Journal of Applied Psychology
1995, Vol. 80, No. 4, 517-524

Effects of Rater Accountability on the Accuracy and the Favorability of Performance Ratings

Neal P. Mero
U.S. Air Force Academy

Stephan J. Motowidlo
University of Florida

The authors tested the effects of holding raters accountable for their performance ratings on the accuracy and the favorability of those ratings. Undergraduate research participants ($N = 247$) completed an inbasket exercise and observed a videotaped simulation during 2 sessions over a 2-week period. The simulation presented performance information on 4 simulated subordinates portrayed through videotaped vignettes. True performance scores were manipulated by varying the proportion of positive and negative performance vignettes presented for each subordinate. Participants who were made to feel accountable by having to justify their ratings to the experimenter in writing rated their simulated subordinates more accurately. In another experimental condition, accountable raters who were told their subordinates' previous performance ratings were too low rated their subordinates more favorably than did raters in the same experimental condition who were not accountable.

Although there is a great deal of research on factors that affect the accuracy of performance ratings, little progress has been made in actually improving rater accuracy (Bernardin & Beatty, 1984; DeNisi & Williams, 1988; Landy & Farr, 1980; Murphy & Cleveland, 1991). This has led to a call for research focusing on the context surrounding the rating process (Ilgen & Feldman, 1983; Murphy & Cleveland, 1991). By including contextual variables in their studies of performance appraisal, researchers should be able to consider how an organization's social system influences the quality of performance ratings. Murphy and Cleveland suggested that little can be accomplished by changing the rater or the rater's task if the context influences the rater to be inaccurate. Research reported here considers how holding raters accountable for their rating decisions influences the quality of their performance ratings in different motivational contexts.

Wherry (1952) proposed that requiring raters to justify their ratings would affect the way they collect information, recall it, and make their ratings. This proposition received some theoretical consideration in the performance appraisal literature but little empirical consideration.

Recently, however, an extensive literature has developed around similar ideas in other decision-making contexts. This literature argues that when accountable decision makers know their audience's views, they make decisions that are consistent with those views (Klimoski & Inks, 1990; Tetlock, 1985), because as cognitive misers (Taylor & Fiske, 1978), people prefer decision-making strategies that involve the least effort. This means that accountable decision makers who know the views of their audience will take the least effortful path by making decisions they think will be acceptable to that audience (Tetlock, 1985).

There is empirical support for this position. Several studies have supported the influence of the need for approval (Jones & Wortman, 1977; Wortman & Linsenmeier, 1977) and the motivation of individuals to present themselves as positively as possible to those to whom they are accountable (Baumeister, 1982; Schlenker, 1980). Tetlock and colleagues (Tetlock, 1983; Tetlock, Skitka, & Boettger, 1989) showed that accountable participants who knew the views of their audience relied on a low-effort acceptability heuristic and shifted their views to-

Neal P. Mero, Department of Management, U.S. Air Force Academy; Stephan J. Motowidlo, Department of Management, University of Florida.

This article is based on Neal P. Mero's doctoral dissertation.

We gratefully acknowledge Barry Schlenker for the guidance and expertise he contributed to this project. We also acknowledge the contributions of Steve Werner, Jeff Katz, Mary Jo Vaughan, Jennifer Burnett, Kevin Banning, Tim DeGroot, and Cheryl Mero to the development and the administration of the managerial simulation used in this study.

The opinions expressed herein are those of the authors and are not necessarily those of any organization of the federal government.

Correspondence concerning this article should be addressed to Neal P. Mero, Headquarters U.S. Air Force Academy/DFM, 2354 Fairchild Drive, Suite 6H94, U.S. Air Force Academy, Colorado Springs, Colorado 80840-5701.

ward those of the audience. In other words, they tended to provide decisions they felt would be more acceptable to their audience.

This acceptability heuristic also leads to predictions about how decision makers are likely to process information when they do not know their audience's views. Tetlock (1985) proposed that under those conditions, accountable decision makers use more complex decision-making strategies, are more sure of their own cognitive processes, and more systematically base their decisions on the available data.

