Journal of Management
http://jom.sagepub.com/

Rater Reactions to Forced Distribution Rating Systems

Deidra J. Schleicher, Rebecca A. Bull and Stephen G. Green

Journal of Management 2009 35: 899 originally published online 5 February 2008
DOI: 10.1177/0149206307312514

The online version of this article can be found at:
http://jom.sagepub.com/content/35/4/899

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
S•M•A
SOUTHERN MANAGEMENT ASSOCIATION
Southern Management Association

Rater Reactions to Forced Distribution Rating Systems†

Deidra J. Schleicher*
Rebecca A. Bull
Stephen G. Green

Krannert School of Management, Purdue University, 100 S. Grant St., West Lafayette, IN 47907-2076

Two experiments examined raters' reactions to a forced distribution rating system (FDRS), which, despite its popularity in organizations, has been largely ignored in the empirical research literature. Greater difficulty and less fairness were reported by raters when the FDRS was used for administrative purposes and when there was reduced variability in ratee performance. In addition, the FDRS was found to be more difficult and less fair than a more traditional rating scale format. Finally, difficulty and fairness reactions had significant implications for raters' confidence in their ability to provide feedback to the ratees and their self-efficacy for using the system going forward.

Keywords: *performance appraisal; forced distribution ratings; rater reactions; fairness*

Performance appraisals (PAs) are ubiquitous, occurring in some form across all types of organizations and jobs (Bannister & Balkin, 1990; Murphy & Cleveland, 1991, 1995). They are equally ubiquitous in research, constituting a large segment of the scholarly organizational literature. Unfortunately, however, there is frequently a disconnect between what is examined by PA researchers and what is of most interest and concern to practitioners and managers, leading to the oft-noted gap between PA research and practice (Banks & Murphy, 1985; Bretz, Milkovich, & Read, 1992; Ilgen, Barnes-Farrell, & McKellin, 1993; Levy & Williams, 2004; Maroney & Buckely, 1992; Smither, 1998). The current article attempts to

†We sincerely thank Ranga Ramanujam for his willingness and assistance in obtaining Study 1's sample.

**Corresponding author: Tel.: 765-496-2048; fax: 765-496-7434*

E-mail address: deidra@purdue.edu

Journal of Management, Vol. 35 No. 4, August 2009 899-927
DOI:10.1177/0149206307312514

address two areas in this gap by examining (a) rater reactions to (b) a forced distribution rating system (FDRS).

FDRS is a particular type of PA approach in which the rater is required to fit evaluations to a particular distribution (often a "normal" distribution or a variant thereof, such as a "20–70–10" distribution). As such, it forces raters to avoid lenient performance ratings. Given reports that 77% of companies believe that lenient appraisals jeopardize the validity of their PA systems (Bretz et al., 1992), it is likely this aspect of FDRS that has contributed to its popularity in organizations. Its use at General Electric (GE) provides a particularly notable example. Jack Welch, GE's former CEO, argued that this type of system forced those responsible for appraising performance to be honest with workers. He truly believed that this approach to performance evaluation was key to the organization's competitive advantage, primarily because it periodically cleared out the "dead wood" (i.e., those employees assigned to the lowest performance category were terminated) and served to motivate those remaining (Welch & Welch, 2005). In the post-Welch era, widespread popularity of this type of evaluation format has continued. In fact, recent estimates are that approximately 20% of *Fortune* 1000 companies use some form of an FDRS (Sears & McDermott, 2003). However, research on FDRS has severely lagged practice.

The use of FDRS has both its proponents and its critics. On one hand, proponents argue that an FDRS forces managers to be more honest and direct in doing PAs and thus is a better approach for accurately identifying both the high-potential employees and the bottom performers (Boyle, 2001; Guralnik, Rozmarin, & So, 2004; HayGroup, 2002; Meisler, 2003). Consequently, both promotions or raises and cuts can be more strategic, which should improve short- and long-term performance (Sears & McDermott, 2003). In addition, an FDRS is believed to help create and sustain a high-performance and high-talent culture, in which poor performance is not tolerated and honest feedback is expected (Guralnik et al., 2004; HayGroup, 2002; Meisler, 2003). On the other hand, critics argue that an FDRS can hinder teamwork and collaboration, foster competition, bring legal challenges (Guralnik et al., 2004), and facilitate political game playing and widespread insecurity (as in its use at Enron; Fusaro & Miller, 2002). Moreover, in the only published empirical research on an FDRS, Scullen, Bergey, and Aiman-Smith (2005) demonstrated with a simulation that initial improvements to individual and organizational performance associated with an FDRS decline sharply over time with successive iterations. Despite the prevalence of these opposing arguments, there currently is "no generally accepted research that gives either side clear superiority in the debate" (Bates, 2003, p. 64).

Particularly noteworthy is the lack of research on *rater reactions* to an FDRS. Although this simply mirrors the broader PA literature, which has largely ignored rater reactions in general (see below), this lack of attention to the reactions of raters, both generally and specifically with regard to FDRS, is at odds with the espoused importance of rater agreement and acceptance for realizing the potential gains of the performance evaluation and feedback process (Fletcher, 2001; Longenecker, Sims, & Gioia, 1987; Meyer, Kay, & French, 1965; Murphy & Cleveland, 1991, 1995; Taylor, Masterson, Renard, & Tracy, 1998; Tziner & Kopelman, 2002; Tziner, Kopelman, & Joanis, 1997). In short, rater support is believed to be a prerequisite for the effectiveness of PA systems.

The relevance of rater reactions as a criterion in PA, combined with the popularity of the FDRS in practice, suggests the value of systematic investigations into raters' reactions to this type of rating approach. Accordingly, the current research, across two studies, examines the reactions of raters completing an FDRS and how such reactions can (a) be affected by rating system characteristics and, in turn, (b) affect rater confidence in using the system in the future.

The Nature of Rater Reactions

In defining and measuring raters' reactions, our goal was to use variables reflective of research in this area and reports of raters using an FDRS in the field. Unfortunately, however, the vast majority of research on PA reactions concerns *ratees*' reactions (e.g., Brett & Atwater, 2001; Brown & Benson, 2003; Fedor, Bettenhausen, & Davis, 1999; Findley, Giles, & Mossholder, 2000; Keeping & Levy, 2000; Korsgaard & Roberson, 1995; Lee & Son, 1998; Levy & Williams, 2004). There has been substantially less work done on raters' reactions (see Levy & Williams, 2004; Tziner et al., 1997; Tziner & Kopelman, 2002), and the work on ratees cannot simply be applied to raters. Although a few theoreticians have discussed qualitative or anecdotal reports of how raters react to PA systems in general (e.g., Longenecker et al., 1987; Meyer et al., 1965; Nalbandian, 1981; Townley, 1999; Tziner & Kopelman, 2002), there is no established taxonomy of rater reactions. Consequently, as Taylor et al. (1998) have noted, existing measures of PA rater reactions are generally not available. These authors created their own two-item, general measure of rater reactions to the PA system (i.e., "I am satisfied with the way performance evaluations are done at . . ." and "I wish . . . would change the performance evaluation system it now uses"; Taylor et al., 1998, p. 578). However, we deemed this measure too general for our purposes; we were more interested in identifying and measuring explicit factors that might underlie such general reactions. Thus, we turned to published anecdotal reports of rater reactions and conducted field interviews with managers ($N = 6$) currently using an FDRS. Our goal was to identify reactions that were valid representations of raters' experiences with an FDRS and had potentially significant implications for the effectiveness of any PA system. As discussed below, these sources and goals converged in suggesting two important dimensions of rater reactions to an FDRS: fairness and difficulty.

First, perceptions of fairness, believed to be omnipresent in organizations in general, are central to reactions to human resource activities, including PA (Erdogan, Kraimer, & Liden, 2001; Folger, Konovsky, & Cropanzano, 1992; Greenberg, 1986; Levy & Williams, 2004; Taylor et al., 1998; Taylor, Tracy, Renard, Harrison, & Carroll, 1995; Thomas & Bretz, 1994). Blume, Baldwin, and Rubin (2005, 2006) have noted that fairness is a salient issue in reactions to an FDRS, but this was in terms of *ratee* reactions. From anecdotal reports, however, fairness concerns also appear to play a role in *raters*' reactions to an FDRS. For example, Meisler (2003) quoted several whose concerns with the FDRS revealed fairness issues: "But perhaps—and maybe a lot more than perhaps—the people at the low end of the bell curve don't *deserve* [italics added] to be . . . fired" and "A lot depends on how *fairly* the system is developed. And how *fair* the people are who carry it out" [italics added].

Hence, both procedural and distributive justice issues (Greenberg, 1990) seem to underlie FDRS reactions. In addition, research on organizational justice suggests that raters' perceptions of the fairness of an FDRS (indeed, any PA approach) could affect commitment to, confidence in, and willingness to support the PA process (Cobb & Frey, 1996; Colquitt, Conlon, Wesson, Porter, & Ng, 2001; Condrey, 1995; Deckop, Mangel, & Cirka, 1999; Erdogan, 2002), thus establishing its relevance in this context.

