will decrease under the high memory demands imposed by a delayed rating task.

Method

Participants

A total of 281 students in their junior or senior year of college participated for course credit. The data for 5 participants were discarded because of incomplete responses, and 14 participants did not return for the rating session, resulting in a sample of 262 students included in our analyses. There were 171 women and 91 men in this sample. The primary majors were business and psychology.

Development of Videotapes

To enhance the generalizability of our results, we chose two jobs that were familiar to most of the participants: bank teller and salesperson (Igen & Faver, 1985). We chose to portray incumbents of two different jobs because differential evaluation and differential accuracy are concerned with differences among rates, and thus, different stimuli might help increase the interrate variance needed to test our hypotheses. Moreover, differences in the level of performance were intentionally built into the videotapes to increase interrate variance (i.e., whereas the bank teller's behavior represented a moderately high level of performance, the performance level of the salesperson was average). Note also that unwanted context effects may be more likely when the same job is used than when two different jobs are used. It is common for supervisors to oversee more than one job title, and therefore this situation seemed ecologically valid.

A group of six research assistants with experience in retail or banking developed scripts based on their own experiences and on critical job tasks, as identified in job analyses conducted in local organizations. One of the videotapes portrayed a customer wanting to cash a check without proper identification. The customer argued that he always cashed his checks that way, while the bank teller repeated that it was the bank's policy not to cash checks without appropriate identification. The customer kept insisting until the bank teller decided to seek a supervisor's approval to cash the check. The other videotape presented a salesperson receiving a call informing him that his coworker was not coming to work. Next, a customer entered the store and requested a demonstration of a particular videocamera. Although the salesperson showed the camera to the customer, the salesperson claimed not to have the time to conduct a demonstration. The salesperson asked the customer to come back when there were more sales personnel. After subsequent arguments, the customer became upset and left the store, promising not to come back. To increase ecological validity, the scripts were filmed in a bank and an electronics store, and the bank teller and the salesperson were portrayed, after numerous rehearsals, by real job incumbents. The videotapes were professionally edited, and each lasted about 10 min.

Behavioral Accuracy Test

A list of job-related behaviors that were either present or absent in the videotapes was assembled. Absent behaviors or foils were written to reflect some degree of realism so they would not be easily recognized. Twenty behaviors (i.e., 10 for each videotape) that were correctly identified as either present or absent in the videotapes by the six individuals involved in the development of the videotapes were kept in the recognition test (e.g., the employee requested two valid IDs from the customer). The number of positive and negative behaviors was balanced to prevent biases due to response sets. Eleven behaviors were actually present in the videotapes, and 9 behaviors were foils.

Behavioral accuracy was scored using the SDT-type discrimination index ($P_t$) proposed by Snodgrass and Corwin (1988). This index takes into account the respondents' tendency to provide either positive or negative identifications, and it can range from −1 to 1:

$$P_t = HR - FAR,$$

where HR is the hit rate and FAR is the false-alarm rate. A hit was designated as a yes response to a previously observed behavior, and the hit rate was the proportion of observed behaviors marked yes. The false-alarm rate was the proportion of foils marked yes. The bias index ($B_t$) was computed to estimate participants' response tendency. This index ranges from 0.0 to 1.0, with scores above 0.5 indicating a tendency to provide responses of yes and scores below 0.5 suggesting a tendency to provide responses of no. The formula used to compute this measure was as follows:

$$B_t = \frac{FAR}{[1 - (HR - FAR)]}.$$

$P_t$ and $B_t$ have shown moderately high levels of reliability in behavioral accuracy research (Borman & Hallam, 1991).

Development of Rating Scales

The behaviorally anchored rating scales (BARS) used in this study were a subset of the nine BARS developed by Sanchez and Fraser (1993) as a generic instrument for appraising employee performance in customer-contact jobs. The procedure outlined by Smith and Kendall (1963) was followed in scale development. Although the videotapes were produced so that they would provide opportunities to evaluate this set of performance
dimensions, we considered that their short duration would not allow a thorough evaluation of all dimensions. Thus, we independently chose the most relevant dimensions for both videotapes. The four dimensions on which we both agreed were tolerance-patience, personable, informative, and cooperative-team player. Matches between videotaped behaviors and BARS anchors, which might bias ratings (Murphy & Constans, 1987), were not apparent.