There is also empirical support for this possibility. Tetlock (1983) reported that decision makers who were not aware of the views of their audience used a preemptive, self-critical strategy by developing counterarguments to potential critics of their decisions. Simonson and Nye (1992) found that accountable participants used a more multidimensional and self-critical information-processing strategy. Ashton (1992) found that auditors who were required to justify their ratings were more accurate and consistent on a task of predicting bond ratings. Tetlock and Kim (1987) found that accountable decision makers formed more complex impressions and made more accurate predictions.

In the context of performance appraisal, accountability can be viewed as a motivating force on the rater. Consistent with accountability theories (Tetlock, 1985), when raters are held accountable for their rating decisions, they should consider the views of those to whom they are accountable. Performance-rating systems in organizations can require raters to justify their rating decisions to a variety of organizational constituencies, including supervisors, subordinates, and researchers (Mohrman & Lawler, 1983; Murphy & Cleveland, 1991). As a result, raters often take into account the views and the attitudes of these constituencies as they approach the rating process (Longenecker, Sims, & Gioia, 1987). In addition to information about the performance of the individual they are evaluating, raters have to be sensitive to many sources of information about the organizational environment. Longenecker et al. suggested that the typical performance appraisal context does not necessarily support an objective or accurate evaluation of subordinates' performance. Executives in their study reported deliberately manipulating ratings to meet their specific objectives or to comply with other organizational pressures. These objectives or pressures form the "motivational context" of performance evaluation.

One important element of the motivational context of performance appraisal is the purpose of the appraisal. Raters are likely to acquire and process performance information differently on the basis of the purpose of the evaluation (Williams, DeNisi, Blencoe, & Cafferty, 1985). In addition, if raters are convinced their ratings have important consequences, they will probably observe and evaluate performance behaviors more carefully (Murphy, Balzer, Kellam, & Armstrong, 1984).

Studies on the effects of purpose on performance ratings have yielded inconsistent results. Empirical evidence (Klimoski & Inks, 1990; Longenecker et al., 1987; Waldman & Thornton, 1988) suggests that effects of rating purpose on performance ratings may be moderated by whether there are personal implications for the rater or the ratee. Because participants in studies of effects of purpose are often guaranteed anonymity, interpersonal consequences in these studies are probably not always particularly salient (Murphy et al., 1984). However, differences between studies in the salience of interpersonal consequences of ratings for the rater might account for inconsistencies in research on leniency as a function of purpose.

The research reported here expands on earlier research by exploring the possibility that raters might be more influenced by motivating contextual factors such as purpose when performance ratings have personal implications for them. One way to heighten personal implications for raters is to require them to justify their ratings. Requiring justification essentially makes raters accountable and should cause them to wonder how they might be affected by the ratings they make. This is the point at which feelings of accountability and cues from the motivational context merge to influence performance ratings.

This study considers several forms of motivational contexts for performance evaluation. One is simply a context in which there are no special pressures on raters to achieve a certain outcome. Raters who are held accountable to supervisors for their ratings in such a motivational context do not know their supervisors' views of what the performance ratings should be. Results of studies of accountability effects suggest that raters in that situation rely on more complex decision-making strategies. This should lead to more accurate performance ratings for three reasons. First, if they have other tasks to perform besides evaluating subordinates' performance, the justification requirement should make the performance appraisal task more salient and cause them to devote more attention to it. Second, participants who have to justify their decisions should process performance information in a way that focuses their attention on the most relevant information (Tetlock, 1985). Third, within the setting of performance appraisal, accountability should improve rating judgment by increasing the consistency with which raters process multiple pieces of performance information (Ashton, 1992).

Our first hypothesis was that raters who are held accountable for their ratings in a motivational context in which there are no special pressures to achieve a certain rating outcome will rate more accurately than raters in

the same motivational context who are not held accountable for their ratings.