Second, perceived difficulty of a decision task is a construct frequently used in the decision-making and other literatures. It refers to the extent to which someone (a rater, in this case) hesitates, deliberates, or vacillates in performing a decision-related task (Chatterjee & Heath, 1996; Cheung, Chan, & Wong, 1999; Trafimow, Sheeran, Conner, & Finlay, 2002). The reports previously mentioned indicate that this is another salient dimension for FDRS raters (see Sears & McDermott, 2003). For example, Bates (2003) noted raters' reports that "Assigning people in a C category was a *difficult* [italics added] decision." In addition, our interviews with managers confirmed that the difficulty of making these ratings is a salient issue and of concern to them. Managers reported spending considerable time on this task and often hesitating and vacillating when trying to discriminate among ratees and make the final assignments to rating categories. Moreover, research in other areas suggests that perceived difficulty could have a number of important implications for the effectiveness of any PA system. That is, perceived difficulty has been found to predict behavioral intentions and behaviors (Cheung et al., 1999; Trafimow et al., 2002), to negatively affect future efficacy expectancies (Terry & O'Leary, 1995), and to prevent individuals from making a decision (Shiloh, Koren, & Zakay, 2001) or lead them to a less than optimal decision (Gati, Krausz, & Osipow, 1996; Zakay & Wooler, 1984). In addition, perceived difficulty of a task can adversely affect confidence in, and motivation to defend, judgments (e.g., Bandura, 1977, 1982; Napier & Latham, 1986; Tziner, Murphy, & Cleveland, 2005) and raters' acceptance of evaluation systems specifically (Fedor et al., 1999). It should be noted, however, that perceived difficulty is not necessarily a uniformly negative reaction, in that greater perceived difficulty could also lead one to be more careful in making a decision (Chatterjee & Heath, 1996, p. 154). Nonetheless, as the above review of research suggests, perceived difficulty can change one's approach to decisions and future behavior, thus making it a relevant and important reaction variable in this context.

In summary, the FDRS appears to provoke both fairness and difficulty reactions among raters using this PA approach, and there is evidence from other areas that such reactions could have relevance for PA system effectiveness. In trying to predict and understand these reactions, we examine two important rating system characteristics as antecedents.

PA System Characteristics

As Murphy and Cleveland (1995) have noted, characteristics of PAs can be viewed as part of the context in which appraisal occurs (see also Levy & Williams, 2004). Both proximal and distal contextual factors can affect PA raters: Proximal factors are those that directly impinge on the individual rater (e.g., aspects of the PA scale, instructions given to raters), whereas distal factors affect the rater more indirectly (e.g., organizational culture that determines norms for evaluating performance; Cleveland, Morrison, & Bjerke, 1986). In Study 1,

we manipulate two more proximal factors (see Findley et al., 2000; Levy & Williams, 2004) that repeatedly emerged in interviews with managerial raters and in the broader PA literature: purpose of ratings and variability of performance among ratees.

Purpose of ratings. As with any PA approach, organizations vary with regard to the purposes for conducting an FDRS (Blume et al., 2005; Meisler, 2003); our interviews with managers using an FDRS confirmed this. At one extreme, some organizations use these ratings to terminate employees placed in the bottom performance category. At GE, for example, these ratings led to the termination of employees in the bottom 10% of the performance distribution (Meisler, 2003). At the other extreme, some organizations collect these ratings for record-keeping purposes only. They are placed in an employee's personnel file (and perhaps conveyed to the employee in a feedback meeting), but no administrative action is taken based on the results. In between these two extremes are organizations that use the ratings to determine promotions or demotions, different assignments, and levels of compensation. These various uses of FDRS ratings correspond to those discussed in the broader PA literature, which has identified purpose of rating as an important contextual factor likely to affect PA outcomes (see Cleveland, Murphy, & Williams, 1989; Levy & Williams, 2004; Murphy & Cleveland, 1991, 1995). Because the most definitive empirical findings in this area pertain to the differences between PA done for administrative purposes (e.g., promotions, raises, terminations) versus nonadministrative purposes (e.g., research, development) (Jawahar & Williams, 1997), these are the categories of purpose we examine in our studies. (The nonadministrative condition is operationalized as research in Study 1 and developmental in Study 2.)

Anecdotal evidence, and the literature in the broader PA area (Jawahar & Williams, 1997; Longenecker et al., 1987; Meyer et al., 1965), suggests that raters will find an FDRS more difficult and see it as less fair if they believe there are associated administrative consequences. For example, one rater cited in Longenecker et al. (1987) noted, "I know that it sounds funny, but the fact that the process is ultimately tied to [administrative decisions] influences the ratings. . . . Whenever a decision involves [that], things can get very . . . ticklish" (p. 185). Because an FDRS constrains raters' responses a priori (i.e., according to the fixed distribution), and because we know raters prefer having flexibility and control when doing PA for administrative purposes (Longenecker et al., 1987), we expect that raters will view the FDRS as less fair when there are administrative consequences. In addition, the decision-making literature has repeatedly found a link between decision consequences and perceived difficulty (London, Casey, Chatterjee, & Hurley, 1997; Zhang & Mittal, 2005), such that decision makers are more likely to hesitate, deliberate, and vacillate when there are weightier consequences to their decisions. Finally, we know that raters tend to inflate their ratings (i.e., be more lenient) when the purpose of the PA is administrative decisions compared to either research or developmental purposes (see Jawahar & Williams, 1997; Levy & Williams, 2004; Murphy & Cleveland, 1991). Because the FDRS format does not allow for such leniency, however, it is likely that raters will experience the FDRS task as more difficult and less fair when there are administrative consequences attached.

Hypothesis 1: Raters completing the FDRS under conditions of administrative purposes, compared to those completing it for nonadministrative purposes, will see the FDRS as more difficult.

Hypothesis 2: Raters completing the FDRS under conditions of administrative purposes, compared to those completing it for nonadministrative purposes, will see the FDRS as less fair.

Performance variability of ratees. The distribution of performance levels across ratees has also been identified as a factor likely to affect rater reactions (see Levy & Williams, 2004). In several of the reports cited above, and our interviews with managers using FDRS in organizations, a frequent theme is that performance across a set of ratees is not variable enough to warrant assignment to the various FDRS categories (e.g., "Sometimes [raters] are forced to identify poor performers even though they don't have the data which indicate they exist"; quoted in Bates, 2003). In other words, raters often report difficulty and perceived unfairness with an FDRS because it forces a particular amount of variability that may not reflect their set of ratees. Specifically, such raters often insist that all of the ratees are average or even above average (Bates, 2003). This may simply reflect a "Lake Woebegone" leniency bias (Jawahar & Williams, 1997) on their part, or it may in fact be true, owing to, for example, range restriction resulting from a valid selection system or the likelihood that supervisors have terminated or transferred any low performing ratees or are effective at motivating them to improve. Scullen et al. (2005) noted that such reduced variability is more likely over time with continued use of an FDRS. Thus, regardless of whether it is perceptual or veridical, it is an important question how such variability in ratee performance affects rater reactions.[1]

Raters forced to rank a group of ratees that they consider all above average are likely to (a) feel that having to place some of them into a lower performance category is unfair (i.e., it is an issue of justice if category assignments do not accurately reflect ratee performance levels; Greenberg, 1986) and (b) experience greater difficulty (i.e., hesitation, deliberation, vacillation) doing so. "Forcing managers to label some as low performers could be arbitrary if everyone in the peer group is doing a good job" (Blume et al., 2005, p. 27). Finer distinctions among alternatives, such as is involved with reduced performance variability, have been shown in the decision-making literature to lead to greater perceived difficulty (Chatterjee & Heath, 1996; also see Shiloh et al., 2001). For these reasons, we predict the following:

Hypothesis 3: Raters assigned to rate a pool of ratees less variable in performance, compared to those assigned ratees more variable in performance, will see the FDRS as more difficult.

Hypothesis 4: Raters assigned to rate a pool of ratees less variable in performance, compared to those assigned ratees more variable in performance, will see the FDRS as less fair.

Implications of Fairness and Difficulty Reactions

An additional goal of this research is to empirically substantiate the importance of rater reactions (specifically, fairness and difficulty) as criteria in PA research. As Murphy and Cleveland (1995) noted, "Reaction criteria are almost always relevant, and an unfavorable reaction may doom the most carefully constructed appraisal system" (p. 314). Yet to our knowledge this has never been empirically tested, certainly not with FDRS specifically. We believe that raters' more immediate reactions to the FDRS task (i.e., fairness and difficulty) are important because they can have implications for use of the PA system going forward, thus suggesting that such reactions could possibly undermine the effectiveness of the FDRS process.

Perceptions of fairness, for example, have been shown to relate to important variables such as satisfaction with, commitment to, and confidence in processes (see Cobb & Frey, 1996; Colquitt et al., 2001; Condrey, 1995; Deckop et al., 1999). Perceived lack of fairness of the FDRS thus could reduce rater confidence in the PA process (see Cobb & Frey, 1996; Colquitt et al., 2001; Condrey, 1995; Deckop, et al., 1999; Erdogan, 2002). Similarly,

research in other areas has shown that when decision or judgment tasks are perceived as being more difficult, one's confidence in one's ability to make a good decision and future efficacy expectations may be weakened (Bandura, 1977, 1982; Napier & Latham, 1986; Terry & O'Leary, 1995; Tziner et al., 2005). Finally, to the extent that an event provokes negative reactions, one is motivated to avoid such circumstances (Bradley, Codispoti, Cuthbert, & Lang, 2001; Dickson & MacLeod, 2004; Elliot & Thrash, 2002; Hillman, Rosengren, & Smith, 2004); this has also been shown in the area of PAs specifically (Harris, 1994). Thus, for all of the above reasons, we predict the following:

Hypotheses 5a and 5b: Fairness reactions to the FDRS will be positively related to confidence in one's ability to provide acceptable feedback to the ratees (Hypothesis 5a) and self-efficacy in using the FDRS in the future (Hypothesis 5b).

Hypothesis 6a and 6b: Difficulty reactions to the FDRS will be negatively related to confidence in one's ability to provide acceptable feedback to the ratees (Hypothesis 6a) and self-efficacy in using the FDRS in the future (Hypothesis 6b).

The Research Context

Study 1 was designed with the primary goal of experimentally studying FDRS processes with a group of raters who were (a) similar to managers on demographics and work experience and (b) facing the same sort of dilemmas inherent in our variables discussed above (e.g., consequences of ratings, variability in performance). Discussed in greater detail below, these requirements led us to study FDRS processes first in the context of MBA students evaluating the classroom performance of their peers in a required MBA course. Although the use of peers admittedly places some limits on the external validity of our study (an issue discussed in more detail later), it should be noted that (a) the primary goal in Study 1 is to strengthen internal validity and that (b) this issue is rectified in Study 2, which does not rely on peer ratings. In addition, the choice of context for Study 1 allowed us to satisfy other important conditions: Such rankings made sense in this context, the administrative consequences mattered to the participants, real working relationships were at stake, and we could use an experimental design (yet, of importance, *without* the raters knowing they were participating in an experiment).

