Journal of Applied Psychology
2010, Vol. 95, No. 3, 546–561

0021-9010/10/$12.00 DOI: 10.1037/a0018866

The Roles of Rater Goals and Ratee Performance Levels in the Distortion of Performance Ratings

Xiaoye May Wang and Kin Fai Ellick Wong
Hong Kong University of Science and Technology

Jessica Y. Y. Kwong
The Chinese University of Hong Kong

The goal-directed perspective of performance appraisal suggests that raters with different goals will give different ratings. Considering the performance level as an important contextual factor, we conducted 2 studies in a peer rating context and in a nonpeer rating context and found that raters do use different rating tactics to achieve specific goals. Raters inflated their peer ratings under the harmony, fairness, and motivating goal conditions (Study 1, $N = 103$). More important, raters inflated their ratings more for low performers than for high and medium performers. In a nonpeer rating context, raters deflated ratings for high performers to achieve the fairness goal, and they inflated ratings for low performers to motivate them (Study 2, $N = 120$).

Keywords: performance evaluation, performance appraisal, rater goals, leniency, performance level

Rating (in)accuracy is one of the most important concerns in performance evaluation (Cronbach, 1955; Murphy & Cleveland, 1995). Researchers have traditionally conceptualized rating inaccuracy as the unwitting result of rating errors. From a psychometric perspective, rating errors are understood to be the results of the rating stimuli that do not trigger reliable and valid responses (Cronbach, 1955). From a cognitive perspective, rating errors are conceptualized to be the result of the limitations of human cognition (DeNisi, 1996), such as memory accessibility (Murphy & Balzer, 1986), cognitive style (Cardy & Kehoe, 1984; Härtel, 1993), and affect (Cardy & Dobbins, 1986). These approaches generally assume that raters involuntarily commit rating errors owing to either poor scale designs or to their own cognitive limitations.

More recently, researchers have begun to consider whether raters may intentionally distort ratings. That is, raters "have specific (and possibly multiple) goals in mind and they intend to provide ratings that are consistent with these goals" (Murphy, Cleveland, Skattebo, & Kinney, 2004, p. 158). Thus, departing from the traditional psychometric and cognitive perspectives, the goal-based approach to performance evaluation conceptualizes that a part of rating inaccuracy is indeed not related to rating error; rather, it is intentionally introduced by the rater to achieve specific goals in organizational contexts (Cleveland & Murphy, 1992; Murphy & Cleveland, 1991, 1995; Murphy et al., 2004; Wong & Kwong, 2007). For example, raters pursuing a harmony goal will increase their mean ratings and will decrease their rating discriminability, and raters pursuing a fairness goal will inflate their mean ratings and decrease the rating discriminability only at the early stage of group building (Wong & Kwong, 2007). These studies suggest that performance evaluation is not just a measurement process, but it is also a social process and a communication process. In other words, raters are not passive participants in the process but are active participants with the ability and motivation to distort ratings intentionally to attain predetermined goals.

Past research has demonstrated convincingly the main effects of rater goals on ratings and has shown that raters who pursue different goals give different ratings (Murphy et al., 2004; Wong & Kwong, 2007). Here, we suggest that studying the main effects of rater goals per se is incomplete because to achieve a goal, raters can have different patterns of ratings according to the characteristics of the ratees (e.g., their performance level, personality, gender). A complete picture of the goal-based approach of performance evaluation should consider raters' goals in conjunction with ratees' characteristics. Essentially, the main effect of raters' goals should be dependent on ratees' characteristics.

Indeed, in discussing the effects of rater goals on rating discriminability, Wong and Kwong (2007) suggested a rationale pertaining to the Rater Goals × Ratee Performance Levels interaction by reasoning that the same goal leads raters to rate differently depending on the performance levels of the ratees. For example, to achieve harmony within groups, raters tend to inflate their ratings. However, "the inflation for good performers may be less than that for poor performers" (Wong & Kwong, 2007, p. 578). Similarly, with respect to motivating ratees, raters "may avoid providing feedback . . . to the good performers . . . by deflating their ratings" and "may minimize the difference between the feedback given to poor performers and the performance standard by inflating ratings

Xiaoye May Wang and Kin Fai Ellick Wong, Department of Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; Jessica Y. Y. Kwong, Department of Marketing, The Chinese University of Hong Kong, Sha Tin, Hong Kong.

Preparation of this article was facilitated by Research Grant Council General Research Grant HKUST644108 and Research Grant Council Earmarked Research Grant HKUST645305H awarded to Kin Fai Ellick Wong and by Research Grant Council General Research Grant CUHK445007 awarded to Jessica Y. Y. Kwong. We are grateful to Virginia Unkefer, Vasavi Dasgupta, and Chong (Alex) Wang for their helpful comments and suggestions on an earlier version of this article.

Correspondence concerning this article should be addressed to Xiaoye May Wang, who is now at the Department of Human Resources and Organizational Behavior, School of Economics and Management, Tsinghua University, Beijing 100084, China. E-mail: wangxiaoye@sem.tsinghua.edu.cn

of poor performers" (Wong & Kwong, 2007, p. 579). Thus, performance ratings are expected to vary as a function of the Rater Goals × Ratee Performance interaction because the influences of rater goals on performance ratings cannot be examined in isolation from ratee characteristics.

The present research contributes beyond previous research (Murphy et al., 2004; Wong & Kwong, 2007) by highlighting the view that raters intentionally distort their ratings not only according to the goals they want to achieve but also by taking ratees' characteristics into consideration. We report two studies that directly examine this general proposition by looking into the interaction between rater goals and ratee performance levels on performance ratings. In addition, our study also makes empirical contributions by complementing a methodological limitation in Wong and Kwong's (2007) study. Specifically, rater goals were manipulated within-participants in Wong and Kwong's study. Although there are several positive features in using the within-participants design (e.g., it is statistically powerful and controls for individual differences), Wong and Kwong acknowledged that this design "could have introduced an experimental demand" in that participants "could have been thinking that they needed to respond somewhat differently to different conditions" (p. 584). We attempt to address this concern by manipulating rater goals between-participants in Study 2.

There are several reasons why performance level is one of the most important characteristics of ratees that needs to be examined. First, as mentioned above, previous research has assumed that raters are likely to have different levels of rating inflation for ratees with different performance levels (Wong & Kwong, 2007). This reasoning, however, has not yet been formally tested. We note that Wong and Kwong (2007) only showed that rater goals determined rating discriminability in general, indicating that different goals may lead to more or less differentiation among the ratings. However, they did not examine how the rater goals effects may be further shaped by ratee performance (i.e., the interaction between rater goals and ratee performance). For example, Wong and Kwong showed that the *harmony goal* reduced rating discriminability. The authors suggested that this reduction is due to a greater rating inflation for low performers than for high performers. This reduction in rating discriminability, however, could also be due to other possibilities, such as rating deflation for high performers with inflation for low performers or rating inflation for low performers together with no change for high performers. To address this ambiguity, in the present study we attempt to examine the effects of rater goals on ratees with different performance levels.

Second, we examine the ratees' performance level because one of the major purposes of performance evaluation is to discriminate among employees who perform at various levels. Such a differentiation is relevant to (a) major personnel decisions, such as promotion and salary increases (Murphy & Cleveland, 1995), and (b) rating validity in terms of differential elevation (Cronbach, 1955). Thus, it is important to examine how various levels of performers are evaluated differently according to different goal conditions.

Third, we examine the ratees' performance level because it is an important contextual factor shaping rating behaviors (Gaugler & Rudolph, 1992; Wexley, 1972; Wexley, Sanders, & Yukl, 1973; Wong & Kwong, 2005). For example, research on contrast effects has shown that the performance of a target ratee is often contrasted with the performance of the preceding candidate (Maurer & Alexander, 1991; Wexley, Yukl, Kovacs, & Sanders, 1972). Research on numerical framing has shown that presenting performance information with a large or a small number will lead to systematic changes in perceived differences across ratees, and different numerical presentations of the tradeoff between performance attributes will engender preference reversal between two focal employees (Wong & Kwong, 2005). Studies on decoy effects have highlighted that the preference for two ratees changes with the addition of another person with inferior performance (Highhouse, 1996; Slaughter, Sinar, & Highhouse, 1999).

The Present Study and Hypotheses

The purpose of this study is to extend the work of Murphy et al. (2004) and Wong and Kwong (2007) in a context that allows us to understand the effects of rater goals on rating scores for ratees with different levels of performance. We conducted two studies to achieve our objective. In Study 1, we manipulated rater goals within participants and examined the effects of rater goals on ratees with different performance levels in a real peer rating context. In Study 2, we manipulated rater goals between participants. Participants were required to play the role of a supervisor and to rate employees from good or from poor performing groups shown in a video. The two studies complement each other in terms of design (within-participant and between-participant) and rating contexts (real peer evaluation and hypothetical supervisory ratings).

We examined the four most common rater goals in our studies, namely identification, harmony, fairness, and motivating (Murphy & Cleveland, 1995; Murphy et al., 2004; Wong & Kwong, 2007). A rater with an identification goal identifies the ratees' strengths and weaknesses. A rater with a harmony goal seeks to maintain group harmony and interpersonal relationships. A rater with a fairness goal seeks to reflect the accurate contribution of each team member. A rater with a motivating goal seeks to increase the future motivation of ratees. Following Wong and Kwong (2007), we used the identification goal condition as the control condition because a typical performance appraisal often requires raters to identify ratees' strengths and weaknesses for administrative decisions, such as promotions and appointments, and for developmental purposes, such as identifying training needs. The effects of various rater goals are captured by comparing the ratings in the other three goal conditions with the rating in the identification goal condition (the control condition). In the following paragraphs, we develop separate hypotheses for each of the three rater goal conditions.

