JOURNAL OF MANAGEMENT SYSTEMS. Vol. 10, No. 1, 39-52

ARTICLE

A Short-Term Longitudinal Study Of The Effects Of Rating Delays On Differential Accuracy

Robert D. Costigan

Assistant Professor Management
St. John Fisher College

■ ABSTRACT

Performance rating models suggest that rating accuracy should suffer with a time delay between observation and judgment. Considering the inconsistent results of past studies, it is unclear what rating delay conditions produce deleterious effects on rating accuracy. Hence, a more systematic examination of the effects of multiple rating delays on accuracy seems needed. A 2 (high vs. low performance) X 6 (no delay vs. 7-day delay vs. 14-day delay vs. 21-day delay vs. 28-day delay vs. 35-day delay) X 2 (one ratee vs. two ratees) X 4 (graphic vs. graphic with behaviors vs. BOS vs. BARS) ANOVA was conducted for each performance dimension. The results of this study showed that the number of ratees observed and the format used influenced rating accuracy whereas the length of the rating delay apparently did not.

■ INTRODUCTION

Over the past few years, performance appraisal researchers have carefully studied the four cognitive processing stages involved in the evaluation process (Ilgen, Barnes-Farrell, & McKellin, 1993). These processing stages include: (1) rater observation of ratee actions; (2) storage of this information in the rater's memory; (3) recall of pertinent information concerning the ratee from memory; and (4) integration of this recalled information so that judgments can be made on the ratee's performance. Bretz, Milkovich, and Read (1992) stated that continued research in the recall domain and related memory characteristics should prove fruitful. In their view, real-world managers spend very little time preparing for, and conducting performance appraisals. Typically, annual reviews place great demands on the rater's memory for performance-related information, especially since keeping an on-going written account (i.e., diary-keeping) appears to be more of the exception than the rule. Thus, long time delays between observation and judgment are characteristic of the rating process. Managers may face the arduous task of recalling and evaluating some subordinate behaviors occurring months ago.

The present study heeds the request by Bretz et al. (1992) for attention to the effects of rating delays on rating accuracy. The current rating delay literature appears to be overly confusing and confounded. Some of the confusion appears to lie with each study's unique operational definition of relevant variables (e.g., length of the delay, format type, number of rating stimuli). Hence, a more systematic look at the effects of multiple rating delays under different appraisal conditions (i.e., varied number of stimuli observed, various rating formats) should provide added clarity to this literature.

Cardy and Dobbins (1994) reported that rater accuracy research in the performance appraisal domain is surprisingly limited. In their view, future research should concentrate more heavily on the cognitive-processing conditions that produce rating accuracy. Thus, the assessment of rating accuracy rather than rating error is the focus of this study. Because Cronbach's (1955) differential accuracy (i.e., rating accuracy for each ratee within each dimension) is a particularly good measure of rating accuracy

Correspondence regarding this article should be addressed to: Robert Costigan, St. John Fisher College, 3690 East Avenue, Rochester, New York 14618. Email: costigan@sjfc.edu

(Hastorf, Schneider, & Polefka, 1970; Borman, 1977), accuracy in this study refers to differential accuracy.

Tulving (1983) concluded that after 100 years of memory research, there is still little agreement on how information is stored in memory. There are, as a consequence, numerous models that attempt to explain the memory process. Wyer and Srull's (1981) model suggests that the information processor first holds observed behavioral information in a temporary work space (i.e., short-term memory). During this brief interlude, observed stimulus information is available for recall. Over a short period of time, the work space's specific behavioral information is either replaced or lost. Wyer and Srull's (1981) interpretation of long-term memory processes suggests the presence of storage-like bins in human memory. Each bin is affixed a bin label that distinguishes the bin from all other bins. This label summarizes the content of the bin. Information stored at the top of the bin is thought to be more accessible than is information that is stored at lower levels of the bin.

Rating models (e.g., Lord, 1985) attempting to elaborate on the work of Wyer and Srull as well as others suggest that the rater first transforms observed ratee behavior to meaningful representations (i.e., observed raw stimulus behaviors are not stored in memory but the encoded representations of these behaviors are). The encoded representation is stored in temporary memory. Factors such as memory decay and interference between competing stimuli bring about short-term memory loss. Thus, temporary memory has limited duration and limited capacity.

The retained representation is then transferred to categories in long-term memory storage. A salient feature of the ratee (e.g., novelty) is often the basis for the rater's unconscious choice of a particular category to store information about the ratee (Feldman, 1981). With the passage of time between observation and judgment, the stimulus information in a long-term memory category loses behavioral specificity. By default, the general characteristics of the category (i.e., prototype information) tend to be remembered and then become the basis of the performance ratings. Both memory decay and the greater accessibility in memory of prototypical features of the category are given as reasons for the bias in recalling prototype information rather than observed stimulus information.

