Journal of Applied Psychology
2017, Vol. 102, No. 3, 421–433

0021-9010/17/$12.00 http://dx.doi.org/10.1037/apl0000085

Performance Appraisal and Performance Management: 100 Years of Progress?

Angelo S. DeNisi
Tulane University

Kevin R. Murphy
University of Limerick

We review 100 years of research on performance appraisal and performance management, highlighting the articles published in JAP, but including significant work from other journals as well. We discuss trends in eight substantive areas: (1) scale formats, (2) criteria for evaluating ratings, (3) training, (4) reactions to appraisal, (5) purpose of rating, (6) rating sources, (7) demographic differences in ratings, and (8) cognitive processes, and discuss what we have learned from research in each area. We also focus on trends during the heyday of performance appraisal research in JAP (1970-2000), noting which were more productive and which potentially hampered progress. Our overall conclusion is that JAP's role in this literature has not been to propose models and new ideas, but has been primarily to test ideas and models proposed elsewhere. Nonetheless, we conclude that the papers published in JAP made important contribution to the filed by addressing many of the critical questions raised by others. We also suggest several areas for future research, especially research focusing on performance management.

Keywords: performance, criteria, ratings, appraisal, evaluation

Supplemental materials: http://dx.doi.org/10.1037/apl0000085.supp

The assessment of people's performance at work—performance appraisal—has been of interest to scholars and practitioners for literally hundreds of years. More recently, there has also been a growing interest in the process of managing performance. The two topics are clearly related, but they are not identical. Performance *appraisal* refers to a formal process, which occurs infrequently, by which employees are evaluated by some judge (typically a supervisor) who assesses the employee's performance along a given set of dimensions, assigns a score to that assessment, and then usually informs the employee of his or her formal rating. Organizations typically base a variety of decisions concerning the employee partially on this rating.

Performance management refers to the wide variety of activities, policies, procedures, and interventions designed to help employees to improve their performance. These programs begin with performance appraisals but also include feedback, goal setting, and training, as well as reward systems. Therefore, performance management systems begin with performance appraisal as a jumping-off point, and then focus on improving individual performance in a way that is consistent with strategic goals and with the ultimate goal of improving firm performance (cf. Aguinis & Pierce, 2008). Performance management is a relatively recent term, however, and throughout the 100-year history reviewed here, the vast majority of articles are concerned with the type of performance appraisal more commonly done in organizations.

Our review will focus on the issues associated with the two processes and how they have developed over the years. We will pay special attention to the role that the *Journal of Applied Psychology* (*JAP*) has played in this history, noting when articles in the journal have made real contributions and when the journal has had less impact. Space limitations make it impossible to describe all of the important articles that contributed to the body of research in performance appraisal and performance management, but the supplemental material for this issue of *JAP* includes more information on articles that had the greatest impact in each area (regardless of where they were published) as well as a bibliography of relevant articles in *JAP* over the last 100 years.

Performance Appraisal Research

Although interest in the evaluation of performance at work dates back to ancient China, and although there were efforts at establishing merit ratings in various settings as far back as the 19th century (Murphy & Cleveland, 1995), psychological research on performance rating did not begin in a serious way until the 1920s. This body of research can be traced back to Thorndike's classic article, "A Constant Error in Psychological Ratings" (Thorndike, 1920). He identified what eventually became known as "halo error." The assumption at the time was that errors of this sort would reduce the accuracy of ratings and therefore make them less useful. This article, along with articles by Rugg (1921) and Remmers (1931), which argued that graphic rating scales were especially prone to this error, helped drive appraisal research for at least the next 50 years.

This article was published Online First January 26, 2017.

Angelo S. DeNisi, A. B. Freeman School of Business, Tulane University; Kevin R. Murphy, Department of Personnel and Employment Relations, University of Limerick.

Correspondence concerning this article should be addressed to Angelo S. DeNisi, A.B Freeman School of Business, Tulane University, 7 McAlister Drive, New Orleans, LA 70118. E-mail: adenisi@tulane.edu

To help characterize the development of performance appraisal research, particularly research published in *JAP*, we sorted performance appraisal articles into eight substantive categories: (a) scale format research, (b) research evaluating ratings—articles examining various criteria for evaluating performance ratings (e.g., rater errors), (c) training research, (d) research on reactions to appraisals, (e) research on purpose of rating, (f) research on rating source, (g) research on cognitive processes—studies of information processing and judgment in rating, and (h) research on demographic effects.

Historical Review of Performance Appraisal Research in JAP

Performance appraisal has been the major focus of a large number of articles published in *JAP*. There were 94 articles dealing primarily with job performance measurement or performance appraisal published in *JAP* prior to 1970, several of which represent important contributions to the literature (e.g., Bingham, 1939; Hollander, 1957; Smith & Kendall, 1963), although most of them deal with rating scale formats and ways to reduce rating "errors." There were also 30 articles dealing with performance appraisal since 2000, but it seems clear that the period 1970 to 2000 represented the heyday of performance appraisal research in *JAP*. During this period, *JAP* published 187 articles in which performance appraisal was the primary topic (we should also note that, during this period, there were also a number, but simply reported on a new rating scale that was developed for some job or reported data on some other type of performance measure [e.g., Ferguson, 1947; Rothe & Nye, 1958], and this is the main focus of our review).

We analyzed the content of these 187 articles to identify trends in the topics covered during the period 1970–2000. We grouped articles into 5-year spans and examined the content of *JAP* articles published 1970–1974 (23 articles), 1975–1979 (29 articles), 1980–1984 (34 articles), 1985–1989 (43 articles), 1990–1994 (30 articles), and 1995–2000 (24 articles). Figure 1 illustrates the proportion in each content category for the six time periods studied. We conducted a similar analysis for the earlier and later periods; those data can be found in the online supplemental materials associated with this article, as can a list of especially influential articles in each area we discuss (Archival Table 1).

There are three particularly noteworthy trends in Figure 1. First, scale format research was very popular in *JAP*, particularly from 1970–1979, representing well over 40% of the performance appraisal articles published in this period (and, as noted in the previous paragraph, this was the case for the earlier period as well). In hindsight, this emphasis might be seen as regrettable, since Landy and Farr's (1980) review of performance appraisal research lamented the unproductive nature of these scale format studies, and called for a moratorium on studies of scale formats, and, in fact, our data suggest this was the case, but as we shall discuss later, this research did help move the field forward.

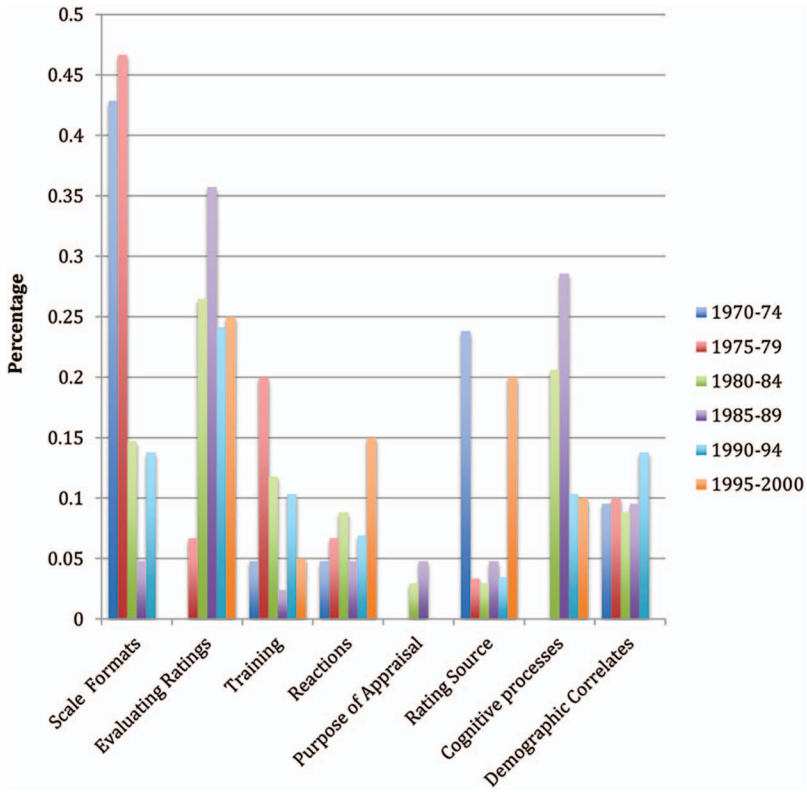

Figure 1. Trends in performance appraisal research, 1970–2000. See the online article for the color version of this figure.