Rater Goals × Ratee Performance on Performance Ratings

Harmony goal. To understand the effect of the harmony goal, it is helpful to analyze the rating patterns that raters believe can reduce conflict. It is generally believed that people with fewer rewards tend to be more unsatisfied and behave more antagonistically (Leventhal, Michaels, & Sanford, 1972). Accordingly, as low performers are likely to get fewer rewards, raters tend to inflate their performance ratings to reduce their likelihood of exhibiting antagonistic behaviors. At the same time, raters are also aware that inflating the ratings of low performers will lead to

dissatisfaction among high performers because such inflation violates the equity norm (Leventhal et al., 1972). Raters, therefore, will try to solve the problem by inflating the ratings of the high performers as well.

However, the inflation of high performers' ratings will not be as great as that of the ratings of the low performers for two reasons. First, high performers are often already close to the top of a rating scale. There will be little room for rating inflation for high performers, but this is not the case for low performers. Second, problems resulting from conflict are more severe for low performers because they receive fewer rewards. This is consistent with the general belief that egalitarianism or the rule of equality can prevent group conflict (Deutsch, 1975, 1985; Marin, 1981). Therefore, we hypothesize the following:

> *Hypothesis 1:* Raters will exhibit rating inflation when they pursue a harmony goal (vs. an identification goal), with the inflation being more pronounced as the performance level of the ratee decreases.

Fairness goal. Following Wong and Kwong (2007, p. 578), we define the fairness goal as "give performance appraisals that fairly and accurately reflected individuals' contributions to their groups." However, this definition does not necessarily imply that raters will always adopt the norm of equity to reflect accuracy for two reasons. First, previous fairness research suggests the perception of accuracy could be very subjective. For example, in an experimental study, Molm, Takahashi, and Peterson (2003) showed that even in an objectively fair exchange relationship, participants will be likely to perceive unfairness and to feel they are being taken advantage of. Brockner, Wiesenfeld, and Martin (1995) also showed that objectively identical outcomes could be perceived to be more or less fair depending on the framing of the outcomes. Second, raters may not seek accuracy through the equity norm. For example, Wong and Kwong (2007) suggested that raters are more likely to be concerned with equality when evaluating in-group ratees and more likely to be concerned with equity when evaluating out-group ratees.

Given the strong subjectivity of the fairness perception (Brockner et al., 1995; Molm et al., 2003), and to understand the effect of the fairness goal, it is useful to start from the people's implicit theories of fairness, which are the raters' naïve beliefs about what indicates fairness. In this regard, equality and equity are the two most commonly used rules for fair resource allocation. The equality rule postulates that rewards are evenly allocated to all members, irrespective of individual inputs (Deutsch, 1985). On the other hand, the idea of the equity rule is that rewards are given according to individual contributions (Adams, 1965).

In performance evaluation, reliance on the equality rule implies that the raters would tend to give equal ratings to all ratees, irrespective of their contributions. In other words, when raters want to pursue a fairness goal and when they represent fairness as equality, they will inflate ratings for poorer performers and deflate ratings for better performers. On the other hand, reliance on the equity rule would essentially be the same as giving ratings to indicate individual contributions. This is conceptually similar to the identification goal of identifying ratees' strengths and weaknesses. Thus, when raters want to pursue a fairness goal and when they represent fairness as equity, they will give ratings in a way similar to that of the ratings under the control (identification) condition.

In the present research, we did not measure an individual's orientation toward equity versus equality. Instead, the equity versus equality orientation is reflected in whether the raters and ratees are from the same group. Specifically, people are more likely to adopt the equity rule for out-group members and to use equality rule for in-group members (Ng, 1984). Thus, we assumed that people tend to use equality to represent fairness when they evaluate ratees who are working with them in the same group, whereas they tend to use equity to represent fairness when they evaluate ratees who are not working in the same group.

> *Hypothesis 2:* In the context of pursuing fairness (vs. pursuing performance identification), raters will inflate ratings for poorer performers and deflate ratings for better performers when they evaluate performers from the same working group; however, raters will have no such rating distortion when they evaluate performers as out-group members.

Motivating goal. To understand the effect of the motivating goal on rating scores for various levels of performers, we start by examining the raters' implicit theory of motivation, which is the raters' naïve theory of how another person will be motivated. Drawing from the feedback intervention theory (FIT) of motivation (Kluger & DeNisi, 1996, p. 263), Wong and Kwong (2007) proposed that raters are likely to have different theories of motivation for low and high performers. Thus, "raters believe that different types of performance feedback are required to motivate good and low performers" (Wong & Kwong, 2007, p. 579).

Under the FIT framework, the theory that is more applicable to better performers (vs. poorer performers) pertains to the cybernetic mechanism of the negative feedback loop (e.g., Carver & Scheier, 1998; Powers, 1973). The cybernetic mechanism implies that people are motivated to reduce the discrepancy between a given performance standard and their current performance, suggesting that an increase in the discrepancy increases motivation. People reduce their effort on the same task when the discrepancy decreases because it allows them to allocate more effort to other tasks (Vancouver, Thompson, Tischner, & Putka, 2002). Better performers, by definition, are those who are approaching, have already reached, or have overshot a performance standard. Raters, therefore, are concerned with better performers being less motivated after knowing that they have already reached the standard. Accordingly, raters are likely to deflate the ratings of better performers. This logic, however, is not applicable to poorer performers because poorer performers, by definition, are farther away from a specific performance standard.

In addition, the FIT framework also suggests that the theory that is more applicable to poorer performers pertains to the concern of maintaining their positive self-evaluation (Seligman, 1975; Vroom, 1964) and self-efficacy (Bandura, 1997; Bandura & Locke, 2003). The key idea is that strongly negative feedback is demotivating because it is "perceived as a threat to the self" (Kluger & DeNisi, 1996, p. 267). In extreme cases, repeated negative feedback will lead to a total loss of motivation, as in the case of learned helplessness (Seligman, 1975). The concern for maintaining a positive self evaluation of a ratee is more applicable to poorer performers than to better performers. This is because a

rater would consider the low performance ratings of poorer performers to be threatening to their positive self evaluation. To reduce this concern, the rater then inflates the ratings of poorer performers. This logic, however, is not applicable to better performers because their good performance would not be considered to be threatening to their positive self evaluation.

> *Hypothesis 3:* Raters will inflate ratings of poorer performers and deflate ratings of better performers when they have a motivating goal (vs. an identification goal).

We conducted two studies to examine the effects of rater goals on the rating scores of ratees with different performance levels. In both studies, we examined the same four rater goals, that is, (a) identifying strengths and weaknesses (the control condition), (b) maintaining harmony, (c) seeking fairness, and (d) motivating future performance. In the first study, the rating context was peer ratings by students of the contributions of their team members to a team project, and the rater goal was treated as a within-subject factor. In the second study, raters evaluated the performances of six team members, who were shown in a video; the rater goal was treated as a between-subjects factor in this case. The two studies were complementary in both contexts and research designs.

Study 1

Method

Participants and procedure. Undergraduate students (N = 103) enrolled in a human resources management course participated in the study. Rater goals were manipulated within participants. The students were from 14 groups, each with seven or eight members. The groups were formed at the beginning of the semester to complete a project examining human resource practices in firms. Students formed the groups voluntarily, and the instructor assigned students randomly to groups if they had not joined a group. The project included a project proposal that had to be approved by the instructor by midsemester, an oral presentation, and a written report to be submitted at the end of the semester. Participants were required to complete peer evaluations twice, once at midsemester and the other at the end of the semester. The scores from the two peer evaluations constituted 20% of the group's final grade. All students were informed at the beginning of the semester that their peer rating scores would be part of their final grades. The peer evaluation outcomes were revealed to students 1 week after the evaluations.

Materials. Students gave ratings on a five-page questionnaire. The instruction on the first page stated the following:

> You are provided with four sets of peer evaluations. Each set requires you to complete the evaluation with a different goal in mind. That is, you need to do the peer evaluation in accordance with the goal mentioned on each page. You need to complete all of the evaluations.

Each following page represented a goal condition. In the identification goal (i.e., control) condition, students were asked to complete a peer evaluation that would help members to identify their strengths and weaknesses. In the harmony goal condition, students were required to complete a peer evaluation that would maintain interpersonal relationships and group harmony. In the fairness goal condition, students were instructed to complete a peer evaluation that reflected fairness and accuracy. In the motivating goal condition, students were reminded that their goal was to complete a peer evaluation that could motivate group members. We randomized the order of the goal conditions in the booklets; after the randomization, there were 24 sets of questionnaires. Students rated their peers on a scale ranging from 1 (*no contribution*) to 7 (*substantial contribution*).

We assessed the reliability of the ratings in two ways, following Wong and Kwong (2007). First, the same rater repeatedly gave performance ratings on the same ratees four times according to goal manipulations. The manipulations should therefore introduce systematic but not random errors. If the performance ratings were reliable, the ratings among the four conditions should be highly interrelated as they did not include a large amount of random errors. Thus, the reliability could be indicated by the internal consistency among ratings of the four conditions. The alphas were .86 and .93 for the midsemester and end of semester, respectively. Second, because each ratee was evaluated by more than one rater, the reliability of the ratings could also be assessed by interrater agreements as indicated by the Average Deviation Indices (Burke, Finkelstein, & Dusig, 1999). A deviation index smaller than 1.2 indicates good interrater agreement on a 7-point scale (Burke & Dunlap, 2002). Here, the Average Deviation Indices of all conditions were better than this cutoff point (all $< .9$), suggesting high interrater reliability on the performance ratings.