■ RATING DELAYS

The previously described storage and recall processes suggest that rating accuracy should suffer with delays between observation and recall. Indeed, three reviews of the performance appraisal literature (Heneman, Wexley, & Moore, 1987; Bretz et al., 1992; Ilgen et al., 1993) concluded that rating accuracy usually suffers with such delays in contrast to ratings made without a time delay. In early rating delay studies, for example, Rush, Phillips, and Lord (1981) and Heneman and Wexley (1983) found that rating delays of two days and one-to-three weeks respectively led to less accurate ratings. Smither and Reilly (1987) tested the effects of a rating delay by incorporating between-subjects and within-subjects designs in the same study. They discovered that, with a one-day delay, rater accuracy decreased for the within-subjects design only. A few studies, such as Nathan and Lord (1983), Barnes-Farrell and Couture (1984), and Murphy, Philbin, and Adams (1989), found no adverse effects from rating delays. The delays in these three studies were two days, one week, and one-to-seven days respectively. One study conducted by Murphy and Balzer (1986) surprisingly found that raters' differential accuracy improved with a one-day rating delay relative to the raters' accuracy in the no-delay condition.

Ilgen et al. (1993) and Bretz et al. (1992) concluded that there is still some confusion concerning the deleterious effects of rating delays. Conflicting findings in the rating delay literature have drawn queries on the influence of rating delays on rating accuracy. For instance, Ilgen et al. (1993) questioned whether rating accuracy depreciates linearly or monotonically with increased time delays between observation and judgment. Previous rating delay studies incorporated different lengths in the time delays, ranging from 24 hours to three weeks. With the exception of the Murphy et al. (1989) study, the effects of rating delays have not been systematically evaluated with multiple time delays in the same study. Murphy et al. (1989) included time delays of one, three, and seven days but found no main effects for the

delay factor, meaning that rating accuracy remained unchanged over the four time-of-rating conditions. It is unclear what rating delay conditions (i.e., length of the delay) produce an ill-effect on rating accuracy.

One purpose of the present study is to revisit this issue by introducing time delays of seven, 14, 21, 28, and 35 days. Assuming that increased time delays produce greater inaccuracy, Ilgen et al.'s (1993) question (i.e., does accuracy depreciate linearly or monotonically with increased time delays between observation and judgment?) will also be considered in this study.

Number of Rates Observed

Previous performance appraisal research has not considered the effects of rating delays along with the effects of the number of stimuli observed. Two studies involving rating delays (i.e., Nathan & Lord, 1983; Rush et al., 1981) required the raters to observe and evaluate one rating stimulus. Focusing attention on one target stimulus should involve minimal interference from competing stimuli and enhance the prospect of accurate memory recall and judgment. Other rating delay studies incorporated multiple stimuli, thereby taxing the cognitive-processing capabilities of raters to a greater extent. For example, Heneman and Wexley's (1983) included three ratees in their study while Murphy and his colleagues (1986, 1989) had subjects' observe four ratees. Smither and Reilly (1987) asked their raters to evaluate the performance of five ratees.

Another purpose of this research effort is to determine the effects of evaluating the performance of a single target stimulus versus multiple target stimuli under varied rating delay conditions. Such effects may perhaps be more pronounced under the more severe recall conditions, such as 28-to-35 days. Observing multiple ratees should: (1) produce increased interference in the observation stage; (2) add to the difficulty of recalling relevant performance-related information; (3) introduce additional bias (e.g., context effects) in the judgment phase; and (4) require additional motivation on the part of the rater in all processing stages.

Rating Formats

Feldman's (1981) seminal work focusing on cognitive information processing in performance appraisals provides a basis for revisiting rating format research, especially in the context of rating delays. Different rating formats require raters to use different cognitive processes; and, therefore, identifying the formats which best complement the particular cognitive processing demands placed on raters is an important area for future research (Murphy & Cleveland, 1995). Significant format findings (Dossett, 1989; Hartel, 1993; Heneman, 1986; Murphy & Constans, 1987; Murphy & Davidshofer, 1988) after Landy and Farr's (1980) call for a moratorium on this kind of research suggest that cognitive-processing issues have not been adequately addressed in rating format research.

At the time of their moratorium, most of appraisal research concentrated on minimizing rating error (Cardy & Dobbins, 1994). Cardy and Dobbin's (1994) noted the dearth of format research in relation to the rating accuracy criterion. The present study attempts to address this concern by assessing the effects of various formats on rating accuracy under different time-of-rating and number-of-stimuli conditions. An effort will be made to identify which formats best cope with the taxing cognitive processing demands imposed on raters in this study (i.e., seven to 35-day time delays, two ratees versus one).

Some rating delay studies (e.g., Nathan & Lord, 1983) had subjects rate with graphic scales (GRS), which focus on the global performance dimensions instead of specific ratee behaviors. A few studies (e.g., Murphy & Balzer, 1986; Murphy et al., 1989) incorporated both behavioral and dimensional ratings. Other performance appraisal studies incorporating a rating-delay variable (e.g., Rush et al., 1981; Heneman & Wexley, 1983) had subjects rate only with behavior-based scales. The results of these studies show no clear pattern as to the effectiveness of one format over another under various rating delay conditions.

