The Effects of Rater Performance and Perspective on Rating Leniency

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Kyle David McNeal

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nathan R. Kuncel, Adviser

August 2019

Acknowledgements

I am immensely grateful to the many people who have helped me to get to this point. I am very fortunate to have benefited from the guidance, encouragement, and support of so many friends, family members, and colleagues throughout this journey. This acknowledgements section seems fully inadequate to express my gratitude. That said, I'll do my best.

First, a thank you to the University of Minnesota and its incredible faculty members. I have had the opportunity and privilege to learn from the very best. In particular, the faculty of the IO program have been a source of inspiration through their deep expertise, commitment to the science, and the impact that their efforts have made on our field and the world of work. I have a deep sense of pride in the legacy and character of our program. A special thanks to my advisor Nathan Kuncel, as well as my committee members Paul Sackett, Aaron Schmidt, and Lou Quast, whose guidance and support are very much appreciated.

I would also like to thank my friends and colleagues in the IO program. Despite the long road and occasional challenges of the graduate experience, my lasting impression of this time in my life will be an overwhelmingly positive one, due in no small part to the camaraderie within the program and the lasting relationships that were built in the process. Amanda Kopydlowski and Melissa Sharpe in particular cannot be thanked enough for their friendship and support as we went through every step of this experience together.

Finally, any accomplishments that have been achieved to date are more a reflection of the love and encouragement I have received from my incredible family and partner than anything I can take credit for on my own. Thank you for everything.

Abstract

Rater errors such as leniency/severity have detrimental effects on the validity of performance ratings. A number of rater characteristics have been examined to understand why some raters tend to be consistently more lenient than others; however, gaps remain in our understanding of these rater characteristics and their influence on rating leniency. The present study examined the previously unexplored characteristic of *rater performance* as a predictor of rater leniency/severity. It was hypothesized that rater performance would be negatively associated with rater leniency, such that high performing employees would be more severe in their evaluations of others. Furthermore, it was hypothesized that this relationship would be particularly pronounced when raters and subjects are peers to one another, and when rater and subject are of the same gender. These hypotheses were tested using a large archival data set including multi-source assessment ratings and annual performance ratings for employees in a multinational healthcare organization. The hypotheses were not supported, and in fact a small positive relationship between rater self-ratings of performance and rater leniency was detected. Implications and future directions are discussed.

Table of Contents

List of Tables

List of Figures

Introduction

Ratings of job performance are intended to reflect the proficiency with which employees perform behaviors that contribute toward organizational goals (Campbell, 2012). However, over a century of research on the topic has demonstrated that performance ratings are influenced by a number of non-performance-related factors. Of these confounding influences, the role of idiosyncratic rater effects has been a topic of particularly extensive study (e.g., Conway & Huffcutt, 1997; Holzbach, 1978; Hoyt & Kerns, 1999; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998). In the process of assigning ratings to a target employee, raters often (either consciously or unconsciously) assign ratings that reflect their own patterns of rating behavior rather than the actual performance of the target employee, or are influenced by extraneous factors immaterial to the employee's performance. Rater errors take many forms, including illusory halo, leniency/severity, and others. The presence of these errors has damaging effects on the validity and usefulness of performance ratings in both academic and applied settings. As a result, addressing the impact of rater-related error variance in ratings of job performance is of critical concern.

Many attempts have been made to mitigate the negative consequences of idiosyncratic rater error. Numerous interventions, including rater training (Roch, Woehr, Mishra, & Kieszcynska, 2012), scale formatting (Landy & Farr, 1980), statistical correction (Lance & Woehr, 1986; Raymond, Harik, & Clauser, 2011), and modifications to the appraisal context (Levy & Williams, 2004) have been introduced with varying degrees of success. However, rater error continues to represent an alarming proportion of variance in performance ratings (Hoyt & Kerns, 1999; Mount et al., 1998). Although

great strides have been made toward understanding the rating process, it is the opinion of this author that efforts to address issues of rater bias have been impeded by an incomplete understanding of the psychological and social causes of rater error. While some solutions (e.g., frame of reference training) have been moderately successful despite this lack of understanding, further progress will require a more comprehensive awareness of the human factors which induce rater error.