Results

Data analyses. We analyzed the data in two ways given two slightly different conceptualizations of ratees' performance levels. First, an absolute rating assumption suggests that a rater may have an absolute standard of performance while giving peer ratings to their group members. The performance of a particular member was compared with this absolute standard. In other words, the rater could consider all members within his/her group to be good, medium, or poor performers. Second, a relative rating assumption suggests that a rater may have a relative standard while giving his/her ratings. The performance of a particular member was compared with the performance of other members in the same group. Accordingly, there should be some members who are considered to be relatively good in performance and some who are considered to be relatively poor in performance.

Although raters in this study were asked to give absolute ratings, we were aware that raters often found it difficult to ignore between-individuals comparisons when giving absolute ratings (Wong & Kwong, 2005). Thus, we were not sure whether the absolute standard assumption or the relative standard assumption was more accurate in capturing the ratings in the current study. Accordingly, we ran two separate sets of analyses that were appropriate for each assumption. Converging the results from these two analyses would point to the robustness of the findings.

To capture the absolute standard assumption and given the nested structure of our data, we used hierarchical linear modeling (HLM) in the first analysis, with ratees' performance ratings in the control condition from a single rater being a continuous variable representing ratees' performance levels across groups. HLM explicitly accounts for the nested nature of the data and can simultaneously estimate the impact of factors at different levels on

individual-level outcomes while maintaining appropriate levels of analysis for the predictors (Bryk & Raudenbush, 1992).

Here, we used two-level models (HLM2), wherein rating scores constituted the Level 1 (L1) cases nested within Level 2 (L2) raters. That is, the outcome variable of L1 was the rating of a particular rater to a particular ratee under a specific rater goal condition, and the outcome variable of L2 was the intercept in L1. The random effect was only for the intercept. Although rating scores were also nested in the rater goal conditions, which means that the low-level units (rating scores) were cross-classified by two higher level units (rater and rater goal conditions), the number of rater goal conditions ($n = 4$) was too small to support an adequate examination of the cross-classified random effects models (HCM2 models; Raudenbush & Bryk, 2002). Thus, we explored the effects of rater goals by including rater goals dummy codes as L1 predictors of the dependent variables (Cohen, Cohen, West, & Aiken, 2003; Raudenbush & Bryk, 2002).

The ratee's performance level was entered as the predictor in L1. The ratee's performance level was operationalized as the rating on individual ratee performance from a particular rater under the control condition, where it was assumed to have the least distortion in the ratings. To examine the rating distortions for various levels of performers under each rater goal condition, we tested the interaction term of the ratees' performance levels and the rater goal conditions in the L1 model. We grand-mean centered the L1 predictors. This centering approach facilitates the interpretation of the HLM results and reduces the multicollinearity in the L2 estimation by reducing the correlation between the L2 intercept and the slope estimates (Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002). We used restricted maximum likelihood estimation (Raudenbush & Bryk, 2002) and reported the values based on the robust standard errors.

In the second analysis, we classified ratees within raters into low, medium, or high performers according to their ratings in the control condition to capture the relative rating assumption. We then tested the hypotheses using analyses of variance (ANOVAs) to examine the inflation and deflation of the ratings in each of the three performance levels. Specifically, in the ANOVA tests, the ratee performance level was regarded as a moderator. We differentiated between low, medium, and high performers within each rater by ranking ratings from the rater for each group member under the control condition. As each group contained seven or eight team members, the bottom two members and the top two members were classified to be low performers and high performers, respectively. The remaining members were classified to be medium performers. When there were ties within a group, averaged rank scores were assigned to each ratee. If the averaged rank was smaller than or equal to two, then the ratee was classified as a low performer; if the averaged rank was greater than five, then the ratee was classified as a high performer; the rest were classified as medium performers. For tie cases that could not be resolved by the above method, ratees' performance rankings were arbitrarily determined by the alphabetical order of their names. Of the ratees, 12.5% and 10.3% were randomly assigned into a different performer category at midsemester and at end of semester, respectively. We also used the 33% and 66% percentile scores as cutoffs to differentiate high, medium, and low performers, and the result patterns are the same as the results reported here using the ranking method.

Table 1 presents the means, standard deviations, and zero-order correlations of all variables used in the ANOVAs. Table 2 summarizes the HLM results, and Table 3 presents the ANOVA results. Figure 1 presents the slope plot of the Goal Condition × Ratee Performance Level interactions of the HLM analyses. As shown in Table 2, all the random effects (τ_{00}) were significant ($p < .01$). This indicates that different raters did give different rating means, suggesting that the random effects approach of the HLM was appropriate.

Tests of hypotheses.

Harmony goal. On the midsemester data, the HLM2 results (see Table 2, Model 4) indicated a significant relationship between the harmony goal and rating scores ($\gamma_{20} = .69$, $p < .01$). More important, the interaction between the harmony goal and ratees' performance levels was also significant ($\gamma_{50} = -.36$, $p < .01$). To further explore this interaction, we plotted rating scores at plus and minus one standard deviation above and below the sample mean of ratees' performance levels (Cohen et al., 2003). As shown in Figure 1 (left panel), the positive relationship between ratees' performance level and rating scores was attenuated under the harmony goal condition versus the control condition. Simple slope tests indicated that the relationship between ratee performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the harmony condition ($\beta = .63$, $p < .01$). This slope difference indicates that there is greater rating inflation as the performance level decreases, consistent with Hypothesis 1.

We found a similar pattern in the ANOVA results using the midsemester data. A 2 (goal condition: control vs. harmony) × 3 (performance level: low, medium, or high) repeated measures ANOVA revealed that there was a significant Goal Condition × Performance Level interaction, $F(2, 204) = 48.13$, $p < .001$, $\eta_p^2 = .32$. Planned comparisons revealed that the effect of the harmony goal condition was significant for all performance levels ($ps < .001$), indicating that raters inflated ratings for ratees at all performance levels. The significant interaction indicates that the inflation varied across performance levels. Consistent with Hypothesis 1, the inflation was the greatest for low performers (control condition = 4.03, $SD = 1.23$; harmony condition = 5.05, $SD = 1.29$; Cohen's $d = 0.81$), followed by medium performers (control condition = 4.61, $SD = 0.97$; harmony condition = 5.35, $SD = 1.09$; Cohen's $d = 0.72$), and it was the least for high performers (control condition = 5.25, $SD = 0.86$; harmony condition = 5.62, $SD = 0.95$; Cohen's $d = 0.41$).

On the end-of-semester data, the HLM2 results (see Table 2, Model 8) indicated a significant relationship between the harmony goal and rating scores ($\gamma_{20} = .56$, $p < .01$). More important, the interaction between the harmony goal and ratees' performance levels was significant ($\gamma_{50} = -.29$, $p < .01$). As shown in Figure 1 (right panel), the positive relationship between ratees' performance levels and rating scores was attenuated under the harmony goal condition versus the control condition. Simple slope tests indicated that the relationship between ratee performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the harmony condition ($\beta = .71$, $p < .01$). This slope difference indicates that there is greater rating inflation as the performance level decreases, consistent with Hypothesis 1.

We found a similar pattern in the ANOVA results using the end-of-semester data. A 2 (goal condition: control vs. harmony) × 3 (performance level: low, medium, or high) repeated measures

Table 1
Means, Standard Deviations, and Zero-Order Correlations Among All Variables Measured at Midsemester (T1) and at End of Semester (T2) in Study 1