For instance, studies incorporating behavior-based formats (e.g., Heneman & Wexley, 1983; Rush et al., 1981) reported that rating accuracy suffered with rating delays while other studies using behavioral formats (e.g., Murphy et al., 1989) reported no ill-effects of time delays on rating accuracy. As reported, one study (i.e., Murphy & Balzer, 1986) even found that behavior-based ratings improved after a one-day delay versus the no-delay ratings. In fairness to this issue of differential format effectiveness under various rating delay conditions, the composition of these behavioral formats differed considerably between studies. For instance, Heneman and Wexley's (1983) study used behavioral scales with absolute frequency anchors (i.e., number of times that a behavior occurred); Murphy and his colleagues' studies (1986, 1989) included behavior observation scales which have relative frequency anchors (i.e., "never," "always"); and Rush et al. (1981) used the Leader Behavior Description Questionnaire which has "agree-disagree" type of anchors.

Considering the variation in the format content between rating delay studies and considering that only a few of these studies provided a direct contrast of competing rating scales, further examination of the effectiveness of commonly used formats under multiple delay conditions appears to be warranted. One graphic dimensional format with evaluative anchors (GRS) and three behavior-based formats (i.e., graphic behavioral format with evaluative anchors (GRBEH), Latham and Wexley's (1977) behavioral observation scale with relative frequency anchors (BOS), and Smith and Kendall's (1963) behaviorally anchored rating scales (BARS)) are included for comparison in the present study.

METHOD

Subjects and Task

Subjects were enrolled in undergraduate management and psychology courses at a northeastern college (N = 466). They received extra credit for participating. Groups of two-to-six subjects viewed either one or two videotapes. They were told that the videotape(s) was a segment of a college course taught during the previous semester. The "job" of instructor was chosen because carefully developed behaviorally-based performance appraisal instruments assessing a number of dimensions of instructor performance were readily available (Hoffman and Dossett, 1982). Thus, raters observed and evaluated either one or two teachers.

Experimental Design

Five factors were employed in a 2 X 2 X 2 X 2 X 6 factorial design. The first three factors consisted of three performance dimensions that were experimentally manipulated (high versus low performance). The fourth factor was the number-of-stimuli variable (one target stimulus observed versus two target stimuli observed). The fifth factor was a time-of-rating variable (no delay vs. seven-day delay vs. 14-day delay vs. 21-day delay vs. 28-day delay vs. 35-day delay). A sixth, within-subjects factor consisted of four rating formats (GRS vs. GRBEH vs. BARS vs. BOS) assessing ratee performance on three manipulated performance dimensions and one non-manipulated dimension.

Videotape Development

The first target stimulus was the videotaped performance of a fictitious college professor lecturing on the topic of Landscaped Office Designs (LODs). This 14-minute LOD videotape developed by Dossett and Costigan (1987) served as the primary rating stimulus in this study. The three performance dimensions manipulated in this videotape were: (a) Explanation of Concepts, (b) Student-Teacher Interaction, and (c) Preparedness. Initially, Dossett and Costigan (1987) developed two videotapes: one representing high performance on all three dimensions and one portraying low performance on all three dimensions.

Thirteen graduate management students who were familiar with performance appraisal research served as expert raters in this pilot study. They first reviewed the GRS rating form. Then, six graduate students observed and rated the videotaped performance intended to portray high performance across

the three dimensions, and seven graduate students observed the videotaped performance intended to portray low performance across dimensions. The mean GRS ratings of the six expert raters for the high performance tape are: 3.83 (SD = .75) for Explanation of Concepts, 4.17 (SD = .41) for Student-Teacher Interaction, and 4.67 (SD = .52) for Preparedness. The mean GRS ratings of the other seven expert raters for the low performance tape are: 1.29 (SD = .49) for Explanation of Concepts, 1.57 (SD = .79) for Student-Teacher Interaction, and 1.43 (SD = .53) for Preparedness. Separate independent-samples t-tests indicated significant differences between performance levels for all three dimensions ($p < .001$).

Because performance on three dimensions was being manipulated, six more videotapes were carefully constructed so that there were eight LOD videotapes in all, each videotape portraying a different combination of performance for the three dimensions. Thus, the quality of performance (high vs. low) was manipulated while the number, distribution, and sequencing of target behaviors were held constant across the eight videotaped performance conditions. A fourth performance dimension, "Objectivity: impartiality; is unbiased in treatment of students," was held constant across videotapes (i.e., performance on this dimension was high in the eight LOD videotapes). The mean GRS rating of the six expert raters viewing the high-performance videotape is 4.00 (SD = .63) for the Objectivity dimension, confirming the ratee's relatively high performance on this dimension.

An 11-minute videotape of another professor's lecture on the topic of sales marketing (i.e., sales forecasting) served as a second rating stimulus for raters in the multiple-stimuli condition.