Our second hypothesis involved motivational contexts that do exert special pressures to achieve certain rating outcomes. Accountable raters in these situations will feel the personal implications of their ratings more acutely than nonaccountable raters and should be more motivated to avoid personal consequences that might be aversive for them. The aversive consequence in this case would be the embarrassment of being unable to justify their ratings when required. Raters will want to avoid this embarrassment by being sure to make their ratings in a way they can justify. Ratings should be easiest to justify when they conform to the pressure of the motivational context. Thus, our second hypothesis was that accountable raters will rate more consistently with the specific pressures of their motivational context than will nonaccountable raters. The present study considers three motivational contexts, each with a different form of rating pressure. In one motivational context, raters were specifically urged to rate accurately, and we expected that accountable raters in that context would rate more accurately than nonaccountable raters. In another motivational context, raters were specifically urged to rate more leniently, and we expected that accountable raters would rate more leniently than nonaccountable raters. In the third motivational context, raters were specifically urged to rate women more leniently, and we expected that accountable raters would inflate their ratings of women more than nonaccountable raters would.

Method

Laboratory studies of performance evaluation have been criticized for failing to consider the complexity of actual performance evaluation environments (Ebbesen & Konecni, 1980; Funder, 1987; Ilgen & Favero, 1985). Funder (1987) suggested that "research must let subjects judge real people in real social contexts, and use realistic external criteria for determining when the judgments are right or wrong" (p. 76). Another criticism by Ebbesen and Konecni argued that people may not make judgments in the real world the same way they do in the laboratory. Many of these criticisms were directed at studies that created "paper people" or provided only limited representations of relevant ratee performance. Ilgen and Favero criticized this paper-people paradigm and suggested that procedures that incorporate actual observation of performance directly or through videotaped vignettes were more likely to yield useful results.

Methods used in this study addressed these concerns in several ways. First, participants were exposed to subordinates' performance information during two 2-hr sessions over a 2-week period. Consequently, they observed performance information over an extended period with some opportunity for decay in memory between the first and second experimental sessions. Second, participants worked on a complicated inbasket simulation that embedded them in a realistic supervisory position in which evaluating subordinates' performance was only one of many tasks that demanded their attention. Third, subordinates' behavior was depicted on videotape. This gave participants a chance to observe their simulated subordinates in a wide range of contexts in which their subordinates performed many kinds of behaviors, some relevant for evaluations of their job performance and some irrelevant. The videotaped portrayals also provided different forms of information about job performance, such as direct observation of subordinates' behaviors, samples of their written work, and reports about subordinates' performance from others.

Managerial Simulation

Overview. An inbasket exercise simulated administrative aspects of a manager's job. Inbasket materials put participants in the role of Leslie Wilder, a divisional manager in a federal purchasing agency, and required them to resolve organizational problems; deal with personnel issues; develop policies; participate in special projects; and handle communications from their supervisor, other divisional managers, customers, and subordinates. As Leslie, the participants had five subordinates who were lower level managers with their own supervisory responsibilities.

Written inbasket materials served as a backdrop for performance information about Leslie's subordinates. This performance information was presented through videotaped vignettes. As participants worked on the inbasket materials, they were periodically interrupted by a television monitor showing subordinates entering Leslie's office to report information or other scenes in which subordinates participated in meetings. Just before each scene was presented, it was introduced on the monitor by means of an intercom call from Leslie's secretary, who announced that someone had asked to see Leslie and was about to enter the office, or by means of a brief narrative. The narrative explained that Leslie was now leaving the office to attend a meeting with subordinates in another part of the building, as the camera panned what Leslie saw on the way to the meeting. These videotaped vignettes portrayed various levels of performance for each subordinate on each of several dimensions.