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Mean rating (T1)																										
Control condition																										
1. High performer	5.25	0.86	—																							
2. Middle performer	4.61	0.97	.75	—																						
3. Low performer	4.03	1.23	.56	.85	—																					
Harmony condition																										
4. High performer	5.62	0.95	.55	.51	.40	—																				
5. Middle performer	5.35	1.09	.41	.60	.54	.83	—																			
6. Low performer	5.05	1.29	.33	.54	.67	.64	.84	—																		
Fairness condition																										
7. High performer	5.24	0.95	.71	.63	.50	.47	.48	.41	—																	
8. Middle performer	4.89	1.05	.63	.78	.72	.56	.71	.61	.79	—																
9. Low performer	4.54	1.36	.40	.67	.80	.37	.58	.68	.61	.82	—															
Motivating condition																										
10. High performer	5.23	1.03	.51	.49	.33	.26	.19	.06	.49	.38	.28	—														
11. Middle performer	4.82	1.23	.47	.68	.58	.37	.48	.38	.51	.58	.49	.81	—													
12. Low performer	4.44	1.31	.37	.61	.68	.29	.44	.49	.47	.54	.58	.68	.86	—												
Mean rating (T2)																										
Control condition																										
13. High performer	5.60	0.92	−.01	−.01	−.02	.01	.00	−.07	.09	.06	−.02	−.04	.04	.03	—											
14. Middle performer	4.95	1.11	−.04	.03	.03	.07	.11	.01	.06	.10	.04	.01	.13	.09	.79	—										
15. Low performer	4.06	1.44	−.03	.08	.12	.03	.13	.12	.09	.14	.17	.03	.12	.17	.46	.69	—									
Harmony condition																										
16. High performer	5.80	1.04	.06	.03	−.04	.09	.12	−.03	.14	.14	−.03	.05	.13	.12	.67	.59	.30	—								
17. Middle performer	5.71	0.99	.03	.03	−.07	.04	.09	−.07	.15	.13	−.01	.10	.15	.10	.58	.64	.39	.83	—							
18. Low performer	4.86	1.61	−.05	.02	.05	.03	.16	.10	.11	.18	.17	.04	.12	.21	.32	.45	.78	.47	.55	—						
Fairness condition																										
19. High performer	5.70	0.90	.05	.01	−.06	.08	.06	−.07	.20	.13	.02	.05	.11	.03	.81	.71	.39	.75	.70	.40	—					
20. Middle performer	5.26	1.14	.00	.04	.01	.11	.14	.02	.17	.20	.15	.09	.15	.07	.66	.80	.57	.57	.65	.50	.82	—				
21. Low performer	4.42	1.58	−.02	.02	.07	.06	.13	.13	.17	.17	.20	.00	.09	.14	.38	.57	.88	.34	.45	.83	.47	.65	—			
Motivating condition																										
22. High performer	5.64	0.92	.11	.05	−.04	−.01	−.06	−.16	.08	.04	−.09	.10	.10	.04	.69	.60	.33	.67	.58	.29	.67	.48	.26	—		
23. Middle performer	5.22	1.14	.07	.11	.04	.08	.10	−.01	.10	.17	.10	.17	.21	.15	.59	.73	.51	.62	.70	.48	.63	.67	.46	.75	—	
24. Low performer	4.43	1.58	.03	.06	.09	.00	.08	.04	.09	.14	.15	.07	.11	.19	.41	.55	.81	.40	.46	.84	.40	.49	.77	.47	.66	—

Note. $N = 103$, listwise deletion.
$|.19| = p < .05$. $|.26| = p < .01$ (two-tailed).

Table 2
Hierarchical Linear Modeling Results of the Interaction Effects of Rater Goals and Performance Level on Rating Scores in Study 1

	T1								T2							
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7		Model 8	
Variable	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Fixed effect																
Intercept, γ_{00}	4.95**	0.08	4.95**	0.04	4.67**	0.02	4.67**	0.00	5.17**	0.09	5.16**	0.04	4.91**	0.02	4.91**	0.00
Ratees' performance level, γ_{10}			0.75**	0.02	0.74**	0.02	0.95**	0.02			0.83**	0.02	0.83**	0.03	0.99**	0.02
Harmony goal, γ_{20}					0.69**	0.08	0.69**	0.08					0.56**	0.08	0.56**	0.07
Fairness goal, γ_{30}					0.25**	0.06	0.25**	0.06					0.24**	0.06	0.24**	0.06
Motivating goal, γ_{40}					0.18*	0.09	0.18*	0.08					0.20**	0.07	0.20**	0.07
Ratees' Performance Level × Harmony Goal, γ_{50}							−0.36**	0.05							−0.29**	0.04
Ratees' Performance Level × Fairness Goal, γ_{60}							−0.21**	0.04							−0.14**	0.03
Ratees' Performance Level × Motivating Goal, γ_{70}							−0.27**	0.05							−0.22**	0.05
Variance component	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}
	.91	.68**	.60	.12**	.53	.12**	.50	.13**	1.24	.79**	.54	.12**	.50	.12**	.49	.12**
N (Level 1)[a]	2,651		2,651		2,651		2,651		2,659		2,659		2,659		2,659	
N (Level 2)	103		103		103		103		103		103		103		103	

Note. T1 = midsemester; T2 = end of semester.
[a] *N* refers to the number of cases subjected to hierarchical linear modeling analyses after excluding missing cases. The actual *N* at Level 1 was 2,664 for both T1 and T2. There were 13 missing cases in T1 and 5 missing cases in T2.
* $p < .05$. ** $p < .01$.

Table 3
Analysis of Variance Results of the Interaction Effects of Rater Goals and Performance Level on Rating Scores in Study 1

Effect	T1			T2		
	F	*MSE*	η_p^2	*F*	*MSE*	η_p^2
Hypotheses 1						
Goal condition (harmony vs. control)	72.28**	1.09	.42	54.77**	0.94	.35
Performance level	79.81**	0.52	.44	75.44**	1.09	.43
Goal Condition × Performance Level	48.13**	0.11	.32	37.07**	0.16	.27
Hypotheses 2						
Goal condition (fairness vs. control)	17.98**	0.59	.15	17.83**	0.52	.15
Performance level	80.94**	0.59	.44	93.47**	1.10	.48
Goal Condition × Performance Level	31.91**	0.11	.24	9.58**	0.10	.09
Hypothesis 3						
Goal condition (motivating vs. control)	5.34**	1.16	.05	9.46**	0.73	.09
Performance level	98.98**	0.53	.49	94.41**	1.03	.48
Goal Condition × Performance Level	21.19**	0.11	.17	9.51**	0.17	.09

Note. $N = 103$. T1 = midsemester; T2 = end of semester.
** $p < .01$.

ANOVA revealed that there was a significant Goal Condition × Performance Level interaction, $F(2, 204) = 37.07$, $p < .01$, $\eta_p^2 = .27$. Planned comparisons revealed that the effect of the goal condition was significant at all performance levels ($ps < .01$), indicating rating inflation at all performance levels. The significant interaction effect indicates that the inflation varied across performance levels. The inflation for low performers (control condition = 4.06, $SD = 1.44$; harmony condition = 4.86, $SD = 1.61$; Cohen's $d = 0.52$) was comparable with that for medium performers (control condition = 4.95, $SD = 1.11$; harmony condition = 5.71, $SD = 0.99$; Cohen's $d = 0.72$), and both were significantly larger than that for high performers (control condition = 5.60, $SD = 0.92$; harmony condition = 5.80, $SD = 1.04$; Cohen's $d = 0.20$). In sum, Hypothesis 1 was supported by the midsemester and end-of-semester data in both the HLM analyses and the ANOVAs.

Fairness goal. On the midsemester peer ratings, the HLM2 results (see Table 2, Model 4) revealed a significant relationship between the fairness goal and rating scores ($\gamma_{30} = .25$, $p < .01$). More important, the interaction between the fairness goal and ratees' performance levels was also significant ($\gamma_{60} = -.21$, $p < .01$). Again, we plotted rating scores at plus and minus one standard deviation above and below the sample mean of ratees' performance levels. As shown in Figure 1 (left panel), the positive relationship between ratees' performance levels and the rating scores is reduced under the fairness goal condition compared with the control condition. Simple slope tests indicated that the relationship between ratees' performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the fairness goal condition ($\beta = .79$, $p < .01$).

A similar pattern was found in the ANOVA results with the midsemester data. A 2 (goal condition: control vs. harmony) × 3 (performance level: low, medium, or high) repeated measures ANOVA revealed that there was a significant Goal Condition × Performance Level interaction, $F(2, 204) = 31.91$, $p < .01$, $\eta_p^2 = .24$. The significant interaction indicates that the inflation varied across performance levels. Planned comparisons revealed that the effect of the goal condition was significant for low and medium performers ($ps < .001$) but not for high performers. That is, there

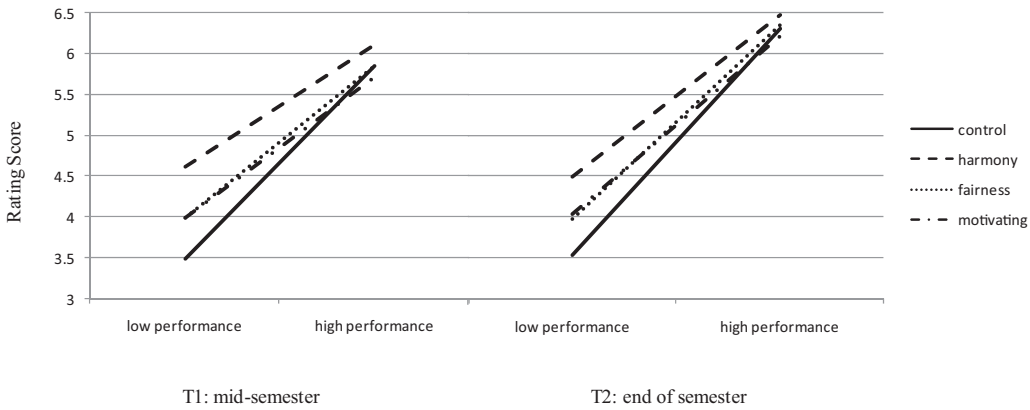

Figure 1. The performance rating as a function of goal condition and performance level at midsemester (T1) and at end of semester (T2) in Study 1.

was significant rating inflation for the low performers (control condition = 4.03, *SD* = 1.23; fairness condition = 4.54, *SD* = 1.36; Cohen's $d = 0.39$) and for the medium performers (control condition = 4.61, *SD* = 0.97; fairness condition = 4.89, *SD* = 1.05; Cohen's $d = 0.28$), but there was no significant inflation for the high performers (control condition = 5.25, *SD* = 0.86; fairness condition = 5.24, *SD* = 0.95; Cohen's $d = 0.01$).