Second, starting in the 1980s, studies of the criteria used for evaluating ratings (i.e., criteria for criteria, such as rater errors, interrater agreement, rating accuracy) started to appear with some regularity; *JAP* published 26 articles on these topics between 1980 and 1997. Articles published in *JAP* during this period featured particularly lively discussions of the nature and consequences of halo error, and of the relationship between halo and other supposed errors in rating. Ultimately, this focus was also shown as somewhat misguided, but it was these articles that helped us understand why reducing rating errors was not the optimum criterion measure for evaluating appraisal systems

Third, there was a dramatic spike in studies on cognitive processes in evaluating performance in the 1980s, primarily because of two influential review articles (Feldman, 1981; Landy & Farr, 1980) calling researchers' attention to the importance of these processes. DeNisi (2006) and Ilgen, Barnes-Farrell, and McKellin (1993) provide excellent reviews of cognitive research on performance appraisal, much of which was published in *JAP*. Starting in 1990, there was a substantial drop in the popularity of studies of cognitive processes in performance appraisal, reflecting, in part, growing concern about the relevance of this work to the practice of performance appraisal in organizations (Ilgen et al., 1993). Nevertheless, this body of research made clear contributions to our understanding of how performance is judged (DeNisi, 2006; Ilgen et al., 1993) and provided useful suggesting for improving performance appraisal in organizations (e.g., behavior diaries).

Fourth, training research showed a brief spike in the late 1970s, but never really emerged as a dominant area of research, and although there were other areas that showed intermittent peaks and valleys, there were no clear trends until recent interest in contextual performance.

The Development of Key Concepts in Performance Appraisal

Figure 1 illustrates some broad trends in performance appraisal research, but it does not fully capture how ideas developed in various parts of this field. A more detailed discussion of the issues dealt with by particular articles provides a more vivid picture of the concepts that came to dominate performance appraisal research.

Rating Scales

As noted above, Thorndike's (1920) and Rugg's (1921) article help set the stage for over 50 years of performance appraisal research. In the 1920s and 1930s, research was primarily concerned with ways to improve graphic rating scales or ranking methods (e.g., Ream, 1921). It is worth noting, however, that although a number of articles dealing with these two types of ratings appeared in *JAP*, the original work introducing the scales appeared elsewhere. Graphic rating scales were introduced by Paterson (1922), who described the scales developed by the Scott Company. Ranking methods were first used operationally by the U.S. Army in World War I, but the original work dates back at least a far as 1906 (Cattell, 1906), and the use of paired comparisons for establishing ranks dates back almost as far (e.g., Barrett, 1914).

Knauft (1948) introduced the idea of using weighted checklists to rate performance. This research proposed a marked improvement over simple lists of behaviors, and other studies seeking further improvements followed (e.g., Meyer, 1951). But the most important development in the area of checklists came with the introduction of the critical incidents technique, introduced by Flanagan in an article appearing in *Personnel Psychology*, and later fully explicated in *Psychological Bulletin* (Flanagan, 1954). Throughout the 1950s and 1960s, there were several articles appearing in *JAP* endorsing this type of scale and/or suggesting improvements (e.g., Kay, 1959). Furthermore, the critical incidents technique also led to the development of behavioral observation scales, which were first introduced by Latham and Wexley (1977) in an article in *Personnel Psychology*. Subsequently, several articles appeared in *JAP* pointing out problems with the use of these scales (e.g., Murphy & Constans, 1987; Murphy, Martin, & Garcia, 1982), and they never gained widespread popularity.

The history and review of forced choice rating scales appeared in *Psychological Bulletin* (Travers, 1951), but the rating method gained popularity with a article by Sisson (1948) and a later article by Berkshire and Highland (1953). Later articles appearing in *JAP* proposed improvements in this technique (e.g., Lepkowski, 1963; Obradovic, 1970).

Smith and Kendall's (1963) classic article adopted methods introduced by Champney (1941) to develop scales for evaluating job performance, but were also based broadly upon critical incidents; these scales are commonly referred to as behaviorally anchored rating scales. There were a variety of rationales for these scales, but it was generally thought that the use of behavioral anchors provided a clear definition and a consistent frame of reference for both the dimensions to be rated and the various levels of performance. Over the years, variations on this scaling format were introduced, including mixed standard scales. This rating format, introduced by Blanz and Ghiselli (1972) in *Personnel Psychology*, asked raters to determine whether the performance of a ratee was better than, worse than, or about the same as the performance level exemplified by a behavioral example, and Saal (1979) introduced scoring systems for potentially inconsistent responses to this type of scale. Figure 2 presents examples of these different rating scale formats.

By the time Landy and Farr's (1980) review of performance appraisal research was published in *Psychological Bulletin*, it was becoming clear that variations in scale formats had only modest effects on the quality of rating data. That review arguably signaled an end to what was the most dominant line of research on performance appraisal—the search for scale formats that would solve the problems of subjectivity, inaccuracy, and lack of credibility that have, for so long, contributed to concerns over performance appraisal in the field. It should be noted, however, that research on rating scale formats has continued, albeit sporadically (e.g., Scullen, Bergey, & Aiman-Smith, 2005; dealing with forced distribution ratings).

Attempts to improve rating scales represented a significant portion of the total body of research on performance appraisal and performance management published in *JAP* during its first 100 years, and from one perspective, these articles might be dismissed as unproductive. We do not agree. Although the search for a truly superior rating scale was not successful, this line of research pushed the field to more carefully anchor performance judgments in behavioral terms, which also improved the clarity and acceptability of performance ratings. Furthermore, given the attention to

scale development in so many other areas of psychology during this time, this was a vein that *had to* be mined. Psychologists may never have moved beyond the view that substandard scales were holding back performance appraisal without this sustained effort to develop better scales.

Evaluating the Quality of Rating Data

Much of the performance appraisal research appearing in *JAP* has been concerned with evaluating the quality of rating data—that is, assessing the reliability, validity, or accuracy of performance ratings. For example, Rugg (1921) used measures of reliability as the criteria for evaluating rating systems. In articles such as this, authors either implied or stated that they were interested in producing more "accurate" ratings, but accuracy was never directly measured. Instead, a system was "good" if it resulted in reliable ratings (e.g., Remmers, 1931), and even "better" if it produced equally reliable ratings with less time and effort (e.g., McCormick & Roberts, 1952).

During the period 1920–1970, the "holy trinity" of rater error measures was developed—that is, measures of halo, leniency/severity, and central tendency/range restriction, starting with the first performance appraisal article published in *JAP* (Thorndike, 1920), which was concerned with halo error. Guilford and Jorgensen's (1938) paper, "Some Constant Errors in Rating," discussed concepts highly similar to leniency and central tendency error, although they did not use these precise terms, and error measures were the most common method of evaluating the quality of rating data, through the 1980s. Alternatives, such as evaluating the convergent and discriminant validity of ratings (Kavanagh, MacKinney, & Wolins, 1971) and applications of factor analysis for evaluating halo (Landy, Vance, Barnes-Farrell, & Steele, 1980), were suggested, but these methods did not displace traditional rater error measures.

There were, however, a few articles that attempted to assess rating accuracy more directly. For example, Mullins and Force (1962) compared ratings of employee characteristics such as carelessness with scores on tests designed to measure carelessness, on the assumption that raters whose evaluations matched test scores were more accurate. Wiley and Jenkins (1964) compared the ratings of individual raters with the mean across many raters, and argued that individuals who were closest to the mean were more accurate. Borman (1977, 1979) developed a method of evaluating rating accuracy that involved comparisons between the ratings obtained from a single rater with the average of ratings obtained from multiple experts operating in optimal rating conditions. In practice, this method required strict control over the stimuli being rated, and typically involved having both subjects and experts rate videotaped performance segments, and numerous *JAP* articles during the 1980s applied these methods.

The development of practical methods of evaluating rating accuracy stimulated debates about precisely how accuracy should be defined. Most studies of rating accuracy relied on one or more of the accuracy indices articulated by Cronbach (1955), such as elevation (accuracy in the overall level of rating) and stereotype accuracy (accuracy in distinguishing dimensions on which ratees tend to perform well from those on which rates tend to perform poorly). Sulsky and Balzer's (1988) review suggested that different accuracy measures were not interchangeable, and also noted that the use of rating accuracy criteria limited appraisal research to laboratory studies and to evaluations of short performance segments.