Procedures

Subjects in the single-stimulus condition were randomly assigned to one of the eight LOD videotape conditions and to a time-of-rating condition. They observed one LOD videotape and then rated the instructor's performance. Likewise, subjects in the multiple-stimuli condition were randomly assigned to one of the eight LOD videotape conditions and to a time-of-rating condition. They observed one of the eight LOD videotapes and the sales forecasting videotape. They then rated the performance of both instructors. The presentation of the LOD and sales forecasting videotapes was counterbalanced across performance-level conditions. In sum, each LOD videotape was observed by four-to-six subjects within each time-of-rating condition and number-of-stimuli condition. The presentation of the four formats was counterbalanced across subjects in each condition. The four rating formats used their respective variations of a five-point scale for each performance dimension. The evaluative anchors in GRS and GRBEH were: (1) "poor" to (5) "excellent." The most extreme evaluative anchors in BARS were: (1) "very low" to (5) "very high." The relative frequency anchors in BOS were: (1) "almost never" to (5) "almost always." The three manipulated dimensions as well as the non-manipulated Objectivity dimension were included in each format.

After observation of the videotape(s), subjects in the no-delay condition immediately evaluated the performance of the instructor(s). Subjects in the delay conditions returned either seven days, 14 days, 21 days, 28 days, or 35 days after observation to complete their evaluations. After completing each of the four formats, subjects completed an item assessing their perceived rating accuracy (i.e., "How accurately does the preceding rating form itself allow you to express your evaluation of the instructor in the videotape?"). Anchors for this 5-point scale were (1) "not at all accurately" to (5) "very accurately." Finally, subjects were given a short debriefing after finishing the assigned written work.

■ RESULTS

To test whether rating accuracy depreciates over time and with multiple rating stimuli and with certain formats and not others, a 2 (performance level) X 2 (number of stimuli) X 6 (time of rating) X 4 (format) within-subjects analysis of variance (ANOVA) was conducted for each manipulated performance dimension in the LOD videotapes. The ratings assigned to the LOD instructor on each performance dimension served as the dependent measure in each analysis. Either a three-way or two-way interaction involving the performance-level variable would indicate significant differential sensitivity (i.e., differential accuracy) to the performance manipulation as a function of time of rating, number of stimuli, and/or

format. Due to interpretation problems, a four-way interaction was not given consideration. Given that the performance manipulation itself constitutes a contrast of "true scores" at the construct level, any differential ability of time of rating, number of stimuli, and/or format to detect the manipulation is, by definition, a measure of the relative accuracy of these variables as operational true scores. A significant main effect for the performance-level factor would indicate that raters detected the performance manipulation of the target dimension across conditions. Even though raters in the multiple-stimuli condition evaluated the performance of both the LOD and sales forecasting instructors, ratings assigned only to the LOD instructor were considered relevant to this study.

The results of the ANOVA for each manipulated performance dimension (see Table 1) indicated significant main effects of the performance-level factor for two of the three dimensions (i.e., Student-Teacher Interaction dimension, F (1,439) = 291.70, $p < .001$, and the Preparedness dimension, F (1,440) = 156.15, $p < .001$). Further attention will not be given to the Explanation of Concepts dimension because raters in this study were unable to detect the performance manipulation (i.e., no main effect or two-way or three-way interactions involving the performance-level variable), indicating little to no accuracy in their ratings of the LOD ratee's performance on this dimension.

The ANOVA results (see Table 1) also indicated no significant performance-level X time-of-rating interaction, suggesting that length of the rating delay had no consistent detrimental effects on rating accuracy for the Student-Teacher Interaction and Preparedness dimensions. Mean ratings for each performance dimension are shown in Table 2.

TABLE 1

Effects of Time of Rating, Stimuli, Format, & Level For Manipulated Dimensional Ratings

	Manipulated Performance Dimension								
	Explanation of Concepts			Student-Teacher Interaction			Preparedness		
Source of variation	df	MS	F	df	MS	F	df	MS	F
Performance level (A)	1	1.63	.54	1	717.00	291.70***	1	570.29	156.15***
Time of rating (B)	5	3.83	1.28	5	10.38	4.22**	5	9.31	2.55*
Number of stimuli (C)	1	22.80	7.63	1	51.73	21.05***	1	27.30	7.47**
Format (D)	3	4.50	15.75	3	17.37	56.72***	3	3.25	12.18***
A X B	5	2.71	.91	5	4.15	1.69	5	1.33	.36
A X C	1	5.07	1.69	1	10.15	4.13*	1	5.57	1.53
A X D	3	.21	.73	3	9.81	32.02***	3	.98	3.66*
B X C	5	1.56	.52	5	3.72	1.51	5	.64	.18
B X D	15	.19	.68	15	.45	1.48	15	.18	.66
C X D	3	.47	1.63	3	.94	3.07*	3	.47	1.76
A X B X C	5	2.03	.68	5	3.90	1.59	5	1.47	.40
A X B X D	15	.11	.40	15	.61	1.98*	15	.38	1.42
A X C X D	3	.12	.41	3	.56	1.83	3	.13	.48
B X C X D	15	.26	.92	15	.26	.85	15	.42	1.59
A X B X C X D	15	.52	1.80*	15	.31	1.01	15	.13	.50

Note. $N = 462$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

TABLE 2

Mean Ratings of Manipulated Performance Dimensions For Level, Time of Rating, & Format