Study 1

Method

Participants

Full-time MBA students (N = 175) from four sections of a required course at a large Midwestern public university served as the participants in this study. Average age of the participants was 28 years, 78% were male, 29% were members of minority groups, and the majority (i.e., 70%) had previous managerial work experience (overall M = 2.5 years of experience) and PA experience (79% reported they had previously appraised another's on-the-job performance, and 30% indicated they had used an FDRS-type system for doing so). Traditionally,

these students are highly motivated to perform well in such required courses and are quite sensitive to factors that might affect their final course grades and grade point averages.

Design and Procedure

To study the effects of the PA system variables, students were asked to participate in a forced distribution rating of classroom participation (i.e., contribution to class discussions). In this course, participation was traditionally considered in determining each student's final course grade and had a significant influence on it (20% of final grade). Thus, this performance dimension was highly salient to the students. The study used a 2 (purpose: administrative consequence vs. no administrative consequence) × 2 (higher vs. lower performance variability) between-subjects design. Two sections of the course were randomly assigned to receive the Administration Consequence condition, and the other two received the No Consequence condition.[2] The other variable, performance variability, was randomly assigned within sections, so that approximately equal numbers of students in each section received each of the two conditions (lower variability vs. higher variability).

Lists of ratees were created in the following way. For the lower variability conditions, course instructors were asked in the 6th week of the 8-week course to identify their top 10 students in each section in terms of classroom participation performance (this represents about 20% of each section's class size). These lists became the lower variability condition for each section. For the higher variability conditions, course instructors *randomly* selected 10 names from a list of *all* students in each section. A manipulation check completed by participants ("Regardless of the rankings you were forced to give, how similar do you believe the students on your list are with regard to class participation"; scaled from 1, *very similar levels of participation*, to 5, *not similar at all*) confirmed that raters in the lower variability condition did in fact perceive their ratees as more similar than those in the higher variability condition (2.78 vs. 3.05, respectively), $F(1, 163) = 4.27$, $p < .05$. We also conducted a manipulation check on the actual final participation scores assigned by the instructors at the end of the semester for the ratees on these lists. The variability in participation scores for those ratees on the lower variability lists was significantly less than that for those ratees on the higher variability lists ($SD = 0.27$ vs. 0.93, respectively), $F = 11.86$, $p < .001$.

Approximately 6 weeks into the 8-week course, the experiment was conducted in class. At the beginning of each section's class, which were all held on the same day, one of the researchers introduced the rating task. Students in all sections were told that the master's program was seeking ways to better grade class participation of the MBA students and that their class was being used to try out a new system of student evaluations of participation. They were told that class members had been randomly divided into sets of students and that each student would get one of these sets to evaluate.[3] They were instructed that, based on their evaluation of each student's class participation, they should assign each student to one of three performance categories: A (for the top 20%), B (for the middle 70%), and C (for the bottom 10%). This distribution was chosen to mimic those commonly found in FDRS in corporate settings (see Blume et al., 2005, 2006; Sears & McDermott, 2003). Students were told they would have to assign all ratees on their list to one of the categories, in the proportions noted. It was emphasized that they should do their own independent evaluations and not talk

about or look at other people's rankings. Finally, they were also told that the administration was interested in student reactions to this new evaluation process, given that it was being considered for adoption across the program. Therefore, they were asked to complete a brief set of questions at the end of the ranking task for feedback to the MBA program office. These reactions measures (described below) were intentionally kept brief to avoid raising participants' suspicions.

As the packets were handed out, students in the administrative consequence sections were told, "Please, take this evaluation seriously. Summaries of your ranking evaluations will be given to the course instructor and he will use these ratings along with his own judgment to determine each student's class participation grade." Students in the no administrative consequence sections were told, "Please, take this evaluation seriously. This is just a pilot test of this rating process, however, and the instructor will not see these evaluations before final grades are assigned in this course. They will not affect any student's participation grade."

Immediately after the students completed the FDRS task and reaction measures, another of the researchers entered the room to do the debriefing. Students were told that the administration was not, in fact, considering this type of change to the MBA evaluation process and, instead, that they had just participated in an experiment, and those in the administrative consequence condition were assured that the rankings would not, in fact, be used to determine course grades. The goals of this study were explained to the students, and they were told why the deception had been necessary; no student suspicions were revealed during this debriefing. It was further explained to them that it was essential they not discuss this with other students in other sections of the course. Finally, the voluntary nature of this study was introduced, and students were told that they were not required to hand in their ranking and reactions packets and that they could choose to opt out of the study at that point; one student declined to include his packet. All students were thanked for participating and reminded of the importance of maintaining the deception through the end of the day.

Measures

Control variables. Following recommendations of Podsakoff, MacKenzie, Lee, and Podsakoff (2003), relationships between self-report measures (i.e., Hypotheses 5a, 5b, 6a, and 6b) were controlled for positive affectivity (PA) and negative affectivity (NA), allowing us to factor out common method variance ascribed to respondents' dispositions to respond positively or negatively. These were measured using 10 items from Diener, Smith, and Fujita (1995) and Shaver, Schwartz, Kirson, and O'Connor (1987) and had been completed by students earlier in the course. Students indicated the frequency (from 1, *never*, to 7, *always*) with which they experienced each of 10 emotions in general. These items were collapsed into PA (four items, $\alpha = .73$) and NA (six items, $\alpha = .83$).

Reactions. Following Chatterjee and Heath (1996) and Trafimow et al. (2002), we measured *difficulty* by asking participants to rate the difficulty (from 1, *extremely difficult*, to 5, *extremely easy*) of the ranking task across four items (to allow calculation of internal consistency reliability): (a) overall, (b) deciding who fit into the A category, (c) deciding who fit into the B category, and (d) deciding who fit into the C category. These four items were reverse coded such that larger numbers represented greater perceived difficulty and were averaged to create the overall

difficulty score (α = .64). *Fairness* was measured with four items, scaled from 1 (*strongly disagree*) to 5 (*strongly agree*). Three items asked respondents about their perceptions of the fairness of the ranking process (i.e., procedural fairness; "This ranking process is a fair way to evaluate class participation," "This ranking process is more fair than other approaches _____ might use to evaluate class participation," and "Most students would see this ranking process as a fair evaluation of class participation"), and one item assessed outcome (i.e., distributive) fairness ("Each student I ranked received a fair and accurate evaluation"). The mean of the four items was used as the fairness score (α = .76).

Confidence in providing acceptable feedback. Participants were asked, "If you had to explain and justify your rankings to the students you just ranked, how would you feel about doing that?" They then responded, on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*), to three statements: "I am confident that I could provide a good explanation for my ranking decisions," "I am confident my classmates would accept my rankings as accurate," and "I am confident my classmates would see my rankings as fair." Reliability was .80 (α).

Self-efficacy for future use of the FDRS. Participants were asked "If _____ adopted this ranking process to evaluate class participation performance going forward, how confident are you that you could continue to . . ." (a) "provide rankings that were an accurate appraisal of each student's participation performance," (b) "provide rankings that you thought were fair to each student," and (c) "provide rankings that all students would accept as fair and reasonable." They responded to each of the three statements on a scale from 1 (*not at all confident*) to 10 (*very confident*). Reliability for this scale was .94 (α).

Results

Table 1 reports descriptive statistics and intercorrelations for Study 1 variables.

Effect of Purpose and Variability on Reactions

Hypotheses 1 through 4, regarding the effect of both purpose and ratee performance variability on reactions, were tested via two two-way ANOVAs, one on difficulty and one on fairness. These ANOVA results are reported below, providing information on the F tests, R^2 values, cell means, and standard deviations for those means. The ANOVA on difficulty revealed no interaction, but significant main effects for both the purpose and ratee variability factors. Those raters in the administrative consequence condition, compared to the no administrative consequence condition, found the FDRS task to be significantly more difficult ($M = 3.47$, $SD = 0.77$, vs. $M = 3.19$, $SD = 0.81$, respectively), $F(1, 174) = 5.38$, $p < .05$ ($R^2 = .03$), thus supporting Hypothesis 1. In addition, those raters assigned a group of ratees with less variability in performance found the FDRS task to be significantly more difficult than those assigned ratees with more variability in performance ($M = 3.46$, $SD = 0.76$, vs. $M = 3.22$, $SD = 0.82$, respectively), $F(1, 174) = 4.11$, $p < .05$ ($R^2 = .02$), thus supporting Hypothesis 3.

Table 1
Intercorrelations and Descriptive Statistics (Study 1)

	M	*SD*	1	2	3	4	5	6	7	8
1. Purpose[a]	1.5	0.50	—							
2. Variability in performance[b]	1.5	0.50	–.03	—						
3. Difficulty reactions	3.3	0.80	.18*	–.15*	.64					
4. Fairness reactions	2.7	0.71	–.07	.17*	–.32**	.76				
5. Confidence in providing feedback	3.3	0.88	.02	.04	–.31**	.63**	.80			
6. Self-efficacy for future use	5.0	2.2	–.04	.05	–.28**	.66**	.60**	.94		
7. Positive affectivity	5.1	0.92	.15	.02	–.01	.07	.04	.04	.73	
8. Negative affectivity	2.8	0.89	.03	–.02	–.02	–.07	.00	.02	–.32**	.83

Note: $N = 175$. Reliabilities (coefficient alpha) are listed on the diagonal.
a. Coded such that *no administrative consequence* = 1 and *administrative consequence* = 2.
b. Coded such that *lower variability* = 1 and *higher variability* = 2.
*$p < .05$, two-tailed. **$p < .01$, two-tailed.