The present study seeks to contribute in this endeavor. First, current knowledge regarding the prevalence of various forms of rater error in performance ratings will be presented. This will be followed by an in-depth focus on a particular form of rater error often labeled *the leniency effect.* The current body of research on the nature, identification, impact, and causes of leniency will be reviewed. This study will then introduce a heretofore unexplored characteristic that may influence a rater's tendency toward leniency or severity: the rater's own level of performance. The theoretical and empirical rationale behind a hypothesized rater performance-rater leniency relationship will be articulated. Then, using a large database of multi-source developmental performance ratings, the presence and magnitude of this effect will be examined, and relevant moderators (e.g., the relationship of the rater to the ratee, rater-ratee gender congruence) will be assessed.

Forms of Idiosyncratic Rater Error

Idiosyncratic rater effects refer to systematic variance in ratings which are attributable to the rater, as opposed to the actual performance of the ratee being assessed (Scullen, Mount, & Goff, 2000). These rater effects can take a variety of forms. One category of idiosyncratic rater effects are systematic rater biases. These include illusory

halo, distributional errors (e.g., restriction of range, central tendency, leniency), and other perceptual errors (e.g., similar-to-me, first impression, systematic distortion). In the following section, each of these is briefly defined and discussed.

Illusory Halo

Halo error is the most extensively researched of the rater biases (Saal, Downey, & Lahey, 1980). Although the concept was initially introduced over a century ago by Wells (1907), Thorndike (1920) was the first to use the term *halo* to refer to a rater's inability to differentiate between conceptually distinct dimensions of a target employee's performance. Halo error manifests as inflated correlations between ratings of different variables made by the same rater—relationships which are not wholly attributable to true correlations between the measured performance dimensions ("true halo") (Hoyt, 2000). Halo error occurs when a rater's general impression of the ratee contaminates ratings across all dimensions, attributable to a number of factors including insufficient observation or recall of the ratee's behavior, highly salient features of the ratee which engulf other observations, inadequate rating instruments, or lack of rater effort (Cooper, 1981). Halo error has been demonstrated to account for a sizeable portion of variance in job performance ratings, and results in the inflation of correlations between rated dimensions (Viswesvaran, Schmidt, & Ones, 2005). Halo has been widely noted as a cause for concern in the interpretation of performance ratings, and a large body of research has sought to prevent or correct for this measurement artifact (Landy & Farr, 1980; Hoyt, 2000).

Distributional Errors

In addition to the problem of inflated covariance among performance dimensions, rater bias also occurs when raters fail to correctly represent the performance distribution of employees being rated. The inappropriate representation of the distribution of individual performance has important consequences on the accuracy and usefulness of performance ratings, potentially preventing the meaningful interpretation and analysis of performance data. The distribution of performance ratings can be adversely impacted by rater bias in two ways: by affecting the variance in ratings (range restriction and central tendency) and by affecting the mean ratings (leniency / severity) (Wildman, Bedwell, Salas, & Smith-Jentsch, 2010).

Range Restriction and Central Tendency. Range restriction is a distributional error in which raters assign ratings which do not adequately discriminate between target employees' performance levels. This occurs when raters assign similar ratings to all employees, failing to utilize the full spectrum of the rating scale (Saal et al., 1980). Kingsbury (1922) was among the first to address this phenomenon, commenting on how certain raters appeared hesitant to make distinctions between employees. Range restriction is often operationalized by the standard deviation of the ratings, with a smaller standard deviation in scores reflecting greater range restriction (e.g., Borman & Dunnette, 1975). A particular case of range restriction, central tendency, refers to raters' reluctance to assign ratings at either extreme of the rating scale, clustering all ratings around the midpoint of the scale (Saal et al., 1980). This adversely affects the rating process by masking performance differences between target employees and attenuating relationships between performance and other variables of importance, such as scores on selection assessments.

Leniency and Severity. In contrast to the phenomena of range restriction and central tendency, which affect the variance in performance ratings, leniency and severity are forms of rater error which primarily impact the location or mean of the distribution (Wildman et al., 2010). Raters provide ratings that are, on average, higher (leniency) or lower (severity) than warranted by the employees' actual behaviors. Although conceptually straightforward, researchers have operationalized leniency and severity by a variety of methods with very different and highly consequential assumptions (Saal et al., 1980). Leniency and severity are of central importance to the present study, and as such will be discussed in greater detail later in this document.