On the end-of-semester peer ratings, the HLM2 results (see Table 2, Model 8) indicated a significant relationship between the fairness goal and rating scores ($\gamma_{30} = .24$, $p < .01$). More important, the interaction between the fairness goal and ratees' performance levels was significant ($\gamma_{50} = -.14$, $p < .01$). As shown in Figure 1 (right panel), the positive relationship between ratees' performance levels and the rating scores was reduced under the fairness goal condition compared with the control condition. Simple slope tests indicated that the relationship between ratees' performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the fairness condition ($\beta = .86$, $p < .01$).

A similar pattern was found in the ANOVA results with the end-of-semester data. A 2 (goal condition: control vs. harmony) × 3 (performance level: low, medium, or high) repeated measures ANOVA revealed that there was a significant Goal Condition × Performance Level interaction, $F(2, 204) = 9.58$, $p < .01$, $\eta_p^2 = .09$. The significant interaction indicates that the inflation varied across performance levels. Planned comparisons revealed that the effect of the goal condition was significant for low and medium performers (*p*s < .01) but not for high performers. That is, there was significant rating inflation for the low performers (control condition = 4.06, *SD* = 1.44; fairness condition = 4.42, *SD* = 1.58; Cohen's $d = 0.24$) and for the medium performers (control condition = 4.95, *SD* = 1.11; fairness condition = 5.26, *SD* = 1.14; Cohen's $d = 0.28$), but there was no significant rating inflation for the high performers (control condition = 5.60, *SD* = 0.92; fairness condition = 5.70, *SD* = 0.90; Cohen's $d = 0.11$).

These results are partly consistent with Hypothesis 2, which predicts that if raters use the equality norm to represent fairness, they would inflate their ratings for low performers while deflating the ratings for high performers. Our results indeed reveal rating inflation for poor performers, yet they do not show the expected deflation for good performers. We discuss the mixed strategy of using equity and equality norms to represent fairness in the Discussion section.

Motivating goal. On the midsemester peer ratings, the HLM2 results (see Table 2, Model 4) indicated a significant relationship between the motivating goal and rating scores ($\gamma_{40} = .18$, $p < .05$). More important, the interaction between the motivating goal and ratees' performance levels was also significant ($\gamma_{70} = -.27$, $p < .01$). As shown in the slope plot in Figure 1 (left panel), the positive relationship between ratees' performance levels and the rating scores was weaker under the motivating goal condition than under the control condition. Simple slope tests indicated that the relationship between ratees' performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the motivating condition ($\beta = .73$, $p < .01$).

The ANOVA showed a similar result. A 2 (goal condition: control vs. motivating) × 3 (performance level: low, medium, or high) repeated measures ANOVA revealed that the Goal Condition × Performance Level interaction was significant, $F(2, 204) = 21.19$, $p < .01$, $\eta_p^2 = .17$. The significant interaction effect indicates that raters inflated ratings differently among the three performance levels. Planned comparisons revealed that the effect of the goal condition was significant for low performers ($p < .01$) and medium performers ($p < .05$), but it was not significant for high performers ($p > .50$). Thus, raters inflated the most for low performers (control condition = 4.03, *SD* = 1.23; motivating condition = 4.44, *SD* = 1.31; Cohen's $d = 0.32$), followed by the medium performers (control condition = 4.61, *SD* = 0.97; motivating condition = 4.82, *SD* = 1.23; Cohen's $d = 0.19$), and did not significantly inflate the rating of high performers (control condition = 5.25, *SD* = 0.86; motivating condition = 5.23, *SD* = 1.03; Cohen's $d = 0.02$).

On the end-of-semester peer ratings, the HLM2 results (see Table 2, Model 8) indicated a significant relationship between the motivating goal and rating scores ($\gamma_{40} = .20$, $p < .01$). More important, the interaction between the motivating goal and ratees' performance levels was also significant ($\gamma_{70} = -.22$, $p < .01$). As shown in Figure 1 (right panel), the positive relationship between the ratees' performance levels and the rating scores was reduced under the motivating goal condition compared with the control condition. Simple slope tests indicated that the relationship between ratees' performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the harmony condition ($\beta = .78$, $p < .01$).

The ANOVA showed a similar result. A 2 (goal condition: control vs. fairness) × 3 (performance level: low, medium, or high) interaction was significant, $F(2, 204) = 9.51$, $p < .01$, $\eta_p^2 = .09$. The significant interaction effect indicates that raters inflated ratings differently among the three performance levels. Planned comparisons revealed that the effect of the goal condition was significant for low performers and medium performers ($p < .01$), but it was not significant for good performers ($p > .50$). Thus, raters inflated more for low performers (control condition = 4.06, *SD* = 1.44; motivating condition = 4.43, *SD* = 1.58; Cohen's $d = 0.24$) and for medium performers (control condition = 4.95, *SD* = 1.11; motivating condition = 5.22, *SD* = 1.14; Cohen's $d = 0.24$), and they did not significantly inflate the ratings of high performers (control condition = 5.60, *SD* = 0.92; motivating condition = 5.64, *SD* = 0.92; Cohen's $d = 0.04$).

These findings partially support Hypothesis 3 in that the raters inflated the ratings for low performers under the motivating goal condition. However, we found that the raters did not deflate ratings for high performers.

Variance explained. It is also interesting to examine how much of the variance in performance ratings was due to true performance, distortion of rating goals, and Rater Goals × Ratee Performance interactions. Here, we assume that ratings in the control condition represent the true rating of performance. We entered the main effects and interaction effects hierarchically (i.e., Step 1: intercept; Step 2: main effect of ratee performance; Step 3: main effects of rater goals; Step 4: interactions). In the midsemester ratings, the "true performance" rating accounted for 29.81% of the variance (in Step 2). Rater goals accounted for an additional 6.73% of the variance in Step 3. In Step 4, the interactions accounted for 2.88% of the variance. In addition, rater effect accounted for 12.50% of the variance, and 48.08% of the total variance is the random effect. In the end-of-semester ratings, the true performance rating accounted for 55.56% of the variance in Step 2. Rater goals accounted for an

additional 3.17% of the variance in Step 3. In Step 4, the interactions accounted for 0.79% of the variance. In addition, rater effect accounted for 9.52% of the variance, and 30.95% of the total variance is the random effect.

Discussion

The above results show that the raters had different distortions for ratees with various performance levels under different raters' goals conditions. Raters inflated ratings under the three goal conditions, which is consistent with Murphy et al.'s (2004) findings that the raters' goals have a positive effect on rating scores. More important, raters tended to inflate ratings more as performance levels decreased. Thus, in general, Study 1 supports the general proposition that to achieve a specific goal, raters "tailor" the degree of rating distortion to ratees with different performance levels.

Under the fairness condition, contrary to our prediction, raters did not deflate the ratings for high performers; rather, they statistically did not distort ratings for high performers. We suspect that under the fairness goal condition, raters may not just use one single norm of justice, that is, either equality or equity to represent fairness. They might use double standards of fairness, such as increasing the ratings for low performers, according to the equality norm to decrease the rating variance, while not distorting ratings for high performers according to the equity norm.

We attempted to further examine Hypothesis 2 by looking into the role of fairness representation (equity vs. equality) in ratings. To do so, we compared the Rater Goals × Ratee Performance interaction at midsemester and at end-of-semester. As mentioned earlier, raters were expected to be more equality oriented at midsemester (because raters and ratees were group members), and they were expected to be more equity oriented at end-of-semester (because raters and ratees were no longer group members). The Time (midsemester vs. end-of-semester) × Rater Goals (harmony vs. control) × Ratee Performance (high, medium, or low) three-way interaction was significant, $F(2, 200) = 4.20$, $p < .05$, $\eta_p^2 = .04$. Raters inflated more for low performers than for high performers, and the rating inflation for low performers compared with high performers was stronger at midsemester than at end-of-semester. The results suggest that raters tended to use the equality norm for low performers, and this tendency was more salient at midsemester than at end-of-semester.

Under the motivating goal condition, raters increased ratings for both medium and low performers, but they did not distort ratings for high performers. Raters may have assumed that high performers were already highly motivated, thus did not need to receive negative feedback to be externally motivated. Further research may help to test these speculations.

Although Study 1 revealed the differential rating patterns for ratees with different performance levels, there are several limitations. First, the within-participant design required students to give ratings under all four rater goal conditions. They might believe that they should respond differently under the four conditions. The rater goals effects might therefore be exaggerated because of experimental demands. To alleviate this concern, we manipulated rater goals between participants in the second study. Second, performance evaluations conducted at different levels (self rating, peer rating, or supervisor rating) may differ (Murphy & Cleveland, 1995). To discern the generalizability of our findings, we replicated our analyses in a nonpeer rating context in Study 2.

Study 2

Method

Participants and procedure. Undergraduate students ($N = 120$) enrolled in an organizational behavior course participated in this study. Among all participants, 60 participants were female (50%), 52 participants were male (43.3%), and eight did not indicate their gender. The mean age of the participants was 18.8 years of age (minimum = 17 years, maximum = 20 years). All participants were randomly assigned to one of the four rater goal conditions, that is, the identification goal (i.e., control condition), fairness goal, harmony goal, or motivating goal.