In fact, both rater error and rating accuracy measures can be problematic, and this point had been made by several authors over the years. For example, Guilford and Jorgensen (1938) noted, error measures required knowledge of the true distributions and intercorrelations among performance dimensions. Therefore, assuming that halo error was present, for example, if the intercorrelations among ratings were too high, made no sense without data concerning the levels of true relationships among dimensions. Bingham's (1939) discussion concerning "true halo" made many of the same points. Furthermore, it was widely assumed that ratings that exhibited more reliability or fewer rater errors must also be more accurate. This assumption persisted for quite a long period of time, despite a number of articles arguing that this was not true (e.g., Buckner, 1959; Cooper, 1981). Murphy and Balzer's (1986) article, demonstrating that traditional rater error measures were uncorrelated with any of the four accuracy measures described by Cronbach (1955), essentially ended the field's reliance upon rater errors, although some recent research on the distribution of ratings suggests that there may be some benefit to less "lenient" ratings (e.g., O'Boyle & Aguinis, 2012).

Research in this area, then, contributed to the field primarily by pointing out the problems with more traditional criteria for evaluating the effectiveness of appraisal systems. Although subsequent research suggested that rating accuracy might not be the best criterion measure (e.g., Ilgen, 1993), this line of research led us away from traditional error measures to accuracy measures and eventually to measures that reflected ratee perceptions of fairness and accuracy.

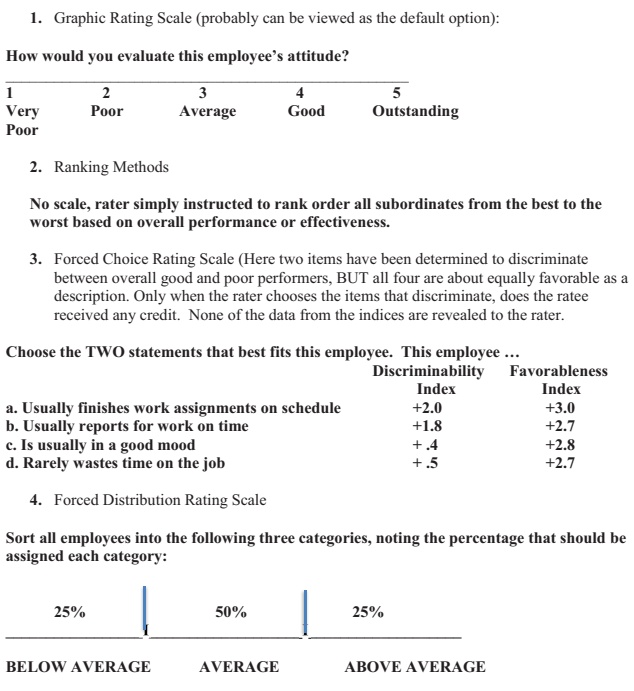

Figure 2. Examples of rating scale formats proposed to improve ratings. See the online article for the color version of this figure.

Training

Levine and Butler (1952) provided the first description of a rater training program in *JAP*. They use a variety of training methods, but one stands out as being prophetic—that is, they lectured raters on the nature of rater errors and cautioned them to avoid them. Similar approaches were applied by a number of other scholars over the years (e.g., Bernardin & Walter, 1977; Borman, 1975). This training "worked" in the sense that it led to lower levels of intercorrelation and lower mean ratings, which was taken to mean less halo and leniency error.

But drawing upon the research described above, training researchers began to realize that training raters to avoid rating errors did not necessarily lead to more accurate ratings and could produce *decreased* accuracy (Bernardin & Pence, 1980; Hedge & Kavanagh, 1988). This led researchers to modify training, for example, by adding detailed discussion of the behavior anchors on rating scales to discussions of rater errors (Ivancevich, 1979) or adding training on how to record behaviors accurately (Thornton & Zorich, 1980), which seemed to help.

Borman (1979) described a training approach that foreshadowed what would become the dominant method of rater training—that is, Frame of Reference (FOR) training. He showed raters videotapes of performance, asked them to rate the tapes, then discussed the most appropriate ratings for each candidate ("true score") and why this was appropriate. Bernardin and Buckley (1981) formally introduced the concept of FOR training, which Woehr (1994) described as including definitions of performance dimensions, samples of behavioral incidents representing each dimension, and an indication of the level of performance represented by each incident, followed by practice and feedback using these standards.

Several studies provide evidence for the effectiveness of FOR training for instilling a uniform theory of work performance in all raters (e.g., Athey & McIntyre, 1987; Gorman & Rentsch, 2009; Pulakos, 1984). However, as Uggerslev and Sulsky (2008) note, FOR training is most successful for raters whose theory of work performance prior to training is already similar to the theory FOR attempts to instill.

Two broad themes have dominated research on rater training: (a) how to train raters, and (b) how to determine whether or not training works. On the whole, progress has been more substantial in determining the content of rater training than in determining its effectiveness. There is consensus that training raters what *not* to do (e.g., training them to avoid rater errors) is ineffective. There is also consensus that training raters to adopt consistent conceptions of what represents good versus poor performance and what behaviors and competencies constitute performance is beneficial.

Reactions to Appraisals

There was little research on reactions to appraisals until the 1970s. The research that began at that time focused on ratee satisfaction and perceptions of fairness, for the most part (e.g., Wexley, Singh, & Yukl, 1973) Research on perceptions of fairness date back to two studies by Landy and his associates (Landy, Barnes, & Murphy, 1978; Landy, Barnes-Farrell, & Cleveland, 1980), which found that identifying goals to improve weaknesses, frequent evaluations, and rater knowledge were important predictors of perceptions of fairness and rating accuracy, while the later study also found that consistency among feedback sources was also important. A later meta-analysis (Cawley, Keeping, & Levy, 1998) reported that participation (in various forms) was highly correlated with employee reactions, and pointed to the importance of justice perceptions in this process. Several articles, appearing elsewhere, also focused on the role of justice perceptions in reactions to performance appraisal (e.g., Folger, Konovsky, & Cropanzano, 1992), which was confirmed in other empirical articles (e.g., M. S. Taylor, Tracy, Renard, Harrison, & Carroll, 1995).

This line of research has been especially important because, when combined with the research on rating errors described above, it helped move the field to consider other types of outcome measures that could be used to evaluate appraisals systems. In fact, justice perceptions have become an important part of later models of performance management (e.g., DeNisi & Smith, 2014), and this is suggested as an important area for research in the future.

Purpose for Appraisal

Meyer, Kay, and French (1965) noted that performance appraisals are often used for multiple reasons, and that these reasons can lead to conflicting goals for appraisal, a finding empirically confirmed by Cleveland, Murphy, and Williams (1989). Studies showed that the purpose for which ratings were to be used affected what kinds of information were sought as well as how that information was used to make decisions (e.g., Williams, DeNisi, Blencoe, & Cafferty, 1985; Zedeck & Cascio, 1982). Also, although organizations have their purpose for collecting ratings, raters also have purposes and goals in mind when making ratings, and Murphy, Cleveland, Skattebo, and Kinney (2004) found that the goal of the rater at the time of the evaluation affected several properties of the ratings given.

This research did not have a large impact on the practice of performance appraisal. The various purposes to which organizations apply appraisal information from Cleveland et al. (1989) continue, and it is extremely rare to find a case in which different appraisals are conducted for different purposes.

Rating Sources

Performance ratings are usually obtained from supervisors, but it has long been recognized that other sources might be tapped to obtain performance evaluations (e.g., subordinates, self-ratings). As early as the 1940s, Ferguson (1947) used peers, supervisors, and subordinates in *developing* appraisal scales, but he did not obtain performance ratings from all three sources. Authors of several other articles (e.g., Bendig, 1953; Fiske & Cox, 1960) discuss the use of self-ratings and/or peer ratings, but do not directly compare the information obtained from self- or peer ratings with supervisory ratings.

Springer (1953) is the first *JAP* article to specifically ask the question of whether "supervisory personnel and co-workers agree in their ratings" (p. 347). His findings (modest positive correlations between ratings of the same dimensions by peers and supervisors) set the pattern for virtually every subsequent study comparing ratings from different sources, but this did not discourage further research on agreement (see review by Heidemeier & Moser, 2009).