Performance vignettes. We developed videotaped episodes showing high and low levels of performance on several dimensions for several subordinates. Starting with behavioral definitions for performance dimensions of administration, supervision, work effort, and coordination and negotiation, we created 4 performance situations for each of four subordinates (two men and two women) and each of the four performance dimensions. (We did not create performance episodes for the fifth subordinate but used him only to facilitate presentation of some personnel problems in the simulation.) This amounted to 64 situations in which we could present a subordinate performing either well or poorly. We prepared two performance scripts for each situation. One presented the subordinate doing something that represented high performance on a particular dimension, and the other showed the same subordinate doing something that represented low performance on the same dimension. Accordingly, we could show the same subordinate performing either well or poorly in the same performance situation.

We developed the 64 high-performance episodes in an effort to have each one represent a performance level of approxi-

mately 6 on a 7-point scale, with 1 representing less effective behaviors and 7 representing more effective behaviors. Similarly, we developed the 64 low-performance episodes in an effort to have each one represent a performance level of approximately 2. Our assumption was that the true score for a subordinate's performance could be estimated reasonably well as the average of the performance levels of all performance episodes presented for the subordinate on a particular dimension.

There were eight performance vignettes for each subordinate in each performance dimension. Only as many as four could be presented in any administration of this simulation, however, because it made no sense to repeat a subordinate's performance in exactly the same situation, presenting him or her as effective in one episode and ineffective in another. Therefore, the four that we presented could include all ineffective episodes (for an estimated true score of 2), three ineffective episodes and one effective episode (for an estimated true score of 3), two ineffective and two effective episodes (for an estimated true score of 4), one ineffective episode and three effective episodes (for an estimated true score of 5), or all effective episodes (for an estimated true score of 6).

Doctoral students acted out the roles of the four subordinates whose performance scores could be manipulated, and we recorded them on videotape.

We conducted a preliminary study to test our assumption that it was reasonable to estimate true scores as approximately 6 for the high-performance vignettes and approximately 2 for the low-performance vignettes. Twenty judges with prior management and rating experience were allowed repeated observations of each performance vignette and were asked to rate each one as an independent source of performance information. Each vignette was observed and rated by five expert judges. For the 64 vignettes presenting high performance, judges' mean ratings ranged from 4 to 7 with an overall mean of 6.18 (mean SD = 0.60). For the 64 vignettes presenting low performance, judges' mean ratings ranged from 1 to 4 with an overall mean of 1.69 (mean SD = 0.57). These values fell close enough to our intended true scores to confirm our assumption that 6 was a reasonable estimate of true scores for high-performance vignettes and 2 was a reasonable estimate of true scores for low-performance vignettes.

An alternative approach might have been to use the judges' ratings as estimates of true scores. Because only five judges rated each vignette, however, their means were not stable enough to inspire confidence that they approximated the true scores any better than our intended true scores did. Our intended true scores had the advantage of computational simplicity and consistency because they were not subject to fluctuations across different samples of judges.

Experimental Performance Displays

We selected 48 performance vignettes for use in this study from the total of 128 that were available. They included 4 vignettes for each of three performance dimensions (supervision, effort, and coordination and negotiation) for each of four subordinates. We selected them to vary performance scores as much as possible across dimensions and subordinates but to still have the average performance for each dimension across subordinates constant. Table 1 shows the number of high- and low-performance vignettes chosen for each dimension separately for each subordinate. It also shows their estimated true dimension scores with the assumption that high-performance vignettes had a true score of 6, low-performance vignettes had a true score of 2, and the true dimension score was the average of true scores for all 4 vignettes representing that dimension.

Experimental Sample and Procedure

Undergraduates (N = 247) who were enrolled in an introductory course in management participated in this research for course credit. The sample included 136 men and 111 women with a mean age of 21 years. Men and women were randomly distributed among eight treatment conditions to control for possible gender effects. There were 16–18 men and 13–15 women in each condition.