Dimension/ Time of Rating	Low Performance Means					High Performance Means				
	BOS	BARS	GRBEH	GRS	N	BOS	BARS	GRBEH	GRS	N
Explanation of Concepts:										
No delay	2.78	2.90	2.49	2.62	42	2.98	3.05	2.94	2.83	40
7-day delay	2.94	3.13	2.95	2.92	37	2.95	2.96	2.78	2.82	38
14-day delay	3.13	3.24	2.96	2.91	35	2.95	2.95	2.81	2.68	41
21-day delay	2.88	2.92	2.78	2.65	40	2.98	3.08	2.91	2.79	38
28-day delay	2.99	3.12	2.99	2.97	36	3.30	3.25	3.21	3.10	40
35-day delay	2.96	3.16	3.00	2.97	36	3.01	3.25	3.14	3.05	40
Entire sample	2.94	3.07	2.85	2.83	226	3.03	3.09	2.96	2.88	237
Student-Teacher Interaction:										
No delay	1.26	2.18	1.38	2.00	40	3.43	3.44	3.17	3.32	41
7-day delay	1.65	2.21	1.82	2.00	36	3.53	3.49	3.18	3.44	39
14-day delay	1.81	2.56	2.00	2.17	41	3.33	3.25	2.90	3.14	36
21-day delay	1.91	2.42	2.00	1.87	39	3.08	3.37	2.96	3.05	39
28-day delay	1.98	2.66	2.26	2.71	38	3.72	3.73	3.43	3.55	38
35-day delay	1.84	2.79	2.07	2.29	34	3.37	3.44	3.23	3.43	42
Entire sample	1.74	2.46	1.92	2.17	228	3.41	3.45	3.15	3.32	235
Preparedness:										
No delay	2.41	2.47	2.22	2.27	41	3.50	3.29	3.11	3.29	41
7-day delay	2.44	2.59	2.36	2.28	39	3.54	3.45	3.24	3.47	36
14-day delay	2.38	2.56	2.28	2.42	41	3.81	3.63	3.71	3.72	36
21-day delay	2.31	2.49	2.19	2.15	39	3.62	3.48	3.26	3.39	39
28-day delay	2.74	2.74	2.74	2.71	41	3.79	3.85	3.63	3.83	35
35-day delay	2.65	2.50	2.62	2.42	38	3.72	3.92	3.54	3.71	38
Entire sample	2.49	2.56	2.40	2.38	239	3.66	3.60	3.40	3.56	225

Note. Greater spread between high & low performance levels indicates higher accuracy.

As shown in Table 1, a significant performance-level X format X delay interaction, F (15,1317) = 1.98, $p < .05$, and a performance-level X format interaction, F (3,1317) = 32.02, $p < .001$, were obtained for the Student-Teacher Interaction performance dimension. According to Cohen and Cohen (1983), a three-way interaction subsumes a two-way interaction. To determine more precisely the nature of this three-way interaction, the four formats were analyzed by pairs in similar performance-level X format ANOVAs for each time-of-rating condition. Table 3 reports the relevant differential accuracy interactions. These performance-level X format interactions in the different time-of-rating conditions (as interpreted by the spread in the means in Table 2) indicate that ratings made with BOS were more accurate than ratings made with the other three formats in the no-delay, seven-day delay, 14-day delay, and 28-day delay conditions. That is, the spread in BOS means is greater than the mean spread for the three other formats in these particular time-of-rating conditions, suggesting higher differential accuracy for BOS. However, no format differences were detected in the 21-day delay condition. In the 35-day delay condition, both BOS and GRS were found to have equally superior accuracy over the other two formats.

TABLE 3

Level by Format (Differential Accuracy) Interactions in Each Time-of-Rating Condition

	Time of Rating											
	No delay		7-day delay		14-day delay		21-day delay		28-day delay		35-day delay	
Format	df	F	df	F	df	F	df	F	df	F	df	F
BOS/ BARS	1,78	30.00***	1,73	15.40***	1,75	21.70***	1,76	1.40	1,74	21.80***	1,75	21.70***
BOS/ GRS	1,79	18.81***	1,73	7.07**	1,75	7.90**	1,76	.01	1,74	16.50***	1,75	2.80
BOS/ GRBEH	1,80	7.50**	1,73	14.00***	1,75	17.40***	1,76	2.15	1,74	20.10***	1,74	5.00*
GRS/ GRBEH	1,79	6.01*	1,73	.16	1,75	.15	1,76	1.54	1,74	2.45	1,74	.01
GRS/ BARS	1,79	.10	1,73	1.07	1,75	2.00	1,76	1.37	1,74	1.27	1,75	3.95*
BARS/ GRBEH	1,80	11.58***	1,73	.24	1,75	1.62	1,76	.00	1,74	.39	1,74	7.19**

* $p < .05$. ** $p < .01$. *** $p < .001$.

A significant performance-level X format interaction, F (3,1320) = 3.66, $p < .05$, was also obtained for the Preparedness dimension. To determine more precisely the nature of this two-way interaction, the four formats were analyzed by pairs in similar performance level X format ANOVAs. Table 4 reports the relevant differential accuracy interactions. The performance level X format interactions (as interpreted by the means in Table 2) show that BOS and GRS were more sensitive than BARS and GRBEH to the performance-level manipulation for this Preparedness dimension. That is, the spread in BOS and GRS means is greater than the spread in the means for the other two formats, suggesting higher differential accuracy for BOS and GRS on this dimension. Table 4 also shows that a significant level-by-format interaction was not detected when BOS and GRS ratings were contrasted, suggesting equivalent rating accuracy for these two formats on this dimension.