This research had its greatest impact in the role played in recommendations for multisource or 360-degree appraisals (i.e.,

rating systems in which feedback is collected from multiple sources and feedback pooled by source is provided to rates). Much of this work was based on research by Lawler (1967), published in *JAP*, including London and Smither's (1995) article suggesting that feedback from multiple sources could influence subsequent goals as well as performance. Smither, London, and Reilly (2005) also presented a theoretical model as well as meta-analysis results supporting multisource appraisals, but Greguras and Robie (1998) reported that none of the rating sources showed high levels of reliability, while Seifert, Yukl, and McDonald (2003) suggested that the use of a coach could improve the effectiveness of these programs.

Research on these programs progressed through a series of steps, first asking whether it was possible to obtain performance ratings from multiple sources, then asking whether ratings from these sources were comparable, and, finally, asking whether multisource feedback was effective. The results of his research suggest that obtaining information from different sources can be useful, but that different sources differ systematically in the conclusions they suggest about overall performance levels, and that evaluations from others are likely to be less favorable than self-evaluations. There remain questions about the effectiveness of these systems, however, and we must conclude that the jury is still out on the last issue (cf. DeNisi & Kluger, 2000).

Demographic Effects

The possibility that performance ratings might be affected by demographic variables such as race, gender, or age has long been recognized as a potential source of employment discrimination. In one of the first published studies of race differences in performance measures, Dejung and Kaplan (1962) examined ratings of combat aptitude as a function of the race of both raters and ratees. They found that Black raters favored Black ratees, but White raters did not favor White ratees, and this same pattern of results was reported in several studies and in a subsequent meta-analysis discussed later in this section.

Bass and Turner (1973) examined Black–White differences in both performance ratings (pooled over multiple raters for most rates) and objective performance measures (number of errors, number of miscounts, attendance) for bank tellers. They reported significant differences, favoring White tellers, but noted that these were generally small, and were smaller for ratings than for objective measures. This pattern of results has been replicated in most subsequent studies.

During the next several years, there were several notable indirect studies on demographic effects on performance evaluations. Hamner, Kim, Baird, and Bigoness (1974) had students observe the performance of confederates (White and Black, male and female) in a work sample task, and they reported small main effects for race and Rater X Ratee interactions. Schmitt and Lappin (1980), in a similar study using videotapes performance samples, replicated this pattern of Rater X Ratee race interactions, with Black raters giving especially high ratings to Black ratees.

Several studies have taken an even more indirect approach, asking subjects in laboratory studies to read vignettes and make judgments about hypothetical ratees. Schwab and Heneman (1978) showed small, but sometimes significant, differences in ratings as a function of rater and ratee age. Rosen and Jerdee published several vignette studies (e.g., Rosen & Jerdee, 1976) suggesting that there were potential sex and age biases in evaluations.

However, there are reasons to believe that the findings of these laboratory studies overestimate the effects of the demographic characteristics in the appraisal process. Wendelken and Inn (1981) argued that demographic differences are made especially salient in laboratory studies in which other ratee characteristics are tightly controlled, and in which raters are untrained and have no prior knowledge of, and no relationship with, ratees. Murphy, Herr, Lockhart, and Maguire's (1986) meta-analysis confirmed that vignette studies do indeed tend to produce larger effects than do studies involving the observation of actual performance.

The results of several large-scale studies provide better insight into the effects of the demographic characteristics of ratees on performance ratings. Pulakos, White, Oppler, and Borman (1989) reported some significant race effects, but also noted that effect sizes were generally quite small and did not always favor White ratees (see also Kraiger & Ford, 1985; Waldman & Avolio, 1991), whereas studies by Sackett, DuBois, and Noe (1991) and Sackett and DuBois (1991) suggested that the overall pattern of race effects might be more complex than had been suspected.

This body of research shows that in some settings (especially laboratory studies), demographic variables *can* influence performance ratings. However, it also suggests that in the field, these variables do not have a large effect on performance ratings. Variables like age, gender, or race influence a number of things that happen in organizations, but the outcomes of performance appraisals do not seem to be strongly influenced by these demographic variables.

Given the evidence or racism, sexism, and ageism in many other settings, the question of *why* performance ratings do *not* seem to show these same biases is an important one. It is possible that as information about performance is acquired over time, that information eventually swamps whatever stereotypes influence judgments made in less information-rich environments, but at present, this is simply a hypothesis awaiting convincing tests.

Cognitive Processes

Several reviews and theory articles (e.g., DeNisi, Cafferty, & Meglino, 1984: Feldman, 1981; Landy & Farr, 1980) sparked interest in the cognitive processes involved in performance rating, particularly in the way raters acquire, organize, recall, and integrate information about rate performance. Over the next 10 to 15 years, a number of studies, mostly laboratory experiments involving relatively short videotapes of performance, were published in *JAP*. These studies dealt primarily with the acquisition, organization, and encoding of performance information, and with the recall and integration of that information.

Various studies examined the role of rating purpose (Kinicki, Hom, Trost, & Wade, 1995; Williams, Cafferty, & DeNisi, 1990), and rater affect (Robbins & DeNisi, 1994), as well as individual differences (e.g., Bernardin, Cardy, & Carlyle, 1982), on cognitive processes, and DeNisi, Robbins, and Cafferty (1989) reported that keeping behavioral diaries could aid recall an encoding. Researchers also found that the organization of information in memory could be affected by categories existing in raters' minds prior to observing behavior (e.g., Kozlowski & Kirsch, 1987). Still others reported that rater schema for organizing information influenced

what was attended to and later recalled (e.g., Nathan & Lord, 1983), and therefore could influence the accuracy if what was actually observed and recalled (e.g., Lord, 1985). Interestingly, these findings tied back to the work on frame of reference training, suggesting that this training could lead raters to adopt a consistent and appropriate set of categories for organizing information, which can, in turn, enhance accuracy in recall and evaluation (e.g., Athey & McIntyre, 1987; Woehr, 1994).

But it should be noted that although DeNisi and Peters (1996) demonstrated similar effects for diary keeping and a structured recall task in one of the few field studies involving cognitive processes, Murphy, Garcia, Kerkar, Martin, and Balzer (1982) reported that accuracy in evaluating performance is mostly unrelated to accuracy in observing and encoding behavior. Furthermore, a number of studies suggested that memory for behaviors was strongly affected by general impressions and overall evaluations (e.g., Murphy & Balzer, 1986), and even the anchors used in behaviorally based rating scales (Murphy & Constans, 1987). More critically, Murphy, Martin, et al. (1982) reported that these effects could overwhelm the actual memory for specific behaviors exhibited by a ratee.

Studies of cognitive processes in performance evaluation made a definitive contribution by considering the ways in which raters obtain, process, and retrieve information about performance. On the other hand, this line of research had a relatively short shelf life, emerging as a major force in the 1980s, but tailing off dramatically by the mid-1990s. This shift was driven, in part, by the recognition that performance appraisal is not a simple judgment task, but rather a task that occurs in a complex and demanding environment and that involves motivational as well as cognitive variables (Murphy & Cleveland, 1995; Murphy & DeNisi, 2008).

Performance Management Research

Given our 100-year perspective, the history of research on performance management is much more limited than that for performance appraisal. Even the term "performance management" is much more recent, and so there is much less history to describe. However, as we shall discuss, there has certainly been a history of research on some of the important components of performance management, such as feedback and goal setting. This research has been focused on improving the performance of individuals, however, and the ultimate goal of performance management systems is to improve firm-level performance. But although it has often been assumed that improving the individual performance would ultimately improve firm-level performance as well, establishing meaningful links between changes in individual performance and changes in firm performance has been an elusive goal (cf. DeNisi & Smith, 2014).

We view the research on performance management as falling into three broad categories. First, there are the articles (or mostly books) that attempt to describe the entire performance management process and suggest how to improve it. Second are the articles that deal with specific aspects of the performance management process. In general, these articles are concerned with improving individual performance, and address one type of performance management intervention in isolation without discussing it in the wider perspective of the entire process. Finally, there are the articles that focus on improving firm-level performance and how Human Resources (HR) practices can influence performance at that level.