Students were assigned to one of four motivational contexts. In three of these conditions, they received information about the effects of previous performance-rating decisions on organizational members through a letter from their simulated supervisor just before they started working on the inbasket and through other letters and memos that were introduced in the inbasket itself. In the inflationary context condition, as Leslie, participants were informed that their subordinates had been rated consistently lower by their previous supervisor than their peers in other divisions. As a result, no one from Leslie's division was promoted in the last 5 years. In the accuracy context condition, Leslie was informed that performance ratings throughout the organization were inflated to the point that performance differences between employees could not be discerned. This made it impossible to use performance evaluations for administrative purposes. Leslie was encouraged to provide ratings that accurately reflected true differences of subordinates within each performance dimension. In the equitable treatment condition, Leslie was informed that women in the organization had been rated consistently lower than men. Leslie's supervisor expressed concern about a pending lawsuit over the low ratings

Table 1
Construction of Performance Displays According to the Number of High- and Low-Performance Vignettes by Dimension and Subordinate

	Subordinates			
Performance dimension	Alice	Bill	Carole	David
Effort				
Number of high vignettes	1	2	3	4
Number of low vignettes	3	2	1	0
Estimated true score	3	4	5	6
Coordination and negotiation				
Number of high vignettes	3	4	1	2
Number of low vignettes	1	0	3	2
Estimated true score	5	6	3	4
Supervision				
Number of high vignettes	2	4	1	3
Number of low vignettes	2	0	3	1
Estimated true score	4	6	3	5

given to women. The fourth motivational context provided no information at all about previous performance ratings. This baseline condition was designed to simulate a motivational context in which there was no particular pressure to rate one way or the other, except for the implicit pressure to rate conscientiously.

Two conditions of accountability, accountable and not accountable, were manipulated by varying a written assignment that was due from the participants at the end of the experiment. At the beginning of the experiment, accountable participants were told their written assignment was to justify their performance evaluations to their supervisors (the researchers). Nonaccountable participants were told that their ratings would remain anonymous and that their written assignment was to critique the simulation. All participants were told that researchers could award up to five points of extra credit for their participation on the basis of their performance in the simulation and on the written assignment. About half the students in each motivational context were assigned to the accountability condition, and half were assigned to the nonaccountable condition.

In the first experimental session, participants were introduced to the inbasket exercise and worked on it for 2 hr while approximately half of the performance vignettes were presented. In the second session, which was scheduled the following week, they worked on the second part of the inbasket for 2 more hr while the rest of the performance vignettes were presented. All participants saw the same set of performance vignettes.

At the beginning of the experiment, participants were told they would be required to rate the performance of their subordinates and were given descriptions of the performance dimensions. At the end of the experiment, they rated performance using a 7-point behavioral rating scale for each of the three performance dimensions. Scale values ranged from 1 to 7. Behavioral examples anchored high, medium, and low ranges of the scales. Scale ratings were used to form three dependent variables.

One dependent variable was Cronbach's (1955) index of differential accuracy, which was used to assess the accuracy of participants' ratings. Differential accuracy measures ratee by dimension interaction and indicates how well raters can differentiate between different levels of ratee performance on different performance dimensions. This measure was appropriate because our performance displays specifically built in differences between performance dimensions separately for each subordinate. There was also some variability between subordinates. Because the main focus in this study was on differences between dimensions within subordinates, however, differential accuracy was a more appropriate index than differential elevation to test differences between accountable and nonaccountable raters in the accuracy and baseline motivational contexts. Lower values of the differential accuracy index represent higher levels of accuracy.

Mean performance rating was the second dependent variable. It was computed as the mean across dimensions and subordinates for each participant in the inflationary motivational context to test differences between accountable and nonaccountable raters.

The third dependent variable was the mean female rating. It was computed as the mean of ratings across dimensions and across the two female subordinates to test differences in the equitable treatment condition between accountable and nonaccountable raters.

Manipulation Checks

We measured three other variables to test whether participants in the accountability conditions responded differently to the managerial simulation. They were based on accountability theory, which posits that accountable decision makers process information more carefully to prepare for the justification requirement. The three variables were attentiveness, note-taking, and engagement.