Materials. A 15-min videotape was edited from *The Apprentice* (Season 2, Episode 11). Participants watched the video and evaluated the performance of the six team members in the video. The six team members in the video clip had been divided into two groups to finish a task of designing a new bottle for Pepsi. One group was named Apex (team members: Kevin, Kelly, and Ivanna), and the other team was named Mosaic (team members: Jennifer, Sandy, and Andy). The video showed the process of how the team members designed the new bottle and how they presented the new design to executive officers in the marketing department of Pepsi. A pilot study ($N = 13$) from undergraduate students showed that the Apex team received higher ratings (5.29) than the Mosaic team (3.88) on a 1–7 rating scale. Therefore, the 15-min video was appropriate for raters to differentiate among ratees with different performance levels. The performance level, in this case, was a within-participant variable. Previous studies have shown that people could make accurate ratings even on the basis of very short video clips. Ambady and Rosenthal (1993) found that students' judgments, on the basis of exposure to video clips lasting less than 30 s, could predict their later evaluations of instructors. We therefore believe our 15-min video provided enough information for raters to evaluate ratees' behaviors and performances.

Participants were asked to evaluate the performance of the six candidates on 10 items including innovation in designing new products, cooperating with team members, presentation skills, team work spirit, and so forth. The scale ranged from 1 (*poor*) to 5 (*excellent*). The reliability of the scale across the four goal conditions was .90.

Participants completed the evaluation in the classroom. First, they were briefly introduced to the backgrounds of the six team members. Then, they watched the 15-min video. They were instructed to observe all individuals closely and to evaluate their performance accordingly. After watching the video, participants were asked to play the role of team manager of the six members and to give ratings for each of them according to their assigned rating goals. The rating instructions for each goal followed those used in Study 1.

Results

Data analysis. Following Study 1, we analyzed the data by using HLM and ANOVA. Again, we used two-level models

(HLM2), wherein rating scores constituted the L1 cases nested within L2 raters. Although raters were also nested in rater goal conditions, which suggests a three-level model (HLM3), the number of rater goal conditions ($n = 4$) was too small to support an adequate examination of HLM3 (Raudenbush & Bryk, 2002). Thus, we explored the effects of rater goals by including rater goals dummy codes as L1 predictors of the dependent variables (Cohen et al., 2003; Raudenbush & Bryk, 2002). To examine the rating distortions for various levels of performers under each rater goal condition, we tested the interaction term of ratees' performance levels and the rater goal condition in the L1 model.

In the HLM analysis, the ratee's performance level was entered as the predictor in L1. The performance rating in the control condition was treated as a continuous variable representing ratees' performance levels. The random effects in HLM were only for the intercepts. As shown in Table 4, all random effects (τ_{00}) but one (in Model 1) were significant ($p < .01$), suggesting a nesting effect in our data, and the use of HLM analyses was thus appropriate. In the ANOVAs, as each rater gave ratings for all six ratees, we ranked the performance ratings by each rater and categorized the top two performers as the high performers, the middle two as the medium performers, and the bottom two as the low performers. Performance level was regarded as a within-participant factor. Table 4 presents the means, standard deviations, and zero-order correlations of all variables used in the ANOVAs. Table 5 summarizes the HLM results, and Table 6 demonstrates the ANOVA results. Figure 2 presents the slope plot of the Goal Condition × Ratee Performance Level interaction of the HLM analyses.

Tests of hypotheses.

Harmony goal. The HLM analysis showed that the interaction between rater goals and ratee performance was not significant ($\gamma_{50} = -.08, p = .38$). A 2 (goal condition: control vs. harmony) × 3 (performance level: high, medium, or low) ANOVA revealed a similar result in that the Goal Condition × Performance Level interaction effect was not significant, $F(2, 120) = 1.91$, $p = .15$, $\eta_p^2 = .03$. Hypothesis 1 was not supported in Study 2 (the nonpeer rating context).

Fairness goal. The HLM results indicated the interaction between the fairness goal and ratees' performance levels was significant ($\gamma_{60} = -.22, p < .05$). To further explore this interaction, we plotted rating scores at plus and minus one standard deviation above and below the sample mean of the ratees' performance level (Cohen et al., 2003). As shown in Figure 2 (top panel), the positive relationship between the ratees' performance level and the rating scores was smaller under the fairness goal condition than under the control condition. Simple slope tests indicated that the relationship between ratee performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the fairness condition ($\beta = .78$, $p < .01$).

A similar result emerged from the ANOVA results. The Goal Condition × Performance Level interaction effect was significant, $F(2, 108) = 3.20$, $p < .05$, $\eta_p^2 = .06$. Planned comparisons revealed that the effect of the goal condition was significant for high performers ($p < .05$) but not for medium and low performers. That is, there was significant rating deflation for high performers (control condition = 4.04, $SD = 0.32$; fairness condition = 3.88, $SD = 0.42$; Cohen's $d = 0.43$) but no significant deflation or inflation for medium and low performers. Hypothesis 2 was partially supported.

Motivating goal. The HLM analysis indicated the interaction between the motivating goal and ratees' performance levels was significant ($\gamma_{70} = -.26, p < .01$). To further explore this interaction, we plotted the rating scores at plus and minus one standard deviation above and below the sample mean of the ratees' performance levels (Cohen et al., 2003). As shown in the slope plot in Figure 2 (bottom panel), the positive relationship between the ratees' performance level and the rating scores is weaker under the motivating goal condition compared with the control condition. Simple slope tests indicated that the relationship between the ratees' performance and the rating scores was stronger under the control condition ($\beta = 1$, $p < .01$) than under the motivating condition ($\beta = .74$, $p < .01$).

The ANOVA revealed a similar result. The Goal Condition × Performance Level interaction effect on the ratings was significant,

Table 4
Means, Standard Deviations, and Zero-Order Correlations Among All Variables Measured in Study 2

Variable	*n*	*M*	*SD*	1	2	3
Control condition	34					
1. High performer		4.04	0.32	—		
2. Medium performer		3.39	0.33	.55**	—	
3. Low performer		2.59	0.38	.06	.57**	—
Harmony condition	29					
1. High performer		3.91	0.30	—		
2. Medium performer		3.25	0.19	.29	—	
3. Low performer		2.60	0.25	−.06	.38*	—
Fairness condition	25					
1. High performer		3.88	0.42	—		
2. Medium performer		3.37	0.44	.71**	—	
3. Low performer		2.71	0.47	.37	.80**	—
Motivating condition	32					
1. High performer		3.91	0.31	—		
2. Medium performer		3.36	0.36	.67**	—	
3. Low performer		2.81	0.38	.41*	.85**	—

Note. Performance ratings range from 1 to 5.
* $p < .05$. ** $p < .01$.

Table 5
Hierarchical Linear Modeling Results of the Interaction Effects of Raters' Goals and Ratees' Performance Level on Rating Scores in Study 2

Variable	Model 1		Model 2		Model 3		Model 4	
	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Fixed effect								
Intercept, γ_{00}	3.30**	0.03	3.30**	0.03	3.32**	0.04	3.32**	0.04
Ratees' performance level, γ_{10}			0.86**	0.04	0.86**	0.04	1.00**	0.07
Harmony goal, γ_{20}					−0.10	0.05	−0.10	0.05
Fairness goal, γ_{30}					−0.04	0.08	−0.04	0.08
Motivating goal, γ_{40}					0.02	0.07	0.02	0.07
Ratees' Performance Level × Harmony Goal, γ_{50}							−0.08	0.10
Ratees' Performance Level × Fairness Goal, γ_{60}							−0.22*	0.11
Ratees' Performance Level × Motivating Goal, γ_{70}							−0.26**	0.09
Variance component	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}	σ^2	τ_{00}
	.44	.00	.15	.05**	.15	.05**	.15	.05**
N (Level 1)	720		720		720		720	
N (Level 2)	120		120		120		120	

Note. $N = 120$.
* $p < .05$. ** $p < .01$.

$F(2, 126) = 9.52$, $p < .01$, $\eta_p^2 = .13$. Planned comparisons revealed that the effect of the goal condition was significant for low performers ($p < .05$) but not for medium and high performers. That is, there was significant rating inflation for low performers (control condition = 2.59, $SD = 0.38$; motivating condition = 2.81, $SD = 0.38$; Cohen's $d = 0.58$) but no significant deflation or inflation for medium and high performers. Hypothesis 3 was partially supported.

Variance explained. We examined how much of the variance in the performance ratings was due to true performance, distortion of raters' goals, and the Rater Goal × Ratee Performance interactions, as in Study 1. The true performance rating accounted for 59.18% of the variance (in Step 2). Raters' goals did not account for any additional variance (i.e., 0%) in Step 3. In Step 4, the interactions accounted for 0.77% of the variance. In addition, rater effect accounted for 10.20% of the variance, and 29.84% of the total variance is the random effect.

Discussion

In Study 2, under the harmony goal condition, contrary to our prediction, raters deflated their ratings for all performers. We suspect that the raters were using the equality norm to reduce conflicts rather than inflating ratings for both high and low performers. We calculated the standard deviation of the ratings among the performers to gauge the rating discriminability (Wong & Kwong, 2007). The ANOVA results showed that with discriminability as the dependent variable, the main effect of the goal was significant, $F(1, 61) = 4.08$, $p < .05$. The discriminability in the harmony condition (0.31) was significantly smaller than that in the

Table 6
Analysis of Variance Results of the Interaction Effects of Rater Goals and Performance Level on Rating Scores in Study 2

Effect	Mean rating		
	F	*MSE*	η_p^2
Hypotheses 1			
Goal condition (harmony vs. control)	2.16	0.15	.04
Performance level	486.15**	0.06	.89
Goal Condition × Performance Level	1.91	0.06	.03
Hypotheses 2			
Goal condition (fairness vs. control)	0.00	0.31	.00
Performance level	331.42**	0.07	.86
Goal Condition × Performance Level	3.20*	0.07	.06
Hypothesis 3			
Goal condition (motivating vs. control)	0.15	0.24	.00
Performance level	459.22**	0.06	.88
Goal Condition × Performance Level	9.52**	0.06	.13

Note. $N = 120$.
* $p < .05$. ** $p < .01$.