The Performance Management Process

Performance appraisal has usually been considered an important aspect of performance management, but it would seem that the best hope of establishing a link between individual performance improvement and firm performance improvement would be to consider performance appraisal (either in its formal, annual guise or in terms of more frequent, less formal assessments) as only one of a broader set of activities that entail aligning staffing, performance feedback, incentives, and supervision with the strategic goals of organizations. Rather than just referring to ways in which organizations use performance appraisal information to improve performance, then, performance management would be defined as this broader set of HR activities, as has been suggested in books (e.g., Aguinis, 2013; Pulakos, Mueller-Hanson, O'Leary, & Meyrowitz, 2012) as well as publications in other journals (e.g., DeNisi & Smith, 2014; Kinicki, Jacobson, Peterson, & Prussia, 2013). But the vast majority of this work describes models and techniques rather than actually testing the effectiveness of these programs.

Improving Individual Performance

There has been a considerable body of research examining ways to improve individual performance and productivity, and much of it has appeared in *JAP*. A good example of this is the research on ProMES, a system that combines feedback, goal setting, and incentives in attempt to improve productivity (e.g., Pritchard, Harrell, DiazGranados, & Guzman, 2008; Pritchard, Jones, Roth, Stuebing, & Ekeberg, 1988). Interestingly, as research began to focus on performance management, the underlying theoretical models switched from measurement-oriented models to motivational models (e.g., DeNisi & Pritchard, 2006). That is, rather than focusing on the accuracy of the ratings, the research began looking at what drove employees to try to improve their performance.

The topic of feedback meant to improve performance has also been featured in several articles from *JAP*, including an influential review of the literature on feedback from the late 1970s (Ilgen, Fisher, & Taylor, 1979) as well as some earlier work on feedback (e.g., Butler & Jaffee, 1974). Feedback research has continued to appear in the journal (e.g., Northcraft, Schmidt, & Ashford, 2011), including articles dealing with conceptual issues raised elsewhere in a major review by Kluger and DeNisi (1996; e.g., Vancouver & Tischner, 2004). Likewise, articles dealing with the role of incentives and goals on performance have been published in *JAP* over the years (e.g., Camman & Lawler, 1973; Ronan, Latham, & Kinne, 1973).

Improving Firm Performance

There is also a considerable body of research indicating how HR practices can influence firm performance when they are implemented as part of a larger scale effort (e.g., Huselid, 1995; Jackson, Schuler, & Jiang, 2014; Ployhart & Moliterno, 2011). Many of these have appeared in other journals, but several have been published in *JAP* (e.g., Aryee, Walumbwa, Seidu, & Otaye, 2012; Crook, Todd, Combs, Woehr, & Ketchen, 2011). What is far from

clear is *what* combinations of HR practices make a difference (and whether performance evaluation and feedback is part of whatever combinations work) and *why* some practices work and others do not. DeNisi and Smith's (2014) model of the performance management process suggests several useful avenues for research on these questions.

Contributions of Performance Appraisal and Performance Management Research

The history of research on performance appraisal and management is not isomorphic with the history of articles on these topics in the *JAP*. As we shall discuss, articles published in *JAP* have more influence in some areas of research than in others. Based on our review of this literature, however, two overall conclusions seem justified. First, the overall contribution of 100 years of research on performance appraisal has been much greater than it has relative to performance management. Second, there are specific areas in which the contribution of articles published in *JAP* has been substantial, but sometimes in quite indirect ways.

For example, we can consider the contribution of studies of rating scale formats. Although we believe that this research failed to demonstrate that the right rating format would make a big difference in the quality of performance ratings, it was necessary to actually conduct this research in a rigorous way in order to reach that conclusion. Furthermore, this work eventually led to an appreciation for the importance of ratee perceptions of fairness and accuracy, and *JAP* is where much of that research was published. This latter contribution was especially important for research on ratee reactions to appraisals (e.g., Taylor et al., 1995), as well as to subsequent models of performance management (e.g., DeNisi & Smith, 2014).

Studies of criteria for evaluating ratings provided a similar pattern of contribution. Two early articles (Bingham, 1939; Thorndike, 1920) stimulated a considerable body of research on halo error (most of which was published between 1970 and 2000), but much of this work proved less useful for evaluating ratings (given the continuing difficulty in sorting true from illusory halo, leniency, and the like) than for helping us understand the way judgments about performance are made. Also, as with research on rating scale formats, this research *had* to be carried out in a rigorous way, in order for us to better understand judgment processes in this area, and, again, the vast majority of that research was published in *JAP*.

Studies of the relationships between the demographic characteristics of raters and rates proved to be very important for understanding the effects of ratings on employees (e.g., Kraiger & Ford, 1985; Schmitt & Lappin, 1980), and, in our view, this stream of research represents one of the real success stories for research published in this journal. Both *JAP* studies and meta-analyses that include many *JAP* articles (e.g., Bowen, Swim, & Jacobs, 2000) suggest that although there are some differences in the performance ratings received by women versus men or by members of different racial and ethnic groups, these differences are usually quite small. These findings loom large in discussions of the construct validity and the fairness of performance ratings.

Research on rater training showed the familiar pattern of fits and starts, which characterize the development of important concepts in many fields of psychology. In particular, much of the early work in this area dealt with what were probably misguided efforts to encourage raters to avoid rater errors. However, subsequent research on rater training documented several effective methods for increasing the accuracy of ratings, particularly frame of reference training. Again, *JAP* has made notable contributions to our understanding of how and why this method of training works.

Finally, we believe that studies of the cognitive processes involved in evaluating others' performance made a modest but worthwhile contribution. This may reflect our bias (as active contributors to this line of research), but we believe this line of research advanced the science of evaluating judgment, while spurring very useful discussions of what can and what cannot be learned from laboratory studies that mimic some aspects of performance appraisal and omitted others (e.g., Murphy & Cleveland, 1995).

Despite the fact articles published in *JAP* have made many meaningful contributions to the literature, we believe it is fair to say *JAP* has not always been highly influential in the development of the main concepts, theories, and approaches that characterize current thinking in performance appraisal. For example, Murphy and Cleveland (1991, 1995) published two widely cited books reviewing performance appraisal and suggesting new directions for performance appraisal research. The broad theme of both books is the need to understand the context in which performance appraisal is carried out, how the motives of the participants (both raters and rates) shape appraisal processes and outcomes, and the implications of these two factors for evaluating the validity and usefulness of performance ratings. Although both books cited a number of articles published in *JAP*, it is fair to say that the research that had the strongest influence on the conceptual development of the models and approaches in these books appeared in outlets other than *JAP*. In fact, we can summarize the contributions of research published in *JAP* to the field of performance appraisal by stating that, with few exceptions, the journal was not the home of the theories or conceptual models that guided the field, but was, instead, where these models were tested and where critical technical issues were addressed.

Understanding the contribution of the *JAP* to the performance management process requires some context—there have been very few empirical articles appearing in *any* journal in our field which have actually tested performance management programs. The work of Pritchard and his associates with PRoMES (cf., Pritchard, Harreell, Diaz-Grandos, & Guzman, 2008), which seeks to improve productivity by applying ideas from several motivational theories, comes close, however, and much of the empirical work concerning ProMES was published in *JAP* (e.g., Pritchard, Jones, Roth, Stuebing, & Ekberg, 1988).

But we could identify no empirical articles in our field (published in *JAP* or elsewhere) that have demonstrated how improving individual level performance can be leveraged to improve firm-level performance. Thus, we are left with conceptual/theoretical work explaining how such systems could be developed, but this work has been confined primarily to books and book chapters by either academicians (e.g., Aguinis, 2013; DeNisi & Smith, 2014) or practitioners (e.g., Pulakos et al., 2012). As noted, though, there have been a number of articles appearing in *JAP* that have added to our understanding of how bundles of HR practices can help improve firm-level performance.

Finally, there have certainly been a large number of articles that have focused upon how various performance management techniques such as feedback, goal setting, or incentive pay can influence individual-level performance, and much of this work has appeared in *JAP*. Thus, it is fair to conclude that *JAP* has made contributions to many of the components of performance management, but there is no credible empirical work, in *JAP* or elsewhere, that allows us to determine how performance management might actually work.

What Have We Learned and What Do We Need to Do?

Our title included a question mark suggesting potential doubts about whether the substantial body of research published in the last 100 years in *JAP* has made a substantial contribution to our understanding of performance appraisal and performance management. The answer is both "yes" and "no." It should be clear that we have come a long way from examining rating scale formats to determine their effects on rating errors, and *JAP* has contributed substantially to this progress. We have certainly learned that the specific format of the rating scale used is not the most important consideration in developing appraisal systems and that traditional error measures are not the best way to evaluate such systems. We have learned that demographic characteristics may have less influence on ratings than we had believed, that some rater cognitive processes are related to appraisal decisions, and that it is possible to train rates to do a better job. Certainly, these accomplishments can be considered progress.