The ANOVA on fairness revealed a significant effect for ratee variability, but not for purpose; again, there was no significant interaction. Although raters in the administrative consequence condition found the FDRS task to be somewhat less fair than those in the no administrative consequence condition ($M = 2.69$, $SD = 0.75$, vs. $M = 2.78$, $SD = 0.65$, respectively), this was not a significant difference, $F(1, 174) = 0.63$, $p > .05$; thus, Hypothesis 2 was not supported. However, those raters assigned ratees less variable in performance did find the FDRS task to be significantly less fair than those assigned ratees more variable in performance ($M = 2.61$, $SD = 0.76$, vs. $M = 2.85$, $SD = 0.64$, respectively), $F(1, 174) = 5.17$, $p < .05$ ($R^2 = .03$), thus supporting Hypothesis 4.

The Relationship Between Reactions and Confidence and Self-Efficacy

Hypotheses 5a and 5b and 6a and 6b predicted that raters' fairness and difficulty reactions, respectively, to the FDRS would be related to confidence in their ability to provide acceptable feedback to the ratees (Hypotheses 5a and 6a) and self-efficacy for using the FDRS in the future (Hypotheses 5b and 6b). Partial correlations, controlling for both PA and NA, supported Hypotheses 5a and 6a (partial r values = .58 and –.28, respectively, $p < .001$; $R^2 = .34$ and .08, respectively); those raters who thought the FDRS was more fair, and less difficult, had greater confidence in their ability to provide acceptable feedback to the ratees. Hypotheses 5b and 6b were also supported (partial r values = .62 and –.27, respectively, $p < .001$; $R^2 = .38$ and .07, respectively); those raters who thought the FDRS process was more fair had greater self-efficacy for future use of the FDRS, and those raters who thought the FDRS was more difficult had less self-efficacy for future use of the FDRS. Thus, as predicted, both perceived fairness and perceived difficulty were significantly associated with confidence in providing acceptable feedback to ratees and self-efficacy for future use of the FDRS system, even after controlling for trait affectivity.

Discussion

The primary goal of Study 1 was to test the effect of two proximal contextual variables (Cleveland et al., 1986; Levy & Williams, 2004; Murphy & Cleveland, 1991) on rater reactions to an FDRS. The results show effects consistent with anecdotal and interview accounts. Specifically, the FDRS task was perceived to be (a) more difficult when raters believed that such ratings were for administrative purposes and (b) more difficult and less fair when the pool of ratees had less variability in performance. Our results further suggest that both difficulty and fairness reactions are potentially important because they can affect raters' confidence in their ability to deliver acceptable feedback to ratees and their self-efficacy for future use of the system. These findings help to empirically establish the relevance of rater reactions as criteria in the PA domain (Levy & Williams, 2004; Murphy & Cleveland, 1995).

Thus, Study 1 makes several noteworthy contributions to the PA literature, particularly regarding the FDRS approach. In addition, there were a number of strengths in the design of this study, including (a) that the two contextual variables (purpose and variability) were manipulated and randomly assigned, thus controlling for alternative explanations, (b) that

participants believed real administrative consequences were attached to the ratings, and (c) that they did not realize they were participating in an experiment. Nonetheless, this first study has both methodological and scope limitations. Thus, we strongly felt that the inclusion of a second study explicitly designed to address these limitations would make a significant value-added contribution. We review each of these specific limitations below to set the stage for Study 2.

The most notable methodological limitations of Study 1 have to do with generalizability concerns. First, Study 1 does not directly measure *managers'* reactions to an FDRS. Although the MBA students used in this study might be representative of "real-world" managers in many respects (e.g., with regard to age, gender, and specific work experience), and although our results were replicated when confining our sample to only those raters with managerial experience, students are not managers. Second, the participants were asked to rate peers as opposed to subordinates, as in a more typical PA situation. Although there is precedent for using peer ratings in PA research (e.g., Bernardin, Cooke, & Villanova, 2000), the context of Study 1 did not allow us to recreate the authority and relational dynamics that might occur between a manager and his or her subordinates. Thus, confirming these results in a context in which managers rate their subordinates is critical. Third, in this rating context, the raters may have assumed that as they were rating others, they were also being rated. This situation may happen in organizations as well (e.g., 360-degree PA), but it is arguably not the norm in practice. Accordingly, the first goal of Study 2 was to confirm the findings from Study 1 in a context that did not have these generalizability constraints. Specifically, Study 2 involves actual working managers who are being asked to rate their current subordinates (and are not under the belief that they are simultaneously being rated by others).

Two additional limitations involve the operationalization of the contextual variables in Study 1. First, for the purpose variable, in some organizational settings (e.g., GE), termination decisions are tied to FDRS evaluations. Obviously, a grade in a course is not an equivalent consequence, even though, in this setting, a low grade (i.e., a C) could lead to probation and dismissal for a student, a salient consequence of some weight. Thus, Study 2 was deemed important for confirming the purpose effects using consequences that are more typical of PA ratings in organizations (e.g., development, promotions, termination).

Second, we chose in Study 1 to operationalize lower performance variability as all high performers, a circumstance that several managers identified in interviews with us as a challenge with FDRS and one that seems more probable, given the reasons mentioned previously (e.g., a "Lake Woebegone" bias on the part of managers, valid selection systems, previous termination or transfer of low-performing employees). However, it is also possible that subordinates may have reduced variability because they are all average or all below-average performers. Thus, in Study 2, we examine variability of subordinate performance in general, independent of level of performance. In addition, given that managers are rating their real subordinates, performance variability is measured rather than manipulated.

Study 2

The first goal of Study 2 was confirming the findings from Study 1 with a managerial sample using a different methodology (reviewed below). It was expected that Study 2's results would further confirm each of the hypotheses supported from Study 1:

Hypothesis 1: Raters completing the FDRS for purposes of making administrative decisions will see the FDRS task as more difficult than those completing it for nonadministrative purposes.[4]

Hypothesis 3: Raters who perceive their pool of ratees to have less variability in performance will see the FDRS task as more difficult than those perceiving their ratees to have more variability in performance.

Hypothesis 4: Raters who perceive their pool of ratees to have less variability in performance will see the FDRS task as less fair than those perceiving their ratees to have more variability in performance.

Hypotheses 5a and 5b: Fairness reactions will be *positively* related to confidence in one's ability to provide acceptable feedback to the ratees (Hypothesis 5a) and self-efficacy for using the FDRS in the future (Hypothesis 5b).

Hypotheses 6a and 6b: Difficulty reactions will be *negatively* related to confidence in one's ability to provide acceptable feedback to the ratees (Hypothesis 6a) and self-efficacy for using the FDRS in the future (Hypothesis 6b).

The second goal of Study 2 was an extension of Study 1, primarily via the inclusion of a manipulated rating format variable. That is, in Study 1, rating format (FDRS) was not manipulated but rather held constant. Although our express intent with Study 1 was to study reactions to *FDRS* ratings, this leaves the possibility that one would see the same effects of the purpose and variability factors on rater reactions with any type of rating format. If that is the case, Study 1's results are less informative about an FDRS specifically. In addition, holding this constant in Study 1 did not allow us to test whether an FDRS is perceived as more difficult and less fair than other PA formats, which is an implicit assumption in most writing on the topic and thus itself deserving of empirical testing.

This extension led to predictions regarding both main effects of format on reactions and interactions between format and the purpose and variability factors. We included two rating format conditions: FDRS and a more traditional rating scale (TRS) format, wherein supervisors indicate the performance of each employee on a rating scale (from 1 to 5). This comparison was chosen because of the relative "generic" nature of this type of rating scale and its frequent usage in organizations (Bernardin & Orban, 1990); a full 60% of our respondents indicated their organization currently uses a TRS or similar approach to appraise performance. In addition, both the FDRS and TRS used in the current study ask raters to rate or rank based on overall performance, thus not confounding format with dimensionality of performance ratings.

Because of the issues noted earlier regarding its forced categorization nature and the concomitant constraints placed on raters, and based on anecdotal reports and interviews with managers, it is expected that raters will report greater difficulty with rating decisions in the FDRS condition than in the TRS condition and that they will see the former as less fair.

Hypotheses 7a and 7b: Managers will perceive the FDRS format as more difficult (Hypothesis 7a) and less fair (Hypothesis 7b) than the TRS format.

In addition, we expect to find interactions between format and both purpose and variability, supporting our assumptions that there is something unique about the FDRS in its effects on rater reactions. First, regarding the purpose variable, we know that managers are more

"ticklish" (Longenecker et al., 1987) when making ratings when there are administrative consequences attached (also see Jawahar & Williams, 1997). However, we believe this is even more true in the case of an FDRS because the forced distribution constraints do not allow raters to "manage" the system (e.g., assigning high ratings to all subordinates so that everyone gets a raise) and could even force a distribution of ratings that does not accurately reflect merit. Thus, managers should hesitate, deliberate, and vacillate more in making their ratings. Conversely, in a TRS format, those distributional constraints do not exist; whatever the purpose or consequences of ratings are, managers can still assign all high ratings, all low ratings, or a mix, depending on what their goals are with regard to doling out consequences.

Hypothesis 8: There will be an interaction between rating format and purpose on difficulty reactions. Specifically, the effect of purpose (administrative or not) on perceived difficulty will be greater for FDRS raters than for TRS raters.[5]

Second, interactions are also expected between format and variability for both difficulty and fairness reactions. Similar to the above rationale, it is the case that the variability of the subordinate pool should be more of an issue for the FDRS than the TRS format. That is, if there is reduced variability in a manager's subordinate pool, under an FDRS system that manager is still required to assign people to each of the three performance levels. With the TRS format, however, there is nothing to stop that manager from assigning all similar ratings to his or her subordinates, should he or she choose to do so. Accordingly, we expect that reduced variability will have a larger effect on managers' perceptions of both difficulty and fairness of the rating task when using the FDRS as opposed to the TRS format.

Hypotheses 9a and 9b: There will be an interaction between rating format and perceived performance variability on difficulty (Hypothesis 9a) and fairness (Hypothesis 9b) reactions. Specifically, the effects of perceived variability on these reactions will be greater for FDRS raters than for TRS raters.