TABLE 4

Performance Level By Format (Differential Accuracy) Interactions

	Preparedness		
Formats	df	MS	F
BOS/BARS	1,463	1.06	4.14*
BOS/GRBEH	1,462	1.65	7.57**
BOS/GRS	1,463	.01	.02
BARS/GRBEH	1,462	.07	.25
BARS/GRS	1,463	1.23	3.85*
GRBEH/GR	1,462	1.89	7.39**

* $p < .05$. ** $p < .01$.

As shown in Table 1, the ANOVA results indicated a significant performance level X number-of-stimuli interaction for the Student-Teacher Interaction dimension, F (1,439) = 4.13, $p < .05$. Examination of the means and t-values (see Table 5) indicates that observation and evaluation of one target ratee produced more accurate ratings on this dimension than did the observation and evaluation of two ratees. That is, both the bigger spread in means and the higher t-value suggest higher differential accuracy for raters in the single-stimulus condition over raters in the multiple-stimuli condition.

TABLE 5

Descriptive Statistics of Student-Teacher Interaction Ratings For Number of Stimuli

	Performance Level						
	Low Performance			High Performance			
Number of stimuli	M	SD	N	M	SD	N	t
Student-Teacher Interaction:							
One target stimulus	2.18	.78	115	3.57	.81	119	13.42*
Two target stimuli	1.98	.79	114	3.09	.85	117	10.31*

* $p < .001$.

As mentioned, the Objectivity performance dimension was held constant (i.e., consistently high performance across the eight LOD videotapes). Thus, the higher the mean rating on this dimension would be indicative of higher differential accuracy. A 2 (number of stimuli) X 6 (time of rating) X 4 (format) within-subjects ANOVA was conducted for this non-manipulated dimension. As shown in Table 6, the ANOVA results indicated no significant main effects for time of rating, suggesting that length of the rating delays had no consistent detrimental influence on rating accuracy for this dimension. On the other hand, significant main effects for both number of stimuli, F (1,451) = 6.08, $p < .01$, and format, F (3,1353) = 17.86, $p < .001$, were found. The mean Objectivity rating for raters observing one target stimulus is 3.84 (SD = .80) whereas the mean Objectivity rating for raters observing two target stimuli is 3.64 (SD = .85). The significant difference in means suggests that raters observing one target stimulus evaluated this ratee more accurately on this dimension than did raters observing two stimuli. Furthermore, Table 7 shows that BOS and BARS means on the Objectivity dimension are higher than the means for the other two formats, suggesting higher differential accuracy for both BOS and BARS on this dimension. Planned paired t-test comparisons (see Table 7) revealed that ratings made with BOS and BARS on this dimension were significantly more accurate than were ratings made with the other two formats.

TABLE 6

Effects of Time, Stimuli, Format, & Performance Level For Non-Manipulated Dimensional Rating & Perceived Rating Accuracy

Source of variation	Objectivity			Perceived Rating Accuracy		
	df	MS	F	df	MS	F
Time of Rating (A)	5	2.06	.86	5	1.79	.97
Number of Stimuli (B)	1	18.81	6.08*	1	.50	.27
Format (C)	3	5.14	17.86**	3	2.10	4.09*
A X B	5	2.81	1.73	5	1.61	.88
A X C	15	.30	1.04	15	.49	.95
B X C	3	.64	2.22	3	.15	.30
A X B X C	15	.45	1.57	15	.39	.75

Note. $N = 466$.
* $p < .01$. ** $p < .001$.

TABLE 7

Mean Ratings of Non-Manipulated Performance Dimension & of Perceived Rating Accuracy For Time of Rating & Format

Dimension & perceived accuracy /Time of rating	BOS	BARS	GRBEH	GRS	N
Objectivity:[a]					
No delay	3.81	3.65	3.68	3.66	82
7-day delay	3.91	3.81	3.60	3.56	75
14-day delay	3.88	3.77	3.69	3.74	76
21-day delay	3.71	3.69	3.52	3.55	78
28-day delay	3.86	3.96	3.72	3.66	76
35-day delay	3.98	3.94	3.78	3.70	76
Entire sample	3.86[x]	3.80[x]	3.67[y]	3.64[y]	463
Perceived rating accuracy:[b]					
No delay	3.42	3.11	3.33	3.10	82
7-day delay	3.24	3.07	3.05	3.07	75
14-day delay	3.05	3.04	3.04	3.12	77
21-day delay	3.17	3.08	3.04	3.08	78
28-day delay	3.28	3.31	3.14	3.08	74
35-day delay	3.18	3.05	3.05	2.97	76
Entire sample	3.23[v]	3.11[w]	3.11[w]	3.07[v]	462

Note. Higher mean rating indicates higher rating accuracy for Objectivity dimension and higher perceived rating accuracy.