Other Errors of Rater Perception

In additional to the frequently discussed halo and distributional errors present in performance ratings, a handful of other rater errors have been the subject of more limited investigation (Borman, 1991). The first of these, *first impression error,* occurs when an observer allows judgments made after an initial meeting with a subject to contaminate ratings of that subject's performance (Latham, Wexley, & Pursell, 1975). This can lead to inappropriately high or low ratings of a subject that are not warranted on the basis of that subject's actual behaviors, and can also undermine the construct validity of a performance measure. *Similar-to-me error* refers to a situation in which a rater identifies with certain characteristics of a subject and, as a result, assigns that subject higher ratings (Latham et al., 1975). For example, raters have been shown to assign higher ratings to targets who share common biographical histories with the rater (Rand & Wexley, 1975). A third rater error, closely related to halo, occurs when covariation among rated dimensions reflects semantic or conceptual similarities between dimension labels rather

than the true relationship between the measured constructs. According to the *systematic distortion hypothesis,* this phenomenon is particularly prevalent in scenarios in which raters have limited information about the performance criteria or the subjects' behavior (Kozlowski & Kirsch, 1987).

Leniency and Severity

The central concern of the present research relates to the rater errors of leniency and severity. In the following section, early research on these errors will be reviewed and alternative methods of defining and operationalizing the phenomena will be presented.

Early Research

The presence of leniency and severity in performance ratings has been acknowledged and discussed in organizational research for nearly a century. Kingsbury (1922) warned of overly lenient "high markers" and inappropriately severe "low markers," advising that each assessor's ratings be juxtaposed against organizational averages and the normal distribution of performance to assess the need for intervention. The term *leniency* was first used by Kneeland (1929) to describe raters who assigned erroneously high ratings, while Ford (1931) was the first to coin the term *severity* (Saal et al., 1980). Ford warned that the inequality of ratings provided by different assessors was of critical concern to the accurate interpretation of personnel data for administrative decision making, advocating for the use of "correction factors" to adjust for leniency and severity (p. 466). An important development pertinent to the current research study occurred when Guilford (1954) first explicitly argued that the tendency of assessors to over-rate or under-rate employees relative to the espoused standards of performance was a stable characteristic of individuals rather than merely a function of rating scales and

appraisal context, and that this tendency could be related to individual differences among raters. These researchers' early works prompted a century of investigation into the definition, measurement, causes, impact, prevention, and correction of leniency and severity in performance ratings, a stream of research which continues to this day.

Definition and Operationalization

Various terminology has been used to refer to the phenomena of leniency and severity. Collectively, leniency and severity have been referred to as rating-level biases (Bernardin, Thomason, Buckley, & Kane, 2016) and elevation accuracy (Cronbach, 1955). The terms rating elevation (Bernardin, Cooke, & Villanova, 2000) and rating inflation (Bernardin, Tyler, & Villanova, 2009) have been used synonymously with "leniency"; stringency (Harari, Rudolph, & Laginess, 2015; Viswesvaran et al., 2005) has also been used as an alternative to "severity." Many of these terms are associated with distinct methods of operationalizing the phenomena of leniency and severity (e.g., elevation accuracy vs. leniency error; Murphy & Balzer, 1989). However, the central premise of each remains the same: raters systematically assign ratings that are higher or lower depending, in part, on characteristics of the rater him or herself or the context in which rating occurs.

While the core notion of leniency and severity errors in performance rating is fairly straightforward, authors have adopted a variety of diverging definitions and methods of operationalizing these phenomena in their research. Discrepancies in how these concepts are defined and operationalized reveal important nuances in how the phenomena of leniency and severity are understood. In the following section,

disagreements in definition and operationalization will be reviewed, and the inherent assumptions and implications of each will be discussed.

In defining leniency and severity, authors often differ in whether leniency and severity are presented as an attribute of the ratings or as an attribute of the rater him or herself (e.g., Guilford, 1954; Sharon & Bartlett, 1969). While this distinction is, on the surface, fairly inconsequential—there are no ratings if there are no raters, and vice a versa—these differences have important ramifications. In defining leniency and severity as a characteristic of the rater, authors make two important assumptions. First, this assumes that raters systematically differ in the degree to which their rating level exceeds or falls below what is warranted. That is, raters have a tendency to be lenient or severe that is consistent across subjects and dimensions. Moreover, asserting that leniency and severity are attributes of the rater implies some degree of stability to these rating patterns— that raters' tendencies toward lenient or severe ratings are consistent over time (Kane, Bernardin, Villanova, & Peyrefitte, 1995). The question of whether leniency can be understood as a systematic and stable pattern of rater behavior, or instead, simply as a property of a particular set of ratings with a defined context, is an important issue to be discussed in a later section of this document.