Method

Participants

The alumni database of a part-time executive MBA program at a large Midwestern university was used to contact potential participants who were currently working as managers. Approximately 400 were contacted via e-mail and invited to participate in an online study of managerial reactions to performance evaluation systems. Completed surveys were received from 125 of these managers, for a response rate of 34% (which is the same as the meta-analytic response rate for Web surveys reported by Cook, Heath, and Thompson [2000] and higher than the 21% reported by Kaplowitz, Hadlock, and Levine [2004]). To be included in this research, managers had to currently have subordinates (75% of the respondents did) or have had subordinates in the recent past (another 18% of the respondents) whose performance they were specifically responsible for evaluating. This criterion resulted in a final study *N* of 116.

The mean age of the managers was 41 years, and their mean managerial work experience was 11.64 years, indicating we had successfully obtained a sample of experienced managers. Managers had worked with their current organization for an average of 6.5 years, had worked in their current positions for an average of 3.1 years, and had on average 6.3 subordinates (SD = 3.2, range = 1 to 17). In terms of size of organization, 83% of the managers worked for a company with at least 500 employees.

Procedure

An online survey, which included demographic variables, the rating task, and reactions to the rating task (each described in more detail below), was created. A week before the survey was sent out, a "pre-notice" (Dillman, 2000) e-mail was sent to all managers describing the upcoming study and encouraging participation. A week after that, the e-mail containing a link to the online survey was sent to all managers. An additional e-mail reminder was sent 1 week later. Managers were not compensated for their participation but could request a copy of the results of the study in exchange for participating; 32% requested such summaries.

Manipulated Variables

Managers were randomly assigned to one of six versions of the rating task portion of the survey, resulting from the 2 (format: FDRS vs. TRS) × 3 (purpose: development only vs. promotion and raises vs. promotion, raises, *and* termination) design. For all versions of the rating task, managers were first asked to list the names of all of their subordinates for whom they had PA responsibility. This was done to make the rating task more salient to them and to focus their thinking on their current subordinates. They were then asked to imagine that their organization had decided to implement a new PA system, one described per the format and purpose manipulations below, and that they would have to use this system to evaluate their employees. After describing this new system per the information below, each manager was asked to actually complete the ratings or rankings for his or her set of subordinates.

Format. Those managers assigned to the FDRS condition were given the following instructions:

> The new procedure for rating your employees is called a "forced distribution" system, wherein you are required to assign a certain number of employees to each of three performance categories: *A* (the top 20% of your employees); *B* (the middle 70% of your employees); or *C* (the bottom 10% of your employees). Looking at your list of subordinates above, please assign each employee to one (and only one) performance category, in the percentages noted. For example, if you have 10 employees, you must assign 2 of them to the A category, 7 of them to the B category, and 1 of them to the C category. (If you have fewer than 10 subordinates, you still must assign one person to the C category.)

This is the same distribution used in Study 1 and mimics that commonly found in FDRS practice in corporate settings (see Blume et al., 2005, 2006; Sears & McDermott, 2003).

Those managers assigned to the TRS condition were given the following instructions:

> The new procedure for rating your employees requires you to evaluate each of them, one at a time, in terms of their overall performance, using the following rating scale: 1 = *far below expectations*, 2 = *somewhat below expectations*, 3 = *meets expectations*, 4 = *above expectations*, and 5 = *truly exceptional*. Looking at your list of subordinates above, please assign each employee an overall performance rating from 1 to 5 using this scale.

Purpose. We originally included three levels of the purpose variable, manipulated via the following instructions: "The ratings you give your subordinates will be used for . . ." (a) "developmental purposes only and will not be placed in the employees' permanent personnel files, nor will they be used for pay raises or promotions or other personnel actions" or (b) "the awarding of pay raises and promotions. That is, employees with higher ratings will be considered for raises and/or promotions, whereas those with lower ratings will not" or (c) "pay raises, promotions, and even termination decisions. That is, employees with higher ratings will be considered for raises and promotions, while employees in the lowest category of ratings will be targeted for termination." For each of these, managers were reminded of the assigned purpose with another statement immediately before the box in which they were to make their ratings or rankings. A manipulation check found that 83% of the managers correctly recalled their assigned purpose. Those not correctly identifying their purpose condition were excluded from further analysis, resulting in a final *N* of 96 (i.e., 83% of 116) for hypothesis testing.

For testing Hypotheses 1 to 6, we collapsed these three levels into two: no administrative consequences (the developmental condition) versus administrative consequences (raises or promotions versus termination) to parallel Study 1's conditions. For the extension part of Study 2, we analyzed all three levels in an exploratory fashion to see whether making termination explicit as an administrative consequence affected reactions beyond promotion consequences. However, finding no significant differences on reactions between the two administrative conditions, we opted to keep purpose dichotomized (administrative consequences vs. no administrative consequences) for all Study 2 hypotheses.

Measures

Perceived variability of subordinate performance. Managers responded to one item indicating how they would characterize the performance of their group of subordinates: "They all perform at about the same level; levels of performance vary somewhat from employee to employee; or levels of performance vary a great deal from employee to employee." Because of a very skewed distribution across these three levels (i.e., 8%, 52%, and 40%, respectively), and to replicate the variability conditions of Study 1, two levels of variability were created from this item: lower variability (the first two response options; 60% of respondents) and higher variability (the third response option; 40% of respondents). Follow-up analyses, treating the three levels separately, did not change the results.

Difficulty. The difficulty measure from Study 1 could not be used in Study 2 because it was specific to an FDRS task (i.e., asked about ease or difficulty of making assignments to the A, B, and C categories) and therefore would not apply across both rating formats. Thus, we used an alternate measure of difficulty (see Cheung et al., 1999; Shiloh et al., 2001; Zhang & Mittal, 2005) appropriate for both rating formats. This scale contained four bipolar sets of adjectives (on 7-point scales): "easy–difficult," "required little deliberation–required much deliberation," "no hesitation about these decisions–considerable hesitation about these decisions," and "straightforward–complicated." These four items were averaged to create one difficulty score ($\alpha = .88$), with higher numbers indicating greater difficulty.

We also included a measure of time to complete the rating task as an additional dimension of difficulty. This variable provides an alternative assessment of the extent to which managers may have hesitated, deliberated, or vacillated in making their ratings (Chatterjee & Heath, 1996; Houston, Sherman, & Baker, 1991) but one that is more behavioral than attitudinal. As expected, the time variable was significantly and positively correlated with perceived difficulty ($r = .22$, $p < .05$). Immediately after making the performance ratings, respondents were asked, "Approximately how many minutes did it take you to do the requested performance ratings?" Estimates ranged from 1 to 10 minutes, with a mean of 5.1 ($SD = 2.4$).

Fairness. In addition to the fairness measure developed and used in Study 1, which was slightly reworded here to apply to subordinates ($\alpha = .83$), we also included Colquitt's (2001) measures of procedural and distributive fairness (again, slightly reworded to apply to a subordinate rating task). The procedural fairness scale included six items (e.g., "The rating system allowed you to be free of bias," "The procedures used to rate your subordinates were based on accurate information"), and the distributive fairness scale included four items (e.g., "Your ratings were appropriate given the work completed by your subordinates," "Your ratings were justified, given your subordinates' performance"). Respondents agreed or disagreed to all items on a 5-point scale. Alphas were .76 and .86, respectively.

Confidence in providing acceptable feedback and self-efficacy for future use of rating system. For these variables, we used the same two measures from Study 1, each containing three items ($\alpha = .84$ and .89, respectively).

Control variables. We included the same measures of PA and NA from Study 1 to control self-report relationships for general affectivity ($\alpha = .70$ for PA and $\alpha = .72$ for NA). In addition, given that number of ratees was not constant across raters, as it was in Study 1, yet could affect reactions to the rating task, especially difficulty and time to complete, we also controlled for number of subordinates in all analyses ($M = 6.3$, range = 1 to 17).

Results and Discussion

Table 2 reports descriptive statistics and intercorrelations for the Study 2 variables.

Table 2
Intercorrelations and Descriptive Statistics for All Participants (Study 2)

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Purpose[a]	1.70	0.46	—											
2. Variability in performance[b]	1.33	0.47	.05	—										
3. Number of subordinates	6.30	3.20	–.08	.42**	—									
4. Difficulty reactions	2.39	1.19	.21*	–.29**	–.14	.88								
5. Time to complete ratings (minutes)	5.10	2.40	.11	.16	.24*	.22*	—							
6. Fairness reactions	2.79	0.91	–.02	–.01	.20	–.12	.00	.83						
7. Procedural fairness	3.07	0.71	–.09	.11	.18	–.19	–.05	.77**	.76					
8. Distributive fairness	3.45	0.84	–.10	.19	.08	–.21*	–.23*	.57**	.60**	.86				
9. Confidence in providing feedback	3.22	0.86	–.09	.13	.17	–.27*	–.18	.59**	.64**	.56**	.84			
10. Self-efficacy for future use	3.03	1.02	–.04	.09	.08	.01	–.05	.78**	.69**	.64**	.67**	.89		
11. Positive affectivity	3.90	0.44	.08	.08	.11	.01	–.01	–.06	.07	.02	.03	–.12	.70	
12. Negative affectivity	2.39	0.44	–.04	–.02	.10	.16	.05	.12	.11	.11	.11	.22*	–.36**	.72
13. Rating format[c]	1.48	0.50	.04	.15	.01	–.19	–.21	.00	–.04	.24*	.26*	.07	.04	.01

Note: $N = 96$. Reliabilities (coefficient alpha) are listed on the diagonal.
a. Coded such that *no administrative consequence* = 1 and *administrative consequence* = 2.
b. Coded such that *lower variability* = 1 and *higher variability* = 2.
c. Coded such that *forced distribution rating system* = 1 and *traditional rating scale* = 2.
*$p < .05$, two-tailed. **$p < .01$, two-tailed.

Confirming Study 1's Hypotheses

To parallel Study 1, these hypotheses (Hypothesis 1 to Hypothesis 6) were tested only on the managers in the FDRS condition. In addition, one-tailed tests were used for these hypotheses, given their confirmatory and thus directional nature; however, most were also significant at two-tailed levels. Finally, number of subordinates was controlled for in each analysis; this was true for both the confirmation and extension hypotheses. Below, we report the results for these statistical tests, including R^2 values, cell means, and standard deviations for those means.