[a] Values with different superscript (x or y) are significantly different at $p < .001$.

[b] Values with different superscript (v or w) are significantly different at $p < .05$. Values with different superscript (x or y) are significantly different at $p < .01$.

To assess perceived rating accuracy, a 6 (time of rating) X 2 (number of stimuli) X 4 (format) within-subjects ANOVA was conducted. As shown in Table 6, the ANOVA results indicated significant main effects of the format factor only, F (3,1350) = 4.09, $p < .01$. Table 7 shows the mean ratings for the format factor in each time-of-rating condition. Planned comparisons were conducted to determine more precisely which format(s) produced the highest level of perceived accuracy across time-of-rating conditions. The results of these paired t-tests (see Table 7) revealed that BOS had significantly higher ratings on the perceived accuracy measure over the other three formats (i.e., BOS > BES = GRS = GRBEH).

■ DISCUSSION

The primary purpose of this study was to revisit the influence of rating delays on rating accuracy. Another interest was to determine whether rating delays have similar ill-effects on ratings made with different formats and on ratings of one versus two observed stimuli. Time delays of seven, 14, 21, 28, and 35 days were incorporated into this study. Contrary to previous findings (Heneman & Wexley, 1983; Rush et al., 1981) and the notions expressed in three recent literature reviews (Bretz et al., 1992; Heneman et al., 1987; Ilgen et al., 1993), this study's findings revealed that rating accuracy on two of the

The results of this particular study suggest that the ill-effects of increased time delays between observation and judgment may be overstated, especially when the performance dimensions and related behaviors are clearly delineated. Considering the recall ability of raters in the four and five-week delay conditions for a brief videotaped lecture, it may be necessary to rethink the importance of cognitive processing variables on ratings. Motivational influences loom as a more important determinant of performance ratings. On the other hand, Cardy and Dobbins (1994) pointed out that cognitive processing studies like this one, which test models, are worthwhile as long as they move our performance appraisal thinking towards more veridical ratings. To this end, it seems that this study has been successful.

The subjects in this study were college students. Therefore, the issue of this study's external validity should be considered. One might argue against the use of college students as raters in performance appraisal research because they lack real-world experience in rating employees. Nathan and Lord's (1983) counter-argument was that college students use the same cognitive processes and are susceptible to the same cognitive biases as those occurring in real-world appraisal settings. Not withstanding, the appraisal literature should be richer when variables present in this study are tested in the applied setting.

■ REFERENCES

Barnes-Farrell, J.L., & Couture, K.A. (1984). Effects of appraisal salience on immediate and memory-based judgments (Tech. Rep. No. 84-1). Honolulu: University of Hawaii, Department of Psychology.

Borman, W.C. (1977). Consistency of rating accuracy and rater errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Bretz, R.D., Milkovich, G.T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, direction, and implications. Journal of Management, 18, 321-352.

Cardy, R.L., & Dobbins, G.H. (1994). Performance appraisal: Alternative perspectives. Cincinnati, OH: South-Western Publishing Co.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/ correlation analysis for the behavioral sciences. Hillside, NJ: Erlbaum.

Cronbach, L.J. (1955). Processes affecting scores in understanding of others and assuming "similarity." Psychological Bulletin, 52, 177-193.

Dossett, D.L. (1989). Dimensional characteristics of behaviorally based performance ratings. Journal of Management Systems, 1, 51-66.

Dossett, D.L., & Costigan, R.D. (1987, April). Illusory halo and accuracy in performance ratings. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Hartel, C.E.J. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. Journal of Applied Psychology, 78, 212-217.

Hastorf, A.H., Schneider, D.J., & Polefka, J. (1970). Person perception. Reading, MA: Addison-Wesley.

Heneman, R.L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Heneman, R.L., & Wexley, K.N. (1983). The effects of time delay in rating and amount of information observed on performance rating accuracy. Academy of Management Journal, 26, 677-686.

Heneman, R.L., Wexley, K.N., & Moore, M.L. (1987). Performance-rating accuracy: A critical review. Journal of Business Research, 15, 431-448.

Hoffman, C.C., & Dossett, D.L. (1982, August). Development and format considerations for behavioral rating scales. Paper presented at the meeting of the American Psychological Association. Washington, D.C.

Ilgen, D.R., Barnes-Farrell, J.L., & McKellin, D.B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? Organizational Behavior and Human Decision Processes, 54, 321-368.

Landy, F.J., & Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

three manipulated dimensions and on one non-manipulated dimension was for the most part unaffected by the lengthy time delays between observation and judgment. More specifically, subjects rated performance on the Student-Teacher Interaction, Preparedness, and Objectivity dimensions in a fairly consistent fashion across the six time-of-rating conditions. The one 3-way interaction for the Student-Teacher Interaction dimension, involving the performance-level and time-of-rating factors, may reflect more of the idiosyncratic rating tendencies of the subjects in the 21-day delay condition than actual memory loss. Recall that BOS ratings of subjects in both the 28-day and 35-day delay conditions return to a more respectable level of accuracy.