The attentiveness variable reflected how attentive participants were to the presentation of performance information. Two judges, naive to the participants' treatment condition, observed while participants worked on the managerial simulation and rated their attentiveness on a 3-point scale on the basis of the level of alertness and interest they showed when viewing performance vignettes. Judges based their ratings on cues such as head position, expression, posture, and note-taking reaction. One judge observed and rated in the first experimental session, and the other judge observed and rated in the second experimental session. The two judges' ratings were summed to form the overall attentiveness score. Intraclass correlation, adjusted according to the Spearman–Brown formula, yielded a reliability estimate of .67 for the two judges combined.

Participants were instructed that if they wished, they could take notes on their subordinates' job performance. They did not know they would not be allowed to refer to their notes when making their performance ratings. The note-taking variable measured the quality and the quantity of notes that participants took. Notes from both sessions were provided to two judges naive to participants' treatment condition. Using a 5-point scale ranging from 1 (*low*) to 5 (*high*), judges independently rated the quality and the quantity of notes taken on performance-related information. Their ratings were summed to form the note-taking score. Intraclass correlation, adjusted by the Spearman–Brown formula, yielded a reliability estimate of .92 for the two judges combined.

The third manipulation check, engagement, was a self-report measure indicating how engaged participants felt when participating in the simulation. The measure consisted of three items that asked how much time they spent thinking about the specific challenges presented in the simulation, whether they discussed the simulation with others between sessions, and whether they debated between alternative responses to problems presented. The sum of the three items formed the engagement score ($\alpha = .69$).

Results

Differences between accountable and nonaccountable raters on manipulation checks and the results of independent sample *t*-test procedures are shown in Table 2. Accountable raters attended more to performance information, according to visible cues displayed while they completed the managerial simulation; took more and better

Table 2
Differences Between Accountable and Nonaccountable Raters on Manipulation Checks

Manipulation check	Accountable raters		Nonaccountable raters	
	M	*SD*	*M*	*SD*
Attentiveness	4.08_a	1.09	3.65_b	1.16
Note-taking	5.46_a	1.95	4.11_b	1.79
Engagement	11.25_a	2.29	10.25_b	2.10

Note. Within rows, means with different subscripts differ significantly ($p < .05$, one-tailed).

notes about their simulated subordinates' job performance; and reported being more engaged in the simulation. These results offer strong support for the construct validity of the accountability manipulation in this experiment.

Consistent with the directional predictions, the independent sample *t* tests used to test hypotheses were one-tailed and were considered significant when the probability of a Type I error for the resulting *t* value was less than or equal to .05. The first hypothesis predicted that raters who were held accountable for their ratings in a motivational context in which there were no special pressures to achieve a certain rating outcome would rate more accurately than raters in the same motivational context who were not held accountable for their ratings. The results presented in Table 3 support this prediction. Differential accuracy scores for accountable raters in the baseline motivational context show they were more accurate ($M = 1.00$, $SD = 0.74$) than nonaccountable raters ($M = 1.34$, $SD = 0.65$), $t(58) = 1.84$, $p < .05$. Lower values of the differential accuracy index represent higher levels of accuracy. This difference represents an effect size (d) of 0.46 *SD* using the formula provided by Cohen (1988).

The second hypothesis predicted that accountable raters would rate according to the specific pressures of their motivational context more than nonaccountable raters would. Results of the *t*-test comparison shown in Table 3 support this prediction in two of the three motivational contexts in which it was tested. In the motivational context that urged accuracy, accountable raters rated more accurately ($M = 0.91$, $SD = 0.53$) than nonaccountable raters ($M = 1.23$, $SD = 0.63$), $t(61) = 2.24$, $p < .05$, $d = 0.51$ *SD*. In the motivational context that urged inflated ratings, accountable raters rated more leniently ($M = 5.00$, $SD = 0.40$) than nonaccountable raters ($M = 4.70$, $SD = 0.39$), $t(61) = 2.99$, $p < .05$, $d = 0.75$ *SD*. In the motivational context that urged higher ratings for women, however, Hypothesis 2 was not supported. There was no indication that accountable raters rated women more favorably. If anything, accountable raters might have rated women less favorably ($M = 3.73$, $SD = 0.64$) than nonaccountable raters did ($M = 4.30$, $SD = 0.74$), but one-tailed tests preclude an interpretation of this effect because it was not in the expected direction.