First, an ANCOVA on difficulty, with number of subordinates as the covariate, revealed a significant effect of purpose, $F(1, 45) = 3.86$, $p < .05$ ($R^2 = .08$), with FDRS managers in the administrative consequence condition finding the rating task significantly more difficult than those in the no administrative consequence condition ($M = 2.84$, $SD = 1.30$, vs. $M = 1.88$, $SD = 0.68$, respectively). Thus, Hypothesis 1 was supported. Similarly, an ANCOVA on the time dependent variable also showed a significant effect of purpose, $F(1, 45) = 4.89$, $p < .05$ ($R^2 = .10$), with those FDRS managers in the administrative consequence condition taking significantly more time to complete their ratings than those in the no administrative consequence condition ($M = 6.03$, $SD = 2.80$, vs. $M = 4.28$, $SD = 1.65$, respectively).

Second, regarding Hypothesis 3, tested via ANCOVA, FDRS managers who perceived their subordinates' performance to be less variable found the rating task more difficult than those who perceived their subordinates' performance to be more variable ($M = 2.69$, $SD = 1.23$, vs. $M = 1.90$, $SD = 1.16$, respectively), and this difference was significant, $F(1, 45) = 2.63$, $p < .05$ ($R^2 = .06$); thus, Hypothesis 3 was supported. In support of Hypothesis 4, tested via MANCOVA because of the multiple fairness variables, FDRS managers who perceived their subordinates' performance to be less variable also found the rating task to be less fair, $F(3, 42) = 3.59$, $p < .05$. Univariate analyses indicated that this was true in terms of procedural fairness ($M = 3.03$, $SD = 0.86$, vs. $M = 3.56$, $SD = 0.80$), $F(1, 45) = 3.68$, $p < .05$ ($R^2 = .08$) and distributive fairness ($M = 3.14$, $SD = 1.04$, vs. $M = 3.77$, $SD = 0.78$), $F(1, 45) = 3.65$, $p < .05$ ($R^2 = .08$); the difference on the other fairness measure, although in the proper direction ($M = 2.73$, $SD = 1.02$, vs. $M = 3.25$, $SD = 0.76$), was not significant, $F(1, 45) = 2.23$, $p = .07$ ($R^2 = .05$). Thus, to summarize, the effects of the manipulated variables from Study 1 were confirmed in Study 2, wherein we found significant effects for both purpose and variability on reactions, supporting Hypotheses 1, 3, and 4.

Third, Hypotheses 5a, 5b, 6a, and 6b were tested controlling for PA and NA as well as number of subordinates. Both fairness and difficulty reactions of the FDRS managers were related to confidence in one's ability to provide acceptable feedback to subordinates and to the managers' self-efficacy for using a system like this in the future (see Table 3). Thus, all of these hypotheses were supported and replicated the results from Study 1, with similar magnitudes as well. Specifically, fairness perceptions were positively and significantly related to confidence in providing feedback (partial $r = .65$, $p < .001$; $R^2 = .42$) and to self-efficacy for future use of the system (partial $r = .85$, $p < .001$; $R^2 = .72$), supporting Hypotheses 5a and 5b. Table 3 also lists the results for the other two fairness measures, procedural and distributive, which paralleled these findings. In addition, difficulty perceptions were negatively and significantly related to confidence in providing feedback (partial

Table 3
Relationships Between Difficulty and Fairness Reactions and Confidence and Self-Efficacy (Hypotheses 5 and 6 in Study 2)

Reactions	Confidence in Providing Feedback	Self-Efficacy for Future Use
Difficulty	–.35**	–.27*
Fairness	.65**	.85**
Procedural fairness	.65**	.79**
Distributive fairness	.58**	.71**

Note: Correlations are controlled for both positive and negative affectivity and for number of subordinates.
*$p < .05$, one-tailed. **$p < .01$, one-tailed.

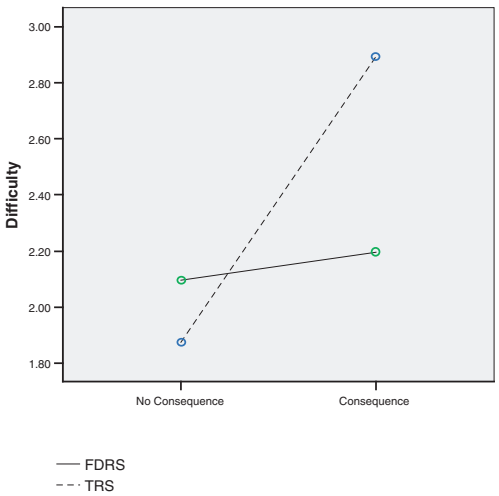
$r = –.35$, $p < .01$; $R^2 = .12$) and to self-efficacy for future use of the system (partial $r = –.27$, $p < .05$; $R^2 = .07$), supporting Hypotheses 6a and 6b. Thus, the FDRS results from Study 2 consistently confirmed those from Study 1 (7 out of 7 hypotheses were supported), indicating that the contextual variables of purpose and performance variability can affect rater reactions to FDRS, which in turn can have repercussions for the future use of the system. We return to the implications of this in the general discussion).

Extension Hypotheses

The results for the extension hypotheses are reported below, providing information on the F tests, R^2 values, cell means, and standard deviations for those means. Hypotheses 7a and 7b, which predicted greater difficulty and lower fairness reactions, respectively, with the FDRS format than with the TRS format, were tested via ANCOVA (Hypothesis 7a) and MANCOVA (Hypothesis 7b), with number of subordinates as the covariate. These hypotheses were supported for each of the reaction variables of interest. Specifically, the FDRS (a) took longer to complete than the traditional rating format ($M = 5.6$, $SD = 2.72$, vs. $M = 4.4$, $SD = 2.11$), $F(1, 92) = 4.08$, $p < .05$ ($R^2 = .04$), (b) was more difficult for raters ($M = 2.62$, $SD = 1.22$, vs. $M = 2.12$, $SD = 1.07$), $F(1, 92) = 4.01$, $p < .05$ ($R^2 = .04$), and (c) was perceived as less fair by raters, $F(3, 87) = 3.56$, $p < .05$ ($R^2 = .04$).

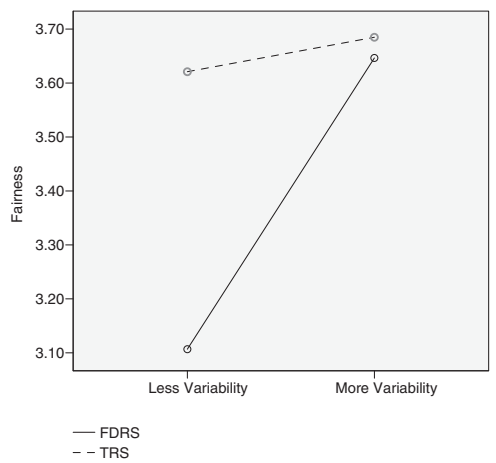
Hypothesis 8 predicted an interaction between rating format and purpose on difficulty reactions. The ANCOVA results supported this prediction, $F(1, 90) = 3.91$, $p < .05$ ($R^2 = .04$). As Figure 1 shows, having an administrative consequence attached to the ratings made little difference in terms of the difficulty of the TRS task ($M = 2.20$, $SD = 1.10$, vs. $M = 2.10$, $SD = 0.80$, for administrative consequence and no administrative consequence conditions, respectively; post hoc comparisons showed that this was a nonsignificant difference, $t = –.30$, $p > .05$). However, having an administrative consequence attached to the ratings made a significantly larger difference in terms of the difficulty of the FDRS rating task ($M = 2.89$, $SD = 1.28$, vs. $M = 1.88$, $SD = 0.68$, respectively; post hoc comparison $t = –2.45$, $p < .01$). Thus, this pattern supports Hypothesis 8. We also tested this interaction hypothesis for the time dependent variable and found parallel results, $F(1, 91) = 5.55$, $p < .05$ ($R^2 = .06$). The interaction patterns for both of these dependent variables suggest an interesting conclusion:

Figure 1
Plot of Significant Rating Format × Purpose Interaction on Perceived Difficulty (Hypothesis 8)



Note: FDRS = forced distribution rating system; TRS = traditional rating scale.

Figure 2
Plot of Significant Rating Format × Variability Interaction on Perceived Fairness (Hypothesis 9)



Note: FDRS = forced distribution rating system; TRS = traditional rating scale.

that the FDRS is somewhat faster and easier than the TRS format when no administrative consequences are attached, but under conditions of administrative consequences, the FDRS is perceived as significantly more difficult and takes longer to do.

Finally, Hypotheses 9a and 9b predicted an interaction between rating format and variability for both difficulty (Hypothesis 9a) and fairness (Hypothesis 9b) reactions. There was no significant interaction for the difficulty variable, $F(1, 90) = 0.08$, $p > .05$; thus, Hypothesis 9a was not supported. However, there was a significant interaction between rating format and perceived variability on fairness, thus supporting Hypothesis 9b. This was supported for the fairness measure from Study 1, $F(1, 90) = 3.86$, $p < .05$ ($R^2 = .04$), and the procedural fairness measure, $F(1, 90) = 3.90$, $p < .05$ ($R^2 = .04$), but not for the distributive fairness measure, $F(1, 90) = 2.28$, $p = .13$. Both of these significant interactions were similar in form (see Figure 2) and indicated that there is a greater effect of perceived variability on perceptions of fairness for the FDRS than for the TRS format, as expected.