From a practical perspective, the results of this study suggest that rating accuracy on certain dimensions may be satisfactory, even with relatively long delays between observation and judgment. Even though raters in this study did not encounter time delays that managers sometimes face (e.g., 35-day delay versus several months delay between observation and judgment), the cognitive processing prowess of this study's raters provide additional hope for veridical ratings in the applied setting.

Subjects were unable to detect the performance manipulation of the Explanation of Concepts dimension, suggesting little to no accuracy for this dimension. Perhaps, this difference in rating accuracy between dimensions suggests that certain dimensions and associated behaviors lend themselves to deeper processing and excellent recall. The behaviors associated with the Student-Teacher Interaction dimension (e.g., "Uses names of students when addressing them in class") may be more salient and/or discrete in comparison to the behaviors associated with the Explanation of Concepts dimension (e.g., "Explains definitions clearly and completely"). Likewise, the behaviors associated with the Preparedness dimension (e.g., "Has examples to use prior to coming to class") appear quite discernible. These results speak to the importance of the proper construction (i.e., clear definitions) of the content of the rating instrument.

This study's findings indicate that the ratings in the single-stimulus condition were better than ratings in the multiple-stimuli condition for the Student-Teacher Interaction and Objectivity dimensions. Single-stimulus raters were probably able to focus their undivided attention on the performance of a single ratee. Introducing a second rating stimulus apparently did not lessen rating accuracy on the Preparedness dimension. The use of certain formats (e.g., GRS) over others (i.e., BOS) may also be a contributor to rating inaccuracy. Caution is needed to avoid exaggerating the importance of the format and number-of-stimuli variables in this study. There are differences in the statistical power for this study's variables, with the time-of-rating factor at a clear disadvantage.

The results of this study, which found BOS ratings to be generally superior to the ratings of the other three formats, appear to support Latham and Wexley's (1977) claim for the effectiveness of BOS. The work of Murphy and Davidshofer (1988) drew attention to the weakness of dimensional formats such as GRS which require raters to recall relevant ratee performance information. The BOS format, on the other hand, may require more of a recognition task and may as a result be cognitively less taxing for the rater.

Cardy and Dobbins (1994) and Miner (1988) concluded that relative rating formats, having anchors such as "One of the best employees = 5," ... , "One of the worst employees = 1," produce better accuracy than do absolute formats such as GRS, GRBEH, and BARS. Heneman (1986) reported in a review of 23 studies that the correlation between supervisory ratings and performance measures was stronger when ratings were made with relative formats than absolute formats. This study's results suggest that the superiority of relative formats over absolute formats may generalize to another kind of relative format, BOS. Behavioral items in BOS are anchored with relative frequency descriptors (i.e., "never," "always"), which differ from the evaluative anchors (e.g., "poor," "excellent") in GRS, GRBEH, and BARS. It should not be overlooked that raters in this study perceived their ratings to be more accurate with BOS relative to the other three formats.

Lathan, G.P., & Wexley, K.N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.

Lord, R.G. (1985). An informational processing approach to social perceptions, leadership and behavioral measurement in organizations. In L.L. Cummings and M.B. Staw (Eds.), Research in Organizational Behavior (Vol. 7, pp. 87-128). Greenwich, CT: JAI Press.

Miner, J.B. (1988). Development and application of the rated ranking technique in performance appraisal. Journal of Occupational Psychology, 61, 291-305.

Murphy, K.R., & Balzer, W.K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Murphy, K.R., & Cleveland, J.N. (1995). Understanding performance appraisal: Social, Organizational, and goal-based perspectives. Thousand Oaks, CA: Sage Publications.

Murphy, K.R., & Constans, J.I. (1987). Behavioral anchors as a source of bias in rating. Journal of Applied Psychology, 72, 573-579.

Murphy, K.R., & Davidshofer, C.O. (1988). Psychological testing: Principles and applications. Englewood Cliffs, NJ: Prentice Hall.

Murphy, K.R., Philbin, T.A., & Adams, S.R. (1989). Effect of purpose of observation on accuracy of immediate and delayed performance ratings. Organizational Behavior and Human Decision Processes, 43, 336-354.

Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Rush, M.C., Phillips, J.S., & Lord, R.G. (1981). The effects of a temporal delay in rating on leader behavior descriptions: A laboratory investigation. Journal of Applied Psychology, 66, 442-450.

Smith, P.C., & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Smither, J.W., & Reilly, R.R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. Organizational Behavior and Human Decision Processes, 40, 369-391.

Tulving, E. (1983). Essentials of Episodic Memory. New York: Oxford University Press.

Wyer, R.S., Jr., & Srull, T.K. (1981). Category accessibility: Some theoretical and empirical issues concerning the processing of social stimulus information. In E.T. Higgins, C.P. Herman, & M.P. Zanna (Eds.), Social cognition: The Ontario Symposium on Personality and Social Psychology. Hillsdale, NJ: Lawrence Erlbaum.

■ ABOUT THE AUTHOR

***Robert D. Costigan** is an Associate Professor of Management at St. John Fisher College in Rochester, New York. He has a Ph.D. in Industrial-Organizational Psychology from the University of Missouri-St. Louis. His research interests include performance appraisal, selection, international human resource management, and organizational trust.*