However, perhaps the most significant progress we have made during this time is to come to better appreciate the critical influence of the context in which performance appraisal occurs on the process and outcomes of appraisal (Murphy & DeNisi, 2008), and the role of *JAP* in this area is smaller and more indirect. Performance appraisal is used for a variety of purposes in organizations (Cleveland et al., 1988), and these purposes influence the way performance is defined (e.g., task performance vs. contextual performance; Podsakoff, Ahearne, & MacKenzie, 1997) and the way raters and ratees approach the task of performance appraisal (Murphy & Cleveland, 1995). The appraisal effectiveness model proposed by Levy and Williams (2004) summarizes much of the research on the role of social context and emphasizes the importance of rate reactions to appraisals and the acceptability of ratings, and some of the work summarized in this review has appeared in *JAP*. However, most of the research published in *JAP* has been decontextualized, examining different facets of the rating process (e.g., cognitive processes, rating scales, rater training) in isolation, and it has become clear that we will not make progress in understanding how or why appraisals succeed without considering why appraisals are done in the first place, and how the climate, culture, norms, and beliefs in organizations shape the appraisal process and the outcomes of appraisals.

Contextualizing performance appraisal research implies paying attention to when and why performance appraisal is carried out and the contextual variables that are likely to be important range from quite distal (e.g., national cultures) to quite proximal (e.g., supervisor-subordinate relationships). For example, there may be aspects of national culture (or organizational culture) that make it less acceptable to give anyone negative feedback, and this may put pressure on raters to intentionally inflate ratings. In fact, we know little about how culture and societal norms really affect appraisal decisions and processes; *JAP* has made few contributions here. There is descriptive research that indicates that different practices and policies are more likely in some parts of the world than in others (e.g., Varma, Budhwar, & DeNisi, 2008), but we do not fully understand how cultural norms may make certain practices more or less effective. Also, we need more research on the effectiveness of individual-level performance management techniques in different cultures. The archive for this issue also includes a model of various factors that might affect performance appraisal processes and changes in individual performance. This model is adapted from Murphy and DeNisi (2008).

At the most fundamental level, the question mark in our title really refers to the uncertainty in moving from the level of individual-level performance to firm-level performance. DeNisi and Smith (2014) concluded that although we have learned a great deal about how to improve individual performance through appraisal and performance management programs, there is no evidence to show that improving individual-level performance will eventually lead to improvements in firm-level performance. As noted earlier, it has always been implied or assumed that improving individual-level performance would eventually lead to improvements in firm-level performance. The ongoing failure to establish a clear link between individual and performance that leads us to raise questions about overall progress in this field. Even if we succeed in using performance appraisal, feedback, and other components of performance management to improve individual job performance, it is not clear that this will lead to more effective organizations. We believe that identifying how (if at all) the quality and the nature of performance appraisal programs contribute to the health and success of organizations is a critical priority. *JAP* is not alone in its failure to address this priority; the research literature in the organizational sciences simply has not grappled with this question in a credible way.

In conclusion, we believe that *JAP* has made some worthwhile contributions to our understanding of performance appraisal and performance management. More important, *JAP* can and should have a critical role in the future progress of our field. *JAP* has always placed a strong emphasis on rigorous empirical test of theories and models, and this is an orientation that is not universally shared across journals in this domain. As a consequence, we believe that *JAP* should be a natural home for rigorous tests of performance management programs and their components. It is disconcerting to see how much discussion of performance management exists, and how little evidence there is about how it actually works. It is our hope that *JAP* can take a lead in combining the concern for the organizational performance, often shown in other parts of the organizational sciences, with its traditional concern for scientific rigor to produce a better understanding of how and why performance appraisal and performance management actually function in organizations, and how attempts to evaluate and improve individual performance influence the lives of employers and employees and the organizations in which they are found.

Specifically, we believe that *JAP* should seek to publish research that (a) is conducted in organizations settings, (b) involves processes and outcomes with real stakes for participants (for example, studies of performance appraisals that are used for pro-

motion and pay decisions), (c) includes assessments of both distal and proximal context variables, and (d) includes assessments of performance/success at a range of levels, including individual, group, and firm performance measures when possible. Research on cognitive processes focused on how raters form judgments, but as Murphy and Cleveland (1995) point out, there is a difference between judgments and actual ratings, and future research also needs to focus on the reasons why raters might not choose to provide ratings consistent with their judgments. All of this will require something that *JAP* once routinely did, but that is now challenging and rare, carrying out research in organizations or in cooperation with practitioners. For decades, we have bemoaned the gap between research and practice. It is time to stop the moaning and start the process of rebuilding the essential links between our research as psychologists and the topic we claim to care about—understanding behavior in organizations.

References

Aguinis, H. (2013). *Performance management* (3rd ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Aguinis, H., & Pierce, C. A. (2008). Enhancing the relevance of organizational behavior by embracing performance management research. *Journal of Organizational Behavior, 29,* 139–145. http://dx.doi.org/10.1002/job.493

Aryee, S., Walumbwa, F. O., Seidu, E. Y., & Otaye, L. E. (2012). Impact of high-performance work systems on individual- and branch-level performance: Test of a multilevel model of intermediate linkages. *Journal of Applied Psychology, 97,* 287–300. http://dx.doi.org/10.1037/a0025739

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72,* 567–572. http://dx.doi.org/10.1037/0021-9010.72.4.567

Barrett, M. A. (1914). A comparison of the order of merit method and the method of paired comparisons. *Psychological Review, 21,* 278–294. http://dx.doi.org/10.1037/h0075829

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology, 57,* 101–109. http://dx.doi.org/10.1037/h0037125

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. *Journal of Applied Psychology, 37,* 38–41. http://dx.doi.org/10.1037/h0057911

Berkshire, J. R., & Highland, R. W. (1953). Forced-choice performance rating: A methodological study. *Personnel Psychology, 6,* 355–378. http://dx.doi.org/10.1111/j.1744-6570.1953.tb01503.x

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? *Journal of Applied Psychology, 66,* 151–160.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60–66. http://dx.doi.org/10.1037/0021-9010.65.1.60

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of Behaviorally Anchored Ratings Scales (BARS). *Journal of Applied Psychology, 66,* 458–463. http://dx.doi.org/10.1037/0021-9010.66.4.458

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62,* 64–69. http://dx.doi.org/10.1037/0021-9010.62.1.64

Bingham, W. V. (1939). Halo, invalid and valid. *Journal of Applied Psychology, 23,* 221–228. http://dx.doi.org/10.1037/h0060918

Blanz, F., & Ghiselli, E. E. (1972). The Mixed Standard Rating Scale: A new rating system. *Personnel Psychology, 25,* 185–199. http://dx.doi.org/10.1111/j.1744-6570.1972.tb01098.x

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* 556–560. http://dx.doi.org/10.1037/0021-9010.60.5.556

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior & Human Performance, 20,* 238–252. http://dx.doi.org/10.1016/0030-5073(77)90004-6

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410–421. http://dx.doi.org/10.1037/0021-9010.64.4.410

Bowen, C., Swim, J. K., & Jacobs, R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology, 30,* 2194–2215.

Buckner, D. N. (1959). The predictability of ratings as a function of interrater agreement. *Journal of Applied Psychology, 43,* 60–64. http://dx.doi.org/10.1037/h0047296

Butler, R. P., & Jaffee, C. L. (1974). Effects of incentive, feedback and manner of presenting the feedback on leader behavior. *Journal of Applied Psychology, 59,* 332–336. http://dx.doi.org/10.1037/h0036655

Camman, C., & Lawler, E. E. (1973). Employee reactions to a pay incentive plan. *Journal of Applied Psychology, 58,* 163–172.

Cattell, J. M. (1906). *American men of science: A biographical dictionary*. New York, NY: Science Press.

Cawley, B. D., Keeping, L. M., & Levy, P. E. (1998). Participation in the performance appraisal process and employee reactions: A meta-analytic review of field investigations. *Journal of Applied Psychology, 83,* 615–633. http://dx.doi.org/10.1037/0021-9010.83.4.615

Champney, H. (1941). The measurement of parent behavior. *Child Development, 12,* 131–166.