General Discussion and Conclusions

The current research makes several distinct contributions to the PA literature. First, we examined an approach to PA that has become increasingly popular in organizations yet has received very little empirical attention in the research literature. In fact, we were able to identify only one other published empirical article (and that was a simulation study; Scullen et al., 2005) and two recent conference presentations (Blume et al., 2005, 2006) on this topic. Although Scullen et al.'s (2005) findings suggested the likely diminishing returns of this type of rating system over time in terms of improved employee and organizational performance, and although Blume et al. (2005, 2006) examined ratee reactions to an FDRS, there has been no empirical research on raters who are asked to use this type of PA rating approach. As Blume et al. (2006) have noted, "A conspicuous gap [exists] in the empirical data with respect to the perceptions and behaviors of the *raters* (not just ratees) involved in an FDRS" (p. 29). Because it was discussion with managers that prompted us to study FDRS in the first place, and because of the relative dearth of research on rater reactions cited previously, we focused our research on raters. We identified two central and relevant dimensions of rater reactions, difficulty and fairness, by reviewing published anecdotal accounts, qualitative data collected from raters using an FDRS in practice, and literature in other areas. In addition, we were able to empirically establish the relevance of these two specific reactions by demonstrating their links to confidence in one's ability to deliver acceptable feedback to the ratees and self-efficacy for future use of the system. Interestingly, there has been an oft-noted gap in the PA area in terms of which criteria are of most interest to practitioners (e.g., appraisal reactions; Cardy & Dobbins, 1994; Murphy & Cleveland, 1995) versus which criteria tend to be studied most often in the research literature (e.g., rater behavior and the psychometric qualities of ratings; Balzer & Sulsky, 1990; Keeping & Levy, 2000). Addressing this gap, by explicitly examining rater reactions, marks the second contribution of our research.

Third, we established the importance of proximal contextual variables (Cleveland et al., 1986; Levy & Williams, 2004; Murphy & Cleveland, 1991) in determining rater reactions to an FDRS. It has been noted that, in practice, an FDRS is implemented in a variety of ways

(Bates, 2003; Blume et al., 2006), a point confirmed by our interviews with FDRS managers, and the current research clearly suggests that such situational differences can significantly affect raters' reactions. Specifically, the two studies were consistent in suggesting that FDRS ratings are (a) more difficult when there are administrative consequences attached and (b) more difficult and perceived to be less fair when there is less variability in performance among the ratees. These findings have clear practical implications for the use of an FDRS in organizations, a point to which we return in the following section. In addition, the interaction findings from Study 2 suggest these differences in specifics of the rating system are more important with the FDRS format than with a TRS format.

Practical Implications

As Taylor et al. (1998) have noted, "Because managers play such a major role in administering HR systems to their employees, their reactions become paramount" (p. 568). Thus, our findings regarding how such reactions can be influenced by PA system factors have some important implications for organizations considering an FDRS.

The first implication is, frankly, such organizations should expect some resistance from managers. Results from Study 2 clearly indicated that managers found the FDRS more difficult and less fair than a TRS format; organizations should be prepared for that. In fact, this resistance may even be viewed as a sign that an FDRS should not be used. However, because the jury is still out with regard to the ultimate effectiveness of an FDRS approach, we confine the remainder of our recommendations to what the current findings suggest regarding ways to mitigate negative reactions to FDRS.

Specifically, we recommend that organizations consider initially attaching less severe administrative consequences to the FDRS evaluations (e.g., avoid making promotion or termination decisions contingent on rankings, at least initially). Our findings across both studies regarding the purpose variable suggest that such a strategy should lead to more positive reactions on the part of raters, which in turn would build their self-efficacy and confidence for future use of the system. In other words, it would give the raters a chance to get comfortable with this process before attaching more severe administrative consequences to their ratings. Although this might be good advice for any new PA intervention, our interaction results from Study 2 suggest this is particularly important for an FDRS. The fact that our findings parallel those from Blume et al. (2006), in which *ratees* also had more positive reactions to an FDRS that had less severe consequences, further strengthens the wisdom of this advice.

Finally, our results also suggest that only organizations, and units within organizations, with reason to believe there is significant variability in performance should use a rating format like FDRS. Indeed, the effects of reduced ratee variability in the current research were particularly strong and consistent. These findings gain increased importance when considered alongside the fact that the longer an FDRS has been in place in an organization, the more uniform (i.e., less variable) ratees' levels of performance are likely to be (Scullen et al., 2005). This is the result of a dual process operating: (a) the lower-performing employees have been terminated under this system and (b) some higher-performing employees may be more likely to leave under this system, as some research suggests

(McBriarty, 1988; Zenger, 1992). Thus, the admonition to ensure that sufficient variability exists before using an FDRS may become particularly important in subsequent iterations of an FDRS in organizations.

Limitations and Future Research Directions

Although we see our research as having a number of strengths, in terms of both contributions to the literature and methodology, there also remain several important limitations to acknowledge. Here, we review these possible limitations and the corresponding directions for future research. The most salient and important limitations stem from the fact that we admittedly did not conduct the "ideal" version of this research, which would have entailed manipulating features of an organization's intact FDRS that managers were currently using to evaluate their subordinates. Collecting data within an operational FDRS in an organizational setting would have provided several improvements over our study. First, and most notably, it would have increased the level of accountability felt by raters. That is, in Study 1, although there was some accountability in terms of grades being attached to the FDRS outcomes, this level of accountability is certainly less than that typically felt by managerial raters in organizations. In Study 2, the accountability was merely hypothetical. Thus, this represents a limitation to our studies, one that should be remedied in future research.

Second, the targeted rating behavior in Study 1 was classroom participation. This may differ in important ways from the typical targets of rating in operational PAs (e.g., leadership, decision making). Although this criticism is less applicable in Study 2 (which had overall performance as the target of rating), it does raise the interesting question of how rater reactions might differ for different types of competencies or targets of rating. One intriguing possibility is that raters may find it easier to complete an FDRS on specific competencies than on overall performance. We would encourage future research aimed at investigating this question.

Third, FDRS research in organizational settings could also build on our findings regarding the purpose variable. Specifically, one might question the generalizability of our operationalization of purpose in Study 2, which included FDRS for developmental purposes (which may be a less typical reason for doing FDRS in organizations). However, it is important to clarify that conclusions from our research regarding the purpose variable should be interpreted as differences between administrative versus nonadministrative purposes, not administrative versus developmental purposes. That is, the nonadministrative purpose category identified in previous PA literature includes both PA done for research (as in our Study 1) and developmental purposes (as in our Study 2) (Jawahar & Williams, 1997). Thus, although operationalized differently across the two studies, both studies compared nonadministrative purposes to administrative purposes. Nonetheless, future research in organizations that takes a more fine-grained approach to rating purpose (e.g., by separating research from developmental from multiple administrative purposes) would further build on the limited operationalizations of purpose employed here.

A final limitation of our study that could be improved on when studying FDRS in organizations involves our difficulty dependent variable. In our study, we anticipated that difficulty would have negative effects on raters and examined only those issues. However, the possibility remains that in organizational settings there could also be positive outcomes

associated with increased perceptions of difficulty (e.g., leading the rater to think more deeply about, or be more careful with, the ratings). Thus, although difficulty may have some negative effects on raters, it may also be associated with managers providing less lenient or more accurate ratings. Future research should take a more comprehensive approach to understanding the full range of outcomes that are associated with raters seeing FDRS as a difficult task.

Future research should also endeavor to study more "macro" (Tziner et al., 2005) or distal (Levy & Williams, 2004) contextual factors (e.g., organizational culture) likely to affect rater reactions (see Blume et al., 2005, p. 19). For example, an FDRS may fit better with some cultures than with others (Guralnik et al., 2004). Finally, for reasons mentioned earlier, our initial focus was on rater reactions to FDRS. Nonetheless, we would strongly encourage dyadic research that examines both rater and ratee reactions to an FDRS (see Levy & Williams, 2004). If the FDRS works to break down trust and leader–member exchange (Graen & Scandura, 1987; Graen & Uhl-Bien, 1995), this type of PA system may be an organization's undoing rather than its strategic advantage.

Notes

1. Because it arguably represents the more typical situation (as identified in our managerial interviews), in Study 1 we manipulate low performance variability as all high performers. However, our hypotheses are with regard to variability per se and are expected to hold across levels of performance. We return to this issue in Study 2.

2. Of importance, students were randomly assigned by the administration to these various sections (i.e., students themselves did not choose their sections), and all four sections met on the same days. In addition, there were no differences observed on any personality or demographic variables across sections or across consequence conditions (for all F values $p > .05$).

3. Given how the ratee lists were created and disseminated, approximately 15% of the participants would have received a list of ratees that included their own name. So that we could code for this, we asked participants, after the debriefing but before collecting their responses, to place an asterisk on the front cover of their packets if they had been asked to rate themselves in this task. We ran the analyses both controlling for this variable and excluding those that had been asked to rate themselves, and the results did not change from those reported here.

4. Hypothesis 2, regarding an effect of purpose on fairness, was not supported in Study 1 and thus is not formally hypothesized here. However, we did test this in Study 2, and, paralleling Study 1, we found no significant effect of purpose on fairness.

5. Because Study 1 did not find an effect of purpose on fairness, we do not predict an interaction between format and purpose for the fairness dependent variable in Study 2. That is, if purpose did not significantly affect fairness in the forced distribution rating system (FDRS), it makes little sense to expect that this effect would be stronger for FDRS than for the traditional rating scale. The data from Study 2 do in fact show, as expected, a nonsignificant interaction.

References

Balzer, W. K., & Sulsky, L. M. 1990. Performance appraisal effectiveness. In K. R. Murphy & F. E. Saal (Eds.), *Psychology in organizations: Integrating science and practice*: 133-156. Hillsdale, NJ: Lawrence Erlbaum.

Bandura, A. 1977. *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. 1982. Self-efficacy mechanisms in human agency. *American Psychologist*, 37: 122-147.

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

Bannister, B. D., & Balkin, D. B. 1990. Performance evaluation and compensation feedback messages: An integrated model. *Journal of Occupational Psychology*, 63: 97-111.

Bates, S. 2003. Forced rankling. *HR Magazine*, 48: 62-68.

Bernardin, H. J., Cooke, D. K., & Villanova, P. 2000. Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, 85: 232-236.