Exploratory analyses were conducted to allow complete consideration of the effects of different motivational contexts within the two conditions of accountability as found in Hypothesis 2. Means and standard deviations for each dependent variable within each motivational context are shown in Table 4. We conducted three 2 × 3 (Accountability × Motivational Context) analyses of variance on differential accuracy, mean rating, and mean female ratings. For differential accuracy, there was a significant effect of accountability, $F(1, 247) = 13.46$, $p < .01$. There was also an overall effect of motivational context, $F(2, 247) = 3.46$, $p < .05$. The interaction was not significant, $F(2, 247) = 0.006$, *ns*. For the overall mean rating, there was also an interaction between motivational context and accountability, $F(2, 247) = 5.37$, $p < .01$. A subsequent Duncan's multiple-range test found that accountable raters in the inflationary condition and nonaccountable raters in the equitable treatment condition rated their subordinates significantly more favorably than raters in all other conditions ($p < .05$). For mean female ratings, there was an interaction between motiva-

Table 3
Mean Scores on Dependent Variables Testing Differences Between Accountable and Nonaccountable Raters Separately for Each Motivational Context

Motivational context	Dependent variable	Accountable raters		Nonaccountable raters	
		M	*SD*	*M*	*SD*
Baseline	Differential accuracy	1.00_a	0.74	1.34_b	0.65
Accuracy	Differential accuracy	0.91_a	0.53	1.23_b	0.63
Inflationary	Mean rating	5.00_a	0.40	4.70_b	0.39
Equitable treatment	Mean female rating	3.73	0.64	4.30	0.74

Note. Within rows, means with different subscripts differ significantly ($p < .05$, one-tailed).

Table 4
Means and Standard Deviations of Dependent Variables by Motivational Context

		Context							
		Baseline		Accuracy		Inflationary		Equitable treatment	
Dependent variable	Accountable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Differential accuracy	Yes	1.00	0.74	0.91	0.53	1.03	0.55	1.20	0.51
	No	1.34	0.65	1.23	0.63	1.38	0.64	1.52	0.80
Mean rating	Yes	4.70	0.46	4.80	0.45	5.00	0.40	4.80	0.43
	No	4.70	0.41	4.70	0.58	4.70	0.39	5.00	0.50
Mean female rating	Yes	3.70	0.70	3.90	0.68	4.10	0.71	3.73	0.64
	No	3.80	0.51	3.60	0.73	3.70	0.57	4.30	0.74

tional context and accountability, $F(2, 247) = 8.70$, $p < .01$. A subsequent Duncan's multiple-range test found significant differences between the ratings of nonaccountable raters in the equitable treatment group and all other conditions except accountable raters in the inflationary group ($p < .05$).

Discussion

Raters who were held accountable for their performance ratings made more accurate ratings than raters who were not held accountable. This was supported by the results of Hypothesis 1 and the analysis of variance that showed that across a variety of motivational rating contexts, accountable raters more accurately evaluated performance than nonaccountable raters. This finding is consistent with theoretical expectations that raters who are held accountable will approach the rating task in a way that will make it easier for them to account for their ratings. Results of manipulation checks showed that accountable raters exhibited behaviors that should lead to better performance evaluations. Accountable raters were more attentive, took more notes, and were more engaged in the simulation than nonaccountable raters.

Accountable raters were also more sensitive to motivational context in two of the three situations. Accountable participants in both the inflationary and accuracy contexts complied more with situational pressures and provided ratings more consistent with those pressures. However, accountable participants in the equitable treatment condition did not comply with situational pressures as expected, perhaps because inflating the ratings of only female subordinates would obscure real performance differences among ratees. Also, although the context encouraged raters to avoid discriminating against female subordinates, accountable raters could comply with this by just rating female subordinates accurately. Inflating the rating for females would have led to performance ratings that would be hard to justify given the performance they were shown. It is interesting to note that nonaccountable raters in this condition did provide inflated ratings for their female subordinates. Perhaps this is due to the emphasis placed in undergraduate management courses on ensuring equitable treatment of all subordinates. Nonaccountable raters complied with contextual pressures to artificially inflate ratings they would not be required to defend.