**Experimental Manipulation**

Participants were asked to rate the videotapes either immediately after viewing them or 10 days after viewing them (i.e., hereinafter referred to as the immediate and delay conditions, respectively). Because 14 participants did not return for the rating session in the delay condition, 124 participants were kept in the delay cell versus 138 in the immediate cell.

**Proxy True Scores and Accuracy Measures**

Proxy true scores were derived from average expert ratings in the manner used by Borman (1978) and others (Murphy et al., 1982; Smither & Reilly, 1987). A group of 14 volunteers with more than 2 years of experience in banking or retail sales served as experts. This panel of experts was given multiple opportunities to view the videotapes, and fractions of them were replaced several times upon request. We analyzed ratings in a Rater \times Ratee \times Dimension design to assess convergent and discriminant validities (Borman, 1978). Whereas the intraclass correlation for the ratee effect (i.e., convergent validity) was .98, the intraclass correlation for the Ratee \times Dimension effect (i.e., discriminant validity) was .66. These correlations compared favorably with the ones obtained by Borman (rs = .69 and .58 for convergent and discriminant validity, respectively) and by Murphy et al. (rs = .58 and .47, respectively). Cronbach's accuracy components representing discrepancies between individual ratings and proxy true scores were computed according to the formulas provided by Murphy et al.

**Procedure**

Participants reported to the assigned room in groups ranging from 7 to 31 participants; groups were randomly assigned to either the immediate or the delay condition. Because expecting to perform a demanding memory task might lead individuals to use a different information-processing strategy (Higgins & Lurie, 1983), all participants were told that they were about to watch two videotapes portraying customer–employee interactions and that they would be asked to rate the employees’ service performance in the near future, without providing them with the specific time frame for this evaluation. The order of presentation of the videotapes was counterbalanced. The order of the behavioral recognition and the rating task was also counterbalanced to control for potential priming effects. Half of the participants first took the behavioral recognition test consisting of 20 behaviors, for which they were instructed to identify whether each behavior was present or absent in the videotapes. Next, they performed the rating task. The other half of the participants completed these tasks in reversed order. Immediately before the rating task, participants were instructed how to use the rating scales.

**Results**

Descriptive statistics and Pearson zero-order correlations among the study variables are presented in Table 1. We computed hierarchical regression equations to test the hypotheses (see Table 2). Cronbach’s accuracy measures were separately regressed on the dummy-coded variable representing the experimental manipulation (i.e., delay) at Step 1. No significant main effects were observed. At Step 2, behavioral accuracy ($P_b$) was entered in the regression equations. The negative beta weight revealed a small but statistically significant effect of $P_b$ on stereotype accuracy in the expected direction: $P_b$ was positively related to stereotype accuracy (note that higher numbers on rating-accuracy measures represent lower levels of accuracy), hence providing partial support for our first hypothesis. To test Hypothesis 2, the interaction between the experimental manipulation and $P_b$ (as represented by the cross-product of Delay \times $P_b$ in Table 2) was entered at Step 3. Statistically significant interactions emerged for both stereotype accuracy and differential accuracy. The positive beta weights suggested that these interactions were in the expected direction, although their effect sizes were small.

Separate regression equations were computed for participants in the delay condition and for those in the immediate condition to facilitate the interpretation of the significant interactions. In support of Hypothesis 2, plotting the predicted value of the rating-accuracy components against the value of $P_b$ revealed that whereas stereotype accuracy and differential accuracy were positively associated with $P_b$ in the immediate condition, this relationship was virtually nonexistent in the delay condition. Because these plots were practically identical, only the one concerning $P_b$ and stereotype accuracy (see Figure 1) is included.

No effects for either order of videotape presentation or order of rating task–behavioral recognition test were revealed by a series of $t$ tests on both behavioral and rating-accuracy measures at $p < .05$. The need to have multiple raters to compute Cronbach's accuracy measures precluded the examination of separate effects for the two videotapes.

**Discussion**

This study examined the relationship between behavioral and rating accuracy in performance appraisal. A distinction was made between components of rating accuracy concerning holistic evaluations of raters and those concerning raters’ strengths and weaknesses. We hypothesized that behavioral accuracy would be associated with
the accuracy of judgments concerning ratees’ strong and weak points and that this relationship would diminish under high memory demands.