Bernardin, H. J., & Orban, J. A. 1990. Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology*, 5: 197-211.

Blume, B. D., Baldwin, T. T., & Ruben, R. S. 2005. *Forced ranking: Who is attracted to it? A study of performance management system preferences*. Paper presented at the 65th annual Academy of Management Conference, Honolulu, HI.

Blume, B. D., Baldwin, T. T., & Ruben, R. S. 2006. *All forced distribution systems are not created equal: A policy capturing study*. Paper presented at the 66th annual Academy of Management Conference, Atlanta, GA.

Boyle, M. 2001. Performance reviews: Perilous curves ahead. *Fortune*, 143(11): 187.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. 2001. Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, 1: 276-298.

Brett, J. F., & Atwater, L. E. 2001. 360-degree feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology*, 86: 930-942.

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

Brown, M., & Benson, J. 2003. Rated to exhaustion? Reactions to performance appraisal processes. *Industrial Relations Journal*, 34: 67-81.

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

Chatterjee, S., & Heath, T. B. 1996. Conflict and loss aversion in multiattribute choice: The effects of trade-off size and reference dependence on decision difficulty. *Organizational Behavior and Human Decision Processes*, 67: 144-155.

Cheung, S. F., Chan, D. K.-S., & Wong, Z. S.-Y. 1999. Reexamining the theory of planned behavior in understanding wastepaper recycling. *Environment and Behavior*, 31: 587-612.

Cleveland, J. N., Morrison, R., & Bjerke, D. 1986. *Rater intentions in appraisal ratings*: *Malevolent manipulation or functional fudging*. Paper presented at the first annual conference of the Society for Industrial and Organizational Psychology, Chicago.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. 1989. Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, 74: 130-135.

Cobb, A. T., & Frey, F. M. 1996. The effects of leader fairness and pay outcomes on superior/subordinate relations. *Journal of Applied Social Psychology*, 26: 1401-1426.

Colquitt, J. A. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, 86: 386-400.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C., & Ng, K. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, 86: 425-445.

Condrey, S. E. 1995. Reforming human resource management systems: Exploring the importance of organizational trust. *American Review of Public Administration*, 25: 341-354.

Cook, C., Heath, F., & Thompson, R. L. 2000. A meta-analysis of response rates in Web- or Internet-based surveys. *Educational and Psychological Measurement*, 60: 821-836.

Deckop, J. R., Mangel, R., & Cirka, C. C. 1999. Getting more than you pay for: Organizational citizenship behavior and pay-for-performance plans. *Academy of Management Journal*, 42: 420-428.

Dickson, J. M., & MacLeod, A. K. 2004. Anxiety, depression, and approach and avoidance goals. *Cognition and Emotion*, 18: 423-430.

Diener, E., Smith, H., & Fujita, F. 1995. The personality structure of affect. *Journal of Personality and Social Psychology*, 69: 130-141.

Dillman, D. A. 2000. *Mail and Internet surveys*: *The tailored design method*. New York: John Wiley.

Elliot, A. J., & Thrash, T. M. 2002. Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology*, 82: 804-818.

Erdogan, B. 2002. Antecedents and consequences of justice perceptions in performance appraisals. *Human Resource Management Review*, 12: 555-578.

Erdogan, B., Kraimer, M. L., & Liden, R. C. 2001. Procedural justice as a two-dimensional construct: An examination in the performance appraisal account. *Journal of Applied Behavioral Science*, 37(2): 205-222.

Fedor, D. B., Bettenhausen, K. L., & Davis, W. 1999. Peer reviews: Employees' dual roles as raters and recipients. *Group & Organization Management*, 24: 92-120.

Findley, H. M., Giles, W. F., & Mossholder, K. W. 2000. Performance appraisal process and system facets: Relationships with contextual performance. *Journal of Applied Psychology*, 85: 634-640.

Fletcher, C. 2001. Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology*, 74: 473-487.

Folger, R., Konovsky, M. A., & Cropanzano, R. 1992. A due process metaphor for performance appraisal. *Research in Organizational Behavior*, 14: 129-177.

Fusaro, P. C., & Miller, R. M. 2002. *What went wrong at Enron*: *Everyone's guide to the largest bankruptcy in U.S. history*. Hoboken, NJ: John Wiley.

Gati, I., Krausz, M., & Osipow, S. H. 1996. A taxonomy of difficulties in career decision making. *Journal of Counseling Psychology*, 43: 510-526.

Graen, G., & Scandura, T. 1987. Toward a psychology of dyadic organizing. *Research in Organizational Behavior*, 9: 175-208.

Graen, G. B., & Uhl-Bien, M. 1995. Relationship-based approach to leadership: Development of leader-member-exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *Leadership Quarterly*, 6: 219-247.

Greenberg, J. 1986. Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology*, 71: 340-342.

Greenberg, J. 1990. Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology*, 75: 561-568.

Guralnik, O., Rozmarin, E., & So, A. 2004. Forced distribution: Is it right for you? *Human Resource Development Quarterly*, 15: 339-345.

Harris, M. M. 1994. Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management*, 20: 737-756.

HayGroup. 2002. *Achieving outstanding performance through a "culture of dialogue."* Working paper, HayGroup, Philadelphia.

Hillman, C. H., Rosengren, K. S., & Smith, D. P. 2004. Emotion and motivated behavior: Postural adjustments to affective picture viewing. *Biological Psychology*, 66: 51-62.

Houston, D. A., Sherman, S. J., & Baker, S. M. 1991. Feature matching, unique features, and the dynamics of the choice process—Predecision conflict and postdecision satisfaction. *Journal of Experimental Social Psychology*, 27: 411-430.

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

Jawahar, I. M., & Williams, C. R. 1997. Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, 50: 905-925.

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. 2004. A comparison of Web and mail survey response rates. *Public Opinion Quarterly*, 68: 94-102.

Keeping, L. M., & Levy, P. E. 2000. Performance appraisal reactions: Measurement, modeling, and method bias. *Journal of Applied Psychology*, 85: 708-723.

Korsgaard, M. A., & Roberson, L. 1995. Procedural justice in performance evaluation: The role of instrumental and non-instrumental voice in performance appraisal discussions. *Journal of Management*, 21: 657-669.

Lee, M., & Son, B. 1998. The effects of appraisal review content on employees' reactions and performance. *International Journal of Human Resource Management*, 9: 203-214.

Levy, P. E., & Williams, J. R. 2004. The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30: 881-905.

London, M., Casey, J., Chatterjee, S., & Hurley, A. 1997. Effects of information frame, response frame, and goal on personnel decision making. *Journal of Psychology*, 131: 225-240.

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

Maroney, B. P., & Buckely, M. R. 1992. Does research in performance appraisal influence the practice of performance appraisal? Regretfully not! *Public Personnel Management*, 21: 185-196.

McBriarty, M. A. 1988. Performance appraisal: Some unintended consequences. *Public Personnel Management*, 17: 421-434.

Meisler, A. 2003. Dead man's curve. *Workforce*, 82: 44-49.

Meyer, H. H., Kay, E., & French, R. P. 1965. Split roles in performance appraisal. *Harvard Business Review*, 43: 123-129.

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

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

Nalbandian, J. 1981. Performance appraisal: If only people were not involved. *Public Administration Review*, 41: 392-396.

Napier, N. K., & Latham, G. P. 1986. Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology*, 39: 827-837.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88: 879-903.

Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. 2005. Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology*, 59: 1-32.

Sears, D., & McDermott, D. 2003. The rise and fall of rank and yank. *Information Strategy: The Executive's Journal*, 19: 6-11.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. 1987. Emotion knowledge—Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52: 1061-1086.

Shiloh, S., Koren, S., & Zakay, D. 2001. Individual differences in compensatory decision-making style and need for closure as correlates of subjective decision complexity and difficulty. *Personality and Individual Differences*, 30: 699-710.

Smither, J. W. 1998. Lessons learned: Research implications for performance appraisal and management practice. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice*: 537-548. San Francisco: Jossey-Bass.

Taylor, M. S., Masterson, S. S., Renard, M. K., & Tracy, K. B. 1998. Managers' reactions to procedurally just performance management systems. *Academy of Management Journal*, 41: 568-579.

Taylor, M. S., Tracy, K. B., Renard, M. K., Harrison, J. K., & Carroll, S. J. 1995. Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly*, 40: 495-523.

Terry, D. J., & O'Leary, J. E. 1995. The theory of planned behaviour: The effects of perceived behavioural control and self-efficacy. *British Journal of Social Psychology*, 34: 199-220.

Thomas, S. L., & Bretz, R. D. 1994. Research and practice in performance appraisal: Evaluating employee performance in America's largest companies. *SAM Advanced Management Journal*, 59: 28-34.

Townley, B. 1999. Practical reason and performance appraisal. *Journal of Management Studies*, 36: 287-306.

Trafimow, D., Sheeran, P., Conner, M., & Finlay, K. A. 2002. Evidence that perceived behavioural control is a multidimensional construct: Perceived control and perceived difficulty. *British Journal of Social Psychology*, 41: 101-121.

Tziner, A., & Kopelman, R. E. 2002. Is there a preferred performance rating format? A non-psychometric perspective. *Applied Psychology: An International Review*, 51(3): 479.

Tziner, A., Kopelman, R., & Joanis, C. 1997. Investigation of raters' and ratees' reactions to three methods of performance appraisal: BOS, BARS, and GRS. *Canadian Journal of Administrative Sciences*, 14(4): 396.

Tziner, A., Murphy, K. R., & Cleveland, J. N. 2005. Contextual and rater factors affecting rating behavior. *Group & Organization Management*, 30: 89-98.

Welch, J., & Welch, S. 2005. *Winning*. New York: Harper Business.

Zakay, D., & Wooler, S. 1984. Time pressure, training and decision effectiveness. *Ergonomics*, 27: 273-284.

Zenger, T. R. 1992. Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, 37: 198-219.

Zhang, Y., & Mittal, V. 2005. Decision difficulty: Effects of procedural and outcome accountability. *Journal of Consumer Research*, 32: 465-472.