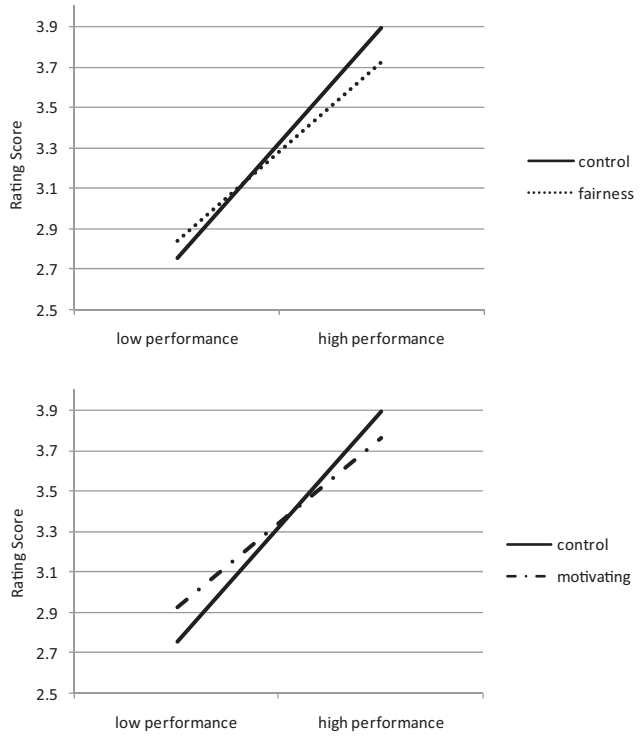

Figure 2. The performance rating as a function of goal condition and performance level (Study 2).

control condition (0.38). Therefore, raters tried to reduce conflict by lowering the discriminability rather than inflating the ratings.

Results for the fairness goal and the motivating goal for Study 2 resemble those found in Study 1. In both studies, raters who wanted to achieve a motivating goal inflated ratings for low performers but did not distort ratings for high performers. Under the fairness goal condition, similar to in Study 1, raters from Study 2 seemed to have double standards about fairness for various levels of performers. They deflated ratings for high performers according to the equality norm, but they did not distort ratings for medium and low performers according to the equity norm. Interestingly and complementarily, in Study 1, raters inflated ratings for medium and low performers, but they did not distort ratings for high performers. The results of the fairness goal effects in Study 1 and Study 2 suggest that raters may adopt different strategies to fulfill the fairness goal under different rating contexts. Equity (Adams, 1965) and equality (Deutsch, 1985) are the two most widely accepted justice rules. Previous research found that people were more likely to adopt the equity rule to represent fairness for out-group members and to apply the equality rule to maintain fairness for in-group members (Ng, 1984). Our studies suggest that people are more likely to adopt the equality rule for low performers in the peer rating context (Study 1); by contrast, people are more likely to adopt the equality rule for high performers in the nonpeer rating context (Study 2).

The major difference in the results between Study 1 and Study 2 is in the harmony condition. We found a strong interaction between the harmony goal and the performance level in Study 1 but not in Study 2. These inconsistent findings may largely be due to the different relationships between raters and ratees in Study 1 and Study 2. In Study 1, raters and ratees belonged to the same group, and they had a common group project to complete. Therefore, the harmony goal was essential. However, in Study 2, raters played the role of manager and gave ratings to the team members (as shown in the video). The goal of maintaining group harmony may not have been as salient as it was in Study 1.

General Discussion

Our studies contribute to research on the goal-based approach to performance appraisal (Cleveland & Murphy, 1992). Our results suggest that rating accuracy is shaped jointly by the ratee's ability (performance level) and the rater's motivation (goals). We found that raters give different ratings under various goal conditions. They inflated their ratings to achieve harmony, fairness, and motivating goal when giving peer ratings at midsemester. More important, rating inflation was stronger for poorer performers. In the nonpeer rating context, raters deflated ratings for high performers to achieve the fairness goal, and they inflated ratings for low performers to motivate them.

Our evidence supports the idea that rating errors are not only the result of inadequate rating skill but they are also the result of the intentions of the raters to pursue specific goals. The variance-explained statistics show that the goal effects account for certain amount of variance of rating scores (9.62% and 3.97% for Study 1, and 0.77% for Study 2). Raters are embedded in contexts in which they may have to achieve multiple goals. From the manager's perspective, performance evaluation is often the means to achieve other high level goals in human resource management, such as maintaining group harmony and promoting group productivity. For example, to build long-term relationships in an organization, reducing conflict and maintaining harmony are essential for groups to function. Therefore, raters have to consider the possible detrimental outcomes if accurate performance rating results are released. When rating accuracy is not the primary goal of the raters, they may intentionally distort their ratings to fulfill other goals. In this sense, rating errors could be attributed to the specific goals that raters are pursuing.

Our article also indicates the importance of examining contextual effects in the performance appraisal. Specifically, we examine the performance level as an important contextual factor in the rating process. To achieve a particular rating goal, raters may adapt their rating strategies to various performers. For example, in Study 1, raters inflated ratings for all performers with the greatest inflation for low performers. In Studies 1 and 2, to motivate ratees, raters inflated ratings for low performers but did not distort ratings for high performers. Note that under two different goal conditions, raters distorted their ratings differently for high and low performers. However, the two different rating distortions led to the same rating pattern; that is, the rating discriminability was reduced. Wong and Kwong (2007) examined the effect of raters' goals on rating patterns in terms of leniency and discriminability; however, they did not explore how the reduced discriminability came about. By examining raters' distortion of ratings for various levels of performers under different goal conditions, we indeed found that the rater's tendency to distort is contingent upon the ratee's characteristics. This provides a clearer picture of the effect of rater goals on rating scores.

Performance level as an important contextual factor is also practically important. One important implication of performance

evaluation is to differentiate the various levels of performers, which helps in administrative decisions, such as promotion and salary increases (Murphy & Cleveland, 1995). Raters will evaluate ratees' performance levels first and then revise their ratings according to the ratees' performance levels and their own specific rating goal. Our results suggest that raters will evaluate the various levels of performers differently if they are pursuing a special goal, such as harmony, fairness, or motivating ratees.

Our research also suggests that there is an alternative understanding of rating accuracy. Kruglanski (1989) suggested there are at least two ways to understand accuracy. First, the traditional and most accepted aspect of accuracy is accuracy as correspondence between judgment and criterion, which is similar to Murphy's (2005) definition, in which rating accuracy refers to the correspondence between performance appraisal ratings and actual performance levels. This conceptualization of accuracy assumes that there is a true score, following the *realist* paradigm. As Kruglanski (1989) argued, the "right" judgment (i.e., accurate ratings in the performance appraisal context) can also be investigated through the *phenomenal* perspective, which means that raters have their internal criteria about what right ratings should be. Raters may distort their ratings upward or downward for different reasons to give an appropriate rating that has a greater pragmatic utility. This conceptualization of rating accuracy suggests that we may put more emphasis on the appropriateness of performance ratings. In addition to rating accuracy, it is also important to consider whether rater goals are aligned with the organization's interests. As Murphy and Cleveland (1995) suggested, several often neglected criteria, such as practicality and the decision process, should also be considered in evaluating a performance appraisal system that focuses more on how much gain in utility the performance appraisal system brings to organizations.

There are three additional interesting observations from the present research. The first one pertains to comparing the main effects of the rater goals of Study 1 and those of Study 2. Study 1 showed the robust main effects of rater goals. The mean ratings in the harmony, fairness, and motivating conditions were all higher than the mean rating in the control condition, indicating that raters generally inflated their ratings regardless of which goals they wanted to achieve. However, we did not find the main effects of rater goals in Study 2, despite the significant Rater Goals $\times$ Ratee Performance interaction effects. This comparison suggests that there could be some factors that lead the main effects of rater goals to be more or less dependent on ratees' performance levels. We speculate that the differences between Studies 1 and 2 in terms of experimental design of goal manipulation (within- vs. between-participants), rating dimensions (a single overall rating vs. ratings on specific components), and ecological validity (real experience and consequences vs. hypothetical evaluation) may be responsible for the disparate rater goal effects across the two studies.

In terms of experimental design, the within-participant manipulation of rater goals in Study 1 (vs. the between-participants manipulation) allowed raters to be more aware of their ratings in one goal condition relative to their ratings in other goal conditions. Thus, stronger rating goal effects were found because Study 1 offered more opportunities for participants to differentiate ratings among all goal conditions. In terms of rating dimensions, the overall performance evaluation with a single-item measurement in Study 1 induced raters to rate on the basis of their general impressions of the ratees, whereas the specific-domain evaluation with multiple-items measurement in Study 2 induced raters to rate on the basis of elaborated memory of ratees' specific behavior. General impressions are often formed by category-based processing that requires less mental effort, whereas the judgment on specific domains often involves feature-based processing that involves more mental effort (Fiske & Neuberg, 1990; Lance, LaPointe, & Stewart, 1994). Thus, the effect of rater goals in Study 2 was weaker probably because raters in Study 2 put less mental effort in achieving the rater goals, because they allocated more mental resources for giving domain-specific ratings. In terms of ecological validity, raters in Study 1 believed that ratees would know the results and that their future behavior and performance could be influenced by the ratings; therefore, raters were more motivated to give ratings to achieve those goals. The stronger rater goal effects in Study 1, therefore, could be due to the participants having a stronger motivation to achieve rater goals in real rather than in hypothetical ratings. It is interesting and important for future research to examine the above three speculations.