Results of this study have several practical implications. First, if subsequent research replicates the finding that accountability leads to more accurate ratings, designers of performance appraisal systems should consider incorporating a justification requirement. Although many current systems require raters to provide feedback to the ratee, fewer systems require raters to justify rating decisions to their supervisors. Although the concept of "true score" as used in this study does not apply to a field rating environment, our results are generalizable to rating contexts where the rater's supervisor is reasonably aware of the actual performance of the ratee. However, as shown in Hypothesis 2, when the supervisor has goals for the performance rating other than accuracy, accountable raters may comply with pressure to achieve that goal. Second, when performance appraisal results are used as input to important personnel decisions, making raters accountable by requiring them to justify their rating decisions in terms of behavioral dimensions relevant to the decision should improve decision quality. Accountability may lead raters to make performance appraisals a higher priority and may reduce their reliance on irrelevant factors.

This research also has implications for the positive effects of accountability. Consistent with Tetlock's (1985) proposal that accountable decision makers would use more complex decision-making strategies, accountable raters in this study exhibited behaviors that suggested a more active and engaged process of gathering information and of considering the implications of that information.

This study answers the call for research focusing on

contextual variables rather than focusing solely on the rater. It provides empirical support for the proposition that rating quality is related to conditions of the rating context. As a result, it supports the argument that rating quality is affected by raters' willingness to give quality ratings in addition to their ability to give quality ratings (Banks & Murphy, 1985).

References

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes, 51,* 416–446.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research–practice gap in performance appraisal. *Personnel Psychology, 38,* 335–345.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin, 91,* 3–26.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Belmont, CA: Wadsworth.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed "similarity." *Psychological Bulletin, 52,* 177–193.

DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resource management* (Vol. 6, pp. 109–155). Greenwich, CT: JAI Press.

Ebbesen, E. B., & Konecni, V. J. (1980). On the external validity of decision-making research: What do we know about decision-making in the real world? In T. S. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 21–45). Hillsdale, NJ: Erlbaum.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10,* 311–321.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141–197). Greenwich, CT: JAI Press.

Jones, E. E., & Wortman, C. (1977). *Ingratiation: An attributional approach.* Morristown, NJ: General Learning Press.

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes, 48,* 70–88.

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

Longenecker, C. O., Sims, H. P., Jr., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1,* 183–193.

Mohrman, A. M., & Lawler, E. E. (1983). Motivation and performance appraisal behavior. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 173–189). Hillsdale, NJ: Erlbaum.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. *Journal of Educational Psychology, 76,* 45–54.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective.* Boston: Allyn & Bacon.

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations.* Monterey, CA: Brooks/Cole.

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51,* 416–446.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attributions: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). New York: Academic Press.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45,* 74–83.

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. In B. M. Staw & L. Cummings (Eds.), *Research in organizational behavior* (Vol. 7, pp. 297–332). Greenwich, CT: JAI Press.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52,* 700–709.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57,* 632–640.

Waldman, D. A., & Thornton, G. C., III (1988). A field study of rating conditions and leniency in performance appraisal. *Psychological Reports, 63,* 835–840.

Wherry, R. J. (1952). *The control of bias in ratings: A theory of ratings.* Columbus: The Ohio State Research Foundation.

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Performance, 35,* 314–339.

Wortman, C. B., & Linsenmeier, J. A. (1977). Interpersonal attraction and techniques of ingratiation in organizational settings. In B. M. Staw & G. Salancik (Eds.), *New directions in organizational behavior* (pp. 133–178). Chicago: St. Clair's Press.

Received May 12, 1994
Revision received February 8, 1995
Accepted February 9, 1995 ■