Cleveland, J. N., Murphy, K. R., & Williams, R. (1988). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74,* 130–135. http://dx.doi.org/10.1037/0021-9010.74.1.130

Cooper, W. H. (1981). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology, 66,* 302–307. http://dx.doi.org/10.1037/0021-9010.66.3.302

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin, 52,* 177–193. http://dx.doi.org/10.1037/h0044919

Crook, T. R., Todd, S. Y., Combs, J. G., Woehr, D. J., & Ketchen, D. J., Jr. (2011). Does human capital matter? A meta-analysis of the relationship between human capital and firm performance. *Journal of Applied Psychology, 96,* 443–456. http://dx.doi.org/10.1037/a0022147

Dejung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and rate on early peer ratings of combat aptitude. *Journal of Applied Psychology, 46,* 370–374. http://dx.doi.org/10.1037/h0048376

DeNisi, A. S. (2006). *A cognitive approach to performance appraisal*. New York, NY: Routledge.

DeNisi, A. S., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior & Human Performance, 33,* 360–396. http://dx.doi.org/10.1016/0030-5073(84)90029-1

DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degreee appraisals be improved? *The Academy of Management Executive, 14,* 129–139.

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field.

Journal of Applied Psychology, 81, 717–737. http://dx.doi.org/10.1037/0021-9010.81.6.717

DeNisi, A. S., & Pritchard, R. D. (2006). Improving individual performance: A motivational framework. *Management and Organization Review, 2,* 253–277. http://dx.doi.org/10.1111/j.1740-8784.2006.00042.x

DeNisi, A. S., Robbins, T., & Cafferty, T. P. (1989). Organization of information used for performance appraisals: Role of diary-keeping. *Journal of Applied Psychology, 74,* 124–129. http://dx.doi.org/10.1037/0021-9010.74.1.124

DeNisi, A. S., & Smith, C. E. (2014). Performance appraisal, performance management, and firm-level performance: A review, a proposed model, and new directions for future research. *The Academy of Management Annals, 8,* 127–179. http://dx.doi.org/10.1080/19416520.2014.873178

Dickinson, Z. C. (1937). Validity and independent criteria. *Journal of Applied Psychology, 21,* 522–527. http://dx.doi.org/10.1037/h0062095

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127–148. http://dx.doi.org/10.1037/0021-9010.66.2.127

Ferguson, L. W. (1947). The development of a method of appraisal for assistant managers. *Journal of Applied Psychology, 31,* 306–311. http://dx.doi.org/10.1037/h0057937

Fiske, D. W., & Cox, J. A. (1960). The consistency of ratings by peers. *Journal of Applied Psychology, 44,* 11–17. http://dx.doi.org/10.1037/h0046278

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327–358. http://dx.doi.org/10.1037/h0061470

Folger, R., Konovsky, M. A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. *Research in Organizational Behavior, 14,* 127–148.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94,* 1336–1344. http://dx.doi.org/10.1037/a0016476

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83,* 960–968. http://dx.doi.org/10.1037/0021-9010.83.6.960

Guilford, J. P., & Jorgensen, A. P. (1938). Some constant errors in ratings. *Journal of Experimental Psychology, 22,* 43–57. http://dx.doi.org/10.1037/h0061118

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59,* 705–711. http://dx.doi.org/10.1037/h0037503

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73,* 68–73. http://dx.doi.org/10.1037/0021-9010.73.1.68

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 94,* 353–370. http://dx.doi.org/10.1037/0021-9010.94.2.353

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41,* 85–90.

Hollander, E. P. (1965). Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology, 49,* 434–438. http://dx.doi.org/10.1037/h0022805

Huselid, M. A. (1995). The impact of human resource practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 38,* 635–672. http://dx.doi.org/10.2307/256741

Ilgen, D. R. (1993). Performance appraisal accuracy: An illusive and sometimes misguided goal. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Industrial and organizational perspectives* (pp. 235–252). Hillsdale, NJ: Erlbaum.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321–368. http://dx.doi.org/10.1006/obhd.1993.1015

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64,* 349–371. http://dx.doi.org/10.1037/0021-9010.64.4.349

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64,* 502–508. http://dx.doi.org/10.1037/0021-9010.64.5.502

Jackson, S. E., Schuler, R. S., & Jiang, K. (2014). Strategic human resource management: A review and aspirational framework. *Academy of Management Annals, 8,* 1–56.

Kavanagh, M., MacKinney, A., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin, 75,* 34–49. http://dx.doi.org/10.1037/h0030412

Kay, B. R. (1959). The use of critical incidents in a forced-choice scale. *Journal of Applied Psychology, 43,* 269–270. http://dx.doi.org/10.1037/h0045921

Kinicki, A. J., Hom, P. W., Trost, M. R., & Wade, K. J. (1995). Effects of category prototypes on performance-rating accuracy. *Journal of Applied Psychology, 80,* 354–370. http://dx.doi.org/10.1037/0021-9010.80.3.354

Kinicki, A. J., Jacobson, K. J., Peterson, S. J., & Prussia, G. E. (2013). Development and validation of a performance management behavior questionnaire. *Personnel Psychology, 66,* 1–45. http://dx.doi.org/10.1111/peps.12013

Kluger, A. N., & DeNisi, A. S. (1996). The effects of feedback interventions on performance: A historical review, meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254

Knauft, E. B. (1948). Construction and use of weighted check list rating scales for two industrial situations. *Journal of Applied Psychology, 32,* 63–70. http://dx.doi.org/10.1037/h0060388

Kozlowski, S. W., & Kirsch, M. P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology, 72,* 252–261. http://dx.doi.org/10.1037/0021-9010.72.2.252

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70,* 56–65. http://dx.doi.org/10.1037/0021-9010.70.1.56

Landy, F. J., Barnes, J. L., & Murphy, K. R. (1978). Correlates of perceived fairness and accuracy of performance evaluations. *Journal of Applied Psychology, 63,* 751–754. http://dx.doi.org/10.1037/0021-9010.63.6.751

Landy, F. J., Barnes-Farrell, J., & Cleveland, J. N. (1980). Perceived fairness and accuracy of performance evaluation: A follow-up. *Journal of Applied Psychology, 65,* 355–356. http://dx.doi.org/10.1037/0021-9010.65.3.355

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107. http://dx.doi.org/10.1037/0033-2909.87.1.72

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology, 65,* 501–506. http://dx.doi.org/10.1037/0021-9010.65.5.501

Latham, G. P., & Wexley, K. N. (1977). Behavior observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255–268. http://dx.doi.org/10.1111/j.1744-6570.1977.tb02092.x

Lawler, E. E. (1967). The multitrait-multirater approach to studying managerial job performance. *Journal of Applied Psychology, 51,* 369–381. http://dx.doi.org/10.1037/h0025096

Lepkowski, J. R. (1963). Development of a forced-choice rating scale for engineer evaluations. *Journal of Applied Psychology, 47,* 87–88. http://dx.doi.org/10.1037/h0044908

Levine, J., & Butler, J. (1952). Lecture vs. group decision in changing behavior. *Journal of Applied Psychology, 36,* 29–33. http://dx.doi.org/10.1037/h0053624

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. http://dx.doi.org/10.1016/j.jm.2004.06.005

London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology, 48,* 803–839. http://dx.doi.org/10.1111/j.1744-6570.1995.tb01782.x

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71. http://dx.doi.org/10.1037/0021-9010.70.1.66

McCormick, E. J., & Roberts, W. K. (1952). Paired comparison ratings: II. The reliability of ratings based on partial pairings. *Journal of Applied Psychology, 36,* 188–192. http://dx.doi.org/10.1037/h0055956

Meyer, H. H. (1951). Methods for scoring a check-list type of rating scale. *Journal of Applied Psychology, 35,* 46–49. http://dx.doi.org/10.1037/h0055890

Meyer, H. H., Kay, E., & French, J. (1965). Split roles in performance appraisal. *Harvard Business Review, 43,* 123–129.