As predicted, the relationship between behavioral and rating accuracy was confined to dimensional judgments of strengths and weaknesses (i.e., the stereotype-accuracy and differential-accuracy measures defined by Cronbach), especially in the immediate condition. The lack of an association between behavioral accuracy and both elevation and differential elevation did not seem attributable to insufficient statistical power because our sample size was capable of detecting even small effect sizes (Cohen, 1992). There is a certain parallelism between these findings and those of recent research on frame-of-reference training (Suls & Day, 1994; Woehr, 1994), suggesting that judgments are related to behavioral memories when such memories are relevant and easily accessible and to global impressions otherwise.

The small effect sizes of behavioral accuracy on stereotype accuracy and differential accuracy in the immediate condition suggest that sources of variance other than memories of behaviors affect rating accuracy, such as (a) cognitive schemata that may facilitate the recognition of prototype-consistent behaviors while hindering the recognition of prototype-inconsistent behaviors (Feldman, 1981; Lance, Woehr, & Fiscarco, 1991; Lord, 1985); (b) behaviors not included in the recognition test, which might not fully represent the scope of behaviors displayed in the videotapes; and (c) individual differences affecting both rating and behavioral accuracy (Borman, 1979; Borman & Hallam, 1991).

The weak association between behavioral and rating accuracy in the delayed condition suggests that when the observation period and the rating task do not immediately follow each other (as usually occurs in annual or semiannual performance appraisals), raters wishing to accurately assess ratees’ strengths and weaknesses may be encouraged to maintain behavioral records instead of relying on their long-term memories. Because of the potential biases affecting behavioral diaries, raters may be better off relying on behavioral checklists (Maurer, Palmer, & Ashe, 1993). According to our results, however, the completion of these checklists should closely follow the observation period to be really effective.

Whereas delay of the rating task reduced behavioral
accuracy, it did not affect rating accuracy. This finding corroborates those of previous studies (Murphy & Balzer, 1986; Murphy et al., 1989; Smither & Reilly, 1987), suggesting that raters overcome memory loss over time by relying on their on-line impressions (Nathan & Lord, 1983; Sulsky & Day, 1994).

Caveats concerning the use of a recognition test, the salience of performance dimensions, and the use of a laboratory setting should be mentioned. First, we measured behavioral accuracy using a recognition test rather than a recall test. One might argue that recall is more central to performance appraisal than is recognition memory. However, because of the usage of behavioral anchors and behavioral incidents in performance-rating tasks, Sulsky and Day (1992) underscored the significance of recognition memory in this context. Indeed, behavioral recognition seems to pervade behaviorally anchored, behavioral observation, forced-choice, and mixed standard scales, as well as behavioral checklists used as rating aids in work samples and assessment centers.

With regard to the salience of performance dimensions, raters remained unaware of the particular dimensions to be used in the appraisal during the observation period, although they were told that they would later evaluate service performance. The relationship between behavioral accuracy and both stereotype accuracy and differential accuracy may decrease when performance dimensions are salient to raters beforehand, because behaviors may be inferred from evaluative impressions rather than evaluations based on memories of behaviors (Murphy & Cleveland, 1991, p. 154; Woehr, 1992). However, only dimensions relevant to service performance were selected in our study, implying that the judgments requested were not totally unexpected. Future research can address this question by including both relevant and irrelevant dimensions.

Turning now to the constraints imposed by the laboratory setting, these results may change because of variations in time frame, rater affect, purpose of observation, and number of and differences among ratees, among
other factors. The restricted generalizability of this line of research to field ratings notwithstanding, Borman and Hallam (1991) suggested its relevance to assessment centers. Because the observation period is closely followed by the rating task in assessment centers and work samples, the relationship between behavioral and rating accuracy observed in our immediate condition supports the emphasis on behavioral observation advocated by proponents of these methods.

In essence, accurate behavioral memories may not necessarily increase the accuracy of all aspects of performance evaluations. These findings do not imply that behavioral record keeping is frequently futile. Precise behavioral records are needed for effective performance feedback (Murphy, 1991) and as a defense against allegations of unfair employment practices. Our results merely suggest that, in the context of performance appraisal, behavioral memories appear to be most beneficial when an accurate identification of strengths and developmental needs is sought.

References


Received February 21, 1995
Revision received July 21, 1995
Accepted July 26, 1995

---

New Editor Appointed

The Publications and Communications Board of the American Psychological Association announces the appointment of Kevin R. Murphy, PhD, as editor of the *Journal of Applied Psychology* for a six-year term beginning in 1997.

As of March 1, 1996, submit manuscripts to Kevin R. Murphy, PhD, Department of Psychology, Colorado State University, Fort Collins, CO 80523-1876.