The second interesting observation is the impressive converging results from HLM and ANOVA, even though the two methods of analysis in this study involve different assumptions about the rating standards the raters employed (i.e., absolute vs. relative standards). Because raters tend to give absolute ratings that can reflect relative performance (Wong & Kwong, 2005), it is difficult to tell for sure whether an absolute standard or a relative standard is more appropriate in the present study. Nonetheless, the converging results from these two methods of analysis indicate that the effects of rater goals are quite robust regardless of the assumptions about the rating standard.

Finally, we also notice that raters in the fairness condition gave equality-oriented ratings to some extent although (a) the instruction in the fairness goal condition (e.g., giving an evaluation that reflected fairness and accuracy) might have highlighted equity, and (b) raters evaluated out-group members (e.g., in Study 2). We propose two possible reasons for this observation. First, it is possible that raters do not consider ex-members in Study 1 or newly met ratees in Study 2 as completely out-group members. Logically, the in-group versus out-group distinction is not completely absolute. Ex-members might be considered as in-group members because the raters and ratees studied in the same university. Similarly, newly met ratees might be considered as in-group members by raters who define a group to include everyone working in the same organization. Therefore, there might be some raters who always used the equality rule to represent fairness in our studies. Second, the equity-oriented responses are conceptually identical to responses in the control condition. Therefore, whenever there is a presence of any effect due to whatever reasons, the only observed direction of this effect would be toward equality. Such an imbalance or asymmetry in the effect direction has been called effect propensity (Simonson, Kramer, & Young, 2004).

Limitations and Future Research

Our study has implications for both researchers studying organizational behavior and practitioners. In many management studies, the manager's performance ratings are used as the employee's performance level indicator. However, these ratings are often significantly shaped by the manager's goals. If rater goals are

explicitly controlled in these studies, the results could be more robust.

Future research may also explore individual differences in the implicit theories of rating strategies. For example, people may have different implicit theories of motivation in terms of whether inflated ratings or deflated ratings will motivate employees more effectively. Some people may adopt the idea that positive reinforcement is an efficient motivating strategy, and they will tend to motivate others by inflating ratings. Others may adopt the idea of control theory, and they will tend to motivate others through deflated ratings. Similarly, to achieve fairness, people may also have different philosophies on justice in terms of equity or equality. People who believe in the equality rule will tend to reduce the rating variance among ratees, and people who equate fairness with equity will tend to maintain the rating variance among ratees. Our study also suggests that the adoption of an implicit theory also depends on the social context of the performance evaluation, such as the rating context, the ratees as in-group or out-group members, and the performance levels of the ratees. All the above ideas suggest that a comprehensive understanding of the goal-based approach to performance appraisal should consider how raters with different implicit theories of achieving a particular goal distort their ratings.

The structure and antecedents of goals also deserve further investigation. In reality, raters may have multiple and hierarchical goals. For example, raters may want to motivate ratees through giving fair ratings or to use the harmony goal to enhance productivity (Meindl, 1989). In Study 2, raters seemed to use the equality norm of justice to achieve the harmony goal. Examining the structure of various goals will help us to have a better understanding of the rater's motivation and rating strategy.

For organizations, the interpretation of rating scores should be conducted in the context of the goals that managers pursue. To reduce rating discrepancies among raters, clearly stating the rating goal for raters may be a solution. Managers who are more conscious of their rating goals can use performance evaluations to manage their employees' performance more effectively without being too concerned about the rating accuracy.

We acknowledge that there are limitations in the present research in terms of generalizing our findings to firm settings. In our two studies, the raters and ratees did not build long-term relationships. In Study 1, students formed the group in one semester. In Study 2, participants finished the evaluation within 45 min and never met the ratees. However, in actual companies, peers or supervisors and subordinates often have long-term relationships. The harmony goal may be even more important in firm settings.

Our undergraduate students' peer ratings and the laboratory setting hindered our exploration of other meaningful raters' goals. For example, performance ratings in firms are inevitably influenced by politics, which were not measured in our settings. The retention goal is another important goal pursued by raters in the organizational context.[1] In actual companies, the potential loss of top talent may be a primary concern for managers, and thus retention may be an important goal as raters approach the performance appraisal process. That is, one of the key goals of the performance appraisal process is to help retain top performers and to often encourage low performers either to improve or to leave. To achieve the retention goal, raters may inflate ratings for high performers and deflate ratings for low performers. Future field studies should take into account this type of goal.

Finally, the performance measure in Study 1 is a single item measure, which hindered our exploration of performance dimensionality. In reality, performance measures are often multiple-dimensional. We suspect that raters' goals are likely to have differential effects on different performance dimensions. This is an area that warrants future research.

Conclusion

The goal-based approach to performance appraisal (Cleveland & Murphy, 1992; Murphy & Cleveland, 1991, 1995) emphasizes the impact of raters' goals on rating scores. However, only a few studies empirically tested the influence of raters' goals on ratings (Murphy et al., 2004; Wong & Kwong, 2007). We extend this line of research and demonstrate that the effects of raters' goals are reflected differently for ratees with various performance levels.

[1] We thank a reviewer for pointing this out to us.

References

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267–299). New York, NY: Academic Press.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64,* 431–441.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: Freeman and Company.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology, 88,* 87–99.

Brockner, J., Wiesenfeld, B. M., & Martin, C. L. (1995). Decision frame, procedural justice, and survivors' reactions to job layoffs. *Organizational Behavior and Human Decision Processes, 63,* 59–68.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5,* 159–172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2,* 49–68.

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology, 71,* 672–678.

Cardy, R. L., & Kehoe, J. F. (1984). Rater selective attention ability and appraisal effectiveness: The effect of a cognitive style on the accuracy of differentiation among ratees. *Journal of Applied Psychology, 69,* 589–594.

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior.* New York, NY: Cambridge University Press.

Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. *Research in Personnel and Human Resources Management, 10,* 121–185.

Cohen, J., Cohen, P., West, L. S., & Aiken, S. G. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

DeNisi, A. S. (1996). *A cognitive approach to performance appraisal: A program of research.* New York, NY: Routledge.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice. *Journal of Social Issues, 31,* 137–149.

Deutsch, M. (1985). *Distributive justice: A social-psychological perspective.* New Haven, CT: Yale University Press.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23,* 1–74.

Gaugler, B. B., & Rudolph, A. S. (1992). The influence of assessee performance variation on assessor's judgments. *Personnel Psychology, 45,* 77–98.

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

Highhouse, S. (1996). Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes, 65,* 68–76.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24,* 623–641.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions of performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Review, 106,* 395–409.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79,* 332–340.

Leventhal, G. S., Michaels, J. W., & Sanford, C. (1972). Inequity and interpersonal conflict: Reward allocation and secrecy about reward as methods of preventing conflict. *Journal of Personality and Social Psychology, 23,* 88–102.

Marin, G. (1981). Perceiving justice across cultures: Equity vs. equality in Colombia and in the United States. *International Journal of Psychology, 16,* 153–159.

Maurer, T. J., & Alexander, R. A. (1991). Contrast effects in behavioral measurement: An investigation of alternative process explanations. *Journal of Applied Psychology, 76,* 3–10.

Meindl, J. R. (1989). Managing to be fair: An exploration of values, motives, and leadership. *Administrative Science Quarterly, 34,* 252–276.

Molm, L. D., Takahashi, N., & Peterson, G. (2003). In the eye of the beholder: Procedural justice in social exchange. *American Sociological Review, 68,* 128–152.

Murphy, K. R. (2005). Rating accuracy. *Blackwell Encyclopedic Dictionary of Human Resource Management, 9,* 1–310.

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. (1991). *Performance appraisal: An organization perspective.* Boston, MA: Allyn & Bacon.

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

Murphy, K., Cleveland, J., Skattebo, A., & Kinney, T. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89,* 158–164.

Ng, S. (1984). Intergroup allocation bias before and after group discussion. *The Journal of Social Psychology, 124,* 95–103.

Powers, W. T. (1973). *Behavior: The control of perception.* Chicago, IL: Aldine.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Seligman, M. E. P. (1975). *Helplessness: On depression development and death.* San Francisco, CA: Freeman.

Simonson, I., Kramer, T., & Young, M. J. (2004). Effect propensity. *Organizational Behavior and Human Decision Processes, 95,* 156–174.

Slaughter, J. E., Sinar, E. F., & Highhouse, S. (1999). Decoy effects and attribute-level inferences. *Journal of Applied Psychology, 84,* 823–828.

Vancouver, J. B., Thompson, C. M., Tischner, E. C., & Putka, D. J. (2002). Two studies examining the negative effect of self-efficacy on performance. *Journal of Applied Psychology, 87,* 506–516.

Vroom, V. H. (1964). *Work and motivation.* New York, NY: Wiley.

Wexley, K. N. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology, 56,* 45–48.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 233–236.

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology, 56,* 45–48.

Wong, K. F. E., & Kwong, J. Y. Y. (2005). Between-individual comparisons in performance evaluation: A perspective from prospect theory. *Journal of Applied Psychology, 90,* 284–294.

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goal on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92,* 577–585.

Received December 22, 2008
Revision received October 29, 2009
Accepted November 5, 2009 ■