Mullins, C. J., & Force, R. C. (1962). Rater accuracy as a generalized ability. *Journal of Applied Psychology, 46,* 191–193. http://dx.doi.org/10.1037/h0044264

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67,* 161–164. http://dx.doi.org/10.1037/0021-9010.67.2.161

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71,* 39–44. http://dx.doi.org/10.1037/0021-9010.71.1.39

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74,* 619–624. http://dx.doi.org/10.1037/0021-9010.74.4.619

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

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

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89,* 158–164. http://dx.doi.org/10.1037/0021-9010.89.1.158

Murphy, K. R., & Constans, J. I. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology, 72,* 573–577. http://dx.doi.org/10.1037/0021-9010.72.4.573

Murphy, K. R., & DeNisi, A. (2008). A model of the performance appraisal process. In A. Varma, P. Budhwar, & A. DeNisi (Eds.), *Performance management systems around the globe* (pp. 81–96). London, UK: Routledge.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. (1982). The relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67,* 320–325. http://dx.doi.org/10.1037/0021-9010.67.3.320

Murphy, K. R., Herr, B. M., Lockhart, M. C., & Maguire, E. (1986). Evaluating the performance of paper people. *Journal of Applied Psychology, 71,* 654–661. http://dx.doi.org/10.1037/0021-9010.71.4.654

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67,* 562–567. http://dx.doi.org/10.1037/0021-9010.67.5.562

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68,* 102–114. http://dx.doi.org/10.1037/0021-9010.68.1.102

Northcraft, G. B., Schmidt, A. M., & Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. *Journal of Applied Psychology, 96,* 1076–1086. http://dx.doi.org/10.1037/a0023221

O'Boyle, E., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65,* 79–119. http://dx.doi.org/10.1111/j.1744-6570.2011.01239.x

Obradovic, J. (1970). Modification of the forced choice method as a criterion for job proficiency. *Journal of Applied Psychology, 54,* 228–233. http://dx.doi.org/10.1037/h0029264

Paterson, D. G. (1922). The Scott Company Graphic Rating Scale. *Journal of Personnel Research, 1,* 361–376.

Ployhart, R. E., & Moliterno, T. P. (2011). Emergence of the human capital resource: A multilevel model. *Academy of Management Review, 36,* 127–150.

Podsakoff, P. M., Ahearne, M., & MacKenzie, S. B. (1997). Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of Applied Psychology, 82,* 262–270. http://dx.doi.org/10.1037/0021-9010.82.2.262

Pritchard, R. D., Harrell, M. M., DiazGranados, D., & Guzman, M. J. (2008). The productivity measurement and enhancement system: A meta-analysis. *Journal of Applied Psychology, 93,* 540–567. http://dx.doi.org/10.1037/0021-9010.93.3.540

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., & Ekeberg, S. E. (1988). Effects of group feedback, goal setting, and incentives on organizational productivity. *Journal of Applied Psychology, 73,* 337–358. http://dx.doi.org/10.1037/0021-9010.73.2.337

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581–588. http://dx.doi.org/10.1037/0021-9010.69.4.581

Pulakos, E. D., Mueller-Hanson, R. A., O'Leary, R. S., & Meyrowitz, M. M. (2012). *Building a high-performance culture: A fresh look at performance management*. SHRM Foundation Effective Practice Guidelines Series. Alexandria, VA: SHRM Foundation.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74,* 770–780. http://dx.doi.org/10.1037/0021-9010.74.5.770

Ream, M. J. (1921). A statistical method for incomplete ordering of merit ratings. *Journal of Applied Psychology, 5,* 261–266. http://dx.doi.org/10.1037/h0066886

Remmers, H. H. (1931). Reliability and halo effect of high school and college students' judgments of their teachers. *Journal of Applied Psychology, 18,* 619–630. http://dx.doi.org/10.1037/h0074783

Robbins, T. L., & DeNisi, A. S. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology, 79,* 341–353. http://dx.doi.org/10.1037/0021-9010.79.3.341

Ronan, W. W., Latham, G. P., & Kinne, S. B. (1973). Effects of goal setting and supervision on worker behavior in an industrial situation. *Journal of Applied Psychology, 58,* 302–307. http://dx.doi.org/10.1037/h0036303

Rosen, B., & Jerdee, T. H. (1976). The nature of job-related age stereotypes. *Journal of Applied Psychology, 61,* 180–183. http://dx.doi.org/10.1037/0021-9010.61.2.180

Rothe, H. C., & Nye, C. T. (1958). Output rates among coil winders. *Journal of Applied Psychology, 42,* 182–186. http://dx.doi.org/10.1037/h0041121

Rugg, H. (1921). Is the rating of human character practicable? *Journal of Educational Psychology, 12,* 425–438, 485–501.

Saal, F. E. (1979). Mixed Standard Rating Scale: A consistent system for numerically coding inconsistent response combinations. *Journal of Applied Psychology, 64,* 422–428. http://dx.doi.org/10.1037/0021-9010.64.4.422

Sackett, P. R., & DuBois, C. L. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76,* 873–877. Retrieved from http://psycnet.apa.org/journals/apl/76/6/873/

Sackett, P. R., DuBois, C. L., & Noe, A. W. (1991). Tokenism in performance evaluation: The effects of work group representation on male-female and White-Black differences in performance ratings. *Journal of Applied Psychology, 76,* 263–267. Retrieved from http://psycnet.apa.org/journals/apl/76/2/263/

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435. http://dx.doi.org/10.1037/0021-9010.65.4.428

Schwab, D. P., & Heneman, H. G. (1978). Age stereotyping in performance appraisal. *Journal of Applied Psychology, 63,* 573–578. http://dx.doi.org/10.1037/0021-9010.63.5.573

Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential. *Personnel Psychology, 58,* 1–32. http://dx.doi.org/10.1111/j.1744-6570.2005.00361.x

Seifert, C. F., Yukl, G., & McDonald, R. A. (2003). Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. *Journal of Applied Psychology, 88,* 561–569. http://dx.doi.org/10.1037/0021-9010.88.3.561

Sisson, E. D. (1948). Forced choice: The new rating. *Personnel Psychology, 1,* 365–381. http://dx.doi.org/10.1111/j.1744-6570.1948.tb01316.x

Smith, P. C., & Kendall, L. M. (1963). Retranlsation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155. http://dx.doi.org/10.1037/h0047060

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58,* 33–66.

Springer, D. (1953). Ratings of candidates for promotion by co-workers and supervisors. *Journal of Applied Psychology, 37,* 347–351.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497–506.

Taylor, M. S., Tracy, K. B., Renard, M. B., Harrison, J. K., & Carroll, S. J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly, 40,* 495–523. http://dx.doi.org/10.2307/2393795

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29. http://dx.doi.org/10.1037/h0071663

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65,* 351–354. http://dx.doi.org/10.1037/0021-9010.65.3.351

Travers, R. M. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin, 48,* 62–70. http://dx.doi.org/10.1037/h0055263

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93,* 711–719. http://dx.doi.org/10.1037/0021-9010.93.3.711

Vancouver, J. B., & Tischner, E. C. (2004). The effect of feedback sign on task performance depends on self-concept discrepancies. *Journal of Applied Psychology, 89,* 1092–1098. http://dx.doi.org/10.1037/0021-9010.89.6.1092

Varma, A., Budhwar, P., & DeNisi, A. (2008). *Performance management systems around the globe*. London, UK: Routledge.

Waldman, D. A., & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76,* 897–901. http://dx.doi.org/10.1037/0021-9010.76.6.897

Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology, 66,* 149–158. http://dx.doi.org/10.1037/0021-9010.66.2.149

Wexley, K. N., Singh, J. P., & Yukl, G. (1973). Subordinate personality as a moderator of the effects of participation in three types of appraisal interview. *Journal of Applied Psychology, 58,* 54–59. http://dx.doi.org/10.1037/h0035411

Wiley, L., & Jenkins, W. S. (1964). Selecting competent raters. *Journal of Applied Psychology, 48,* 215–217. http://dx.doi.org/10.1037/h0045601

Williams, K. J., Cafferty, T. P., & DeNisi, A. S. (1990). Effects of performance appraisal salience on recall and ratings. *Organizational Behavior and Human Decision Processes, 46,* 217–239.

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The effects of appraisal purpose on information acquisition and utilization. *Organizational Behavior and Human Decision Processes, 36,* 314–339. http://dx.doi.org/10.1016/0749-5978(85)90027-5

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79,* 525–534. http://dx.doi.org/10.1037/0021-9010.79.4.525

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67,* 752–758. http://dx.doi.org/10.1037/0021-9010.67.6.752

Received May 20, 2015
Revision received December 11, 2015
Accepted December 17, 2015 ■