Additionally, research on leniency and severity has employed a variety of operationalizations to identify these biases in ratings (Saal et al., 1980). These operationalizations each carry their own assumptions regarding the nature and definition of the performance ratings and rater error. One of the more common approaches to defining leniency has been to compare the average mean ratings assigned by a rater to the midpoint of the rating scale (e.g., Bernardin, Alvares, & Cranny, 1976; Taylor &

Hastman, 1956). While simple and intuitively justifiable, this approach tacitly assumes that all samples of employees are, on average, average performers. Given the considerable resources many organizations devote to recruiting, selecting, developing, and retaining high-performing employees, this would be a disappointing reality. Most performance rating scales use objective standards rather than relative comparisons as the basis of their rating anchors (Landy & Farr, 1980); as such, the expectation that the mean rating should precisely coincide with the midpoint of the rating scale is, in most cases, unfounded.

A second, less common method by which leniency has been operationalized is to assess the degree of skewness within a distribution of ratings (Saal et al., 1980). Using this method, researchers examine the extent to which a set of performance ratings deviate from the normal distribution. A negatively skewed distribution is argued to denote leniency, while a positively skewed distribution reflects rating severity (e.g., Landy, Farr, Saal, & Freytag, 1976). While this method has its merits, using skewness as an indicator of leniency requires two noteworthy assumptions. First, in order to make comparisons between raters in their leniency or severity, one must assume that ratees are randomly assigned to the raters—that there are no systematic differences in performance between the employees being rated by each rater, and that the distribution of performance within each set of ratees is equivalent. In the case of ratings assigned to direct reports by their managers, this assumption is clearly violated; research clearly indicates that managers have influence over the performance of their direct reports (e.g., Dulebohn, Bommer, Liden, Brouer, & Ferris, 2012; Wang, Oh, Courtright, & Colbert, 2011), and as such, the performance ratings of employees with highly effective (or ineffective) managers would

be expected to be more negatively (or positively) skewed than performance ratings of average managers. Secondly, this method of identifying leniency or severity requires that the true distribution of performance conforms to the normal distribution. While strong evidence indicates that the distribution of performance is approximately normal (Beck, Beatty, & Sackett, 2014), a number of factors may contribute to a non-normal distribution of performance within a given sample of employees. As a result, skewed but valid distributions of performance ratings may be erroneously characterized as lenient or severe under this operationalization.

The gold standard in identifying the presence of leniency or severity is to compare the rater's ratings to the employee's true performance. This method of measuring leniency is typically labeled "elevation accuracy" in the literature (Cronbach, 1955; Murphy & Balzer, 1989). Average ratings which exceed the employee's true performance indicate leniency, whereas average ratings below the true performance level denotes rater severity. Of course, accurately specifying an employee's true performance is a significant challenge, and is in fact an unattainable goal—despite the prevalent use of the term "true score" in this body of research, a more accurate label would be *true score estimate* (Sulsky & Balzer, 1988). Researchers have employed a variety of methods to estimate an employee's true performance by which to compare the ratings of a particular rater or set of raters. For the purposes of research, some authors have used objective performance metrics and compared these scores to the ratings provided by raters; for example, Farh and Werbel (1986) compared students' self-ratings of classroom participation with actual participation frequency to identify leniency or severity in self-ratings. While potentially useful for research, this method has little to no practical utility— not all performance

dimensions are amenable to objective measurement, and, if valid objective measures of individual performance are readily available, this essentially eliminates the need for subjective ratings in the first place. A more common approach to computing estimates of an employees' true performance is the method introduced by Borman (1977). This procedure, summarized briefly, involves averaging the ratings of numerous subject matter experts who evaluate the target employees under optimal conditions for observation, note-taking, and recall. The mean expert rating is then used as a true performance score against which raters' scores are compared (e.g., Jawahar & Stone, 1997). Again, however, this method has been subject to numerous criticisms, including its reliance on laboratory settings and its relevance to only certain types of performance behaviors observed in a constrained time frame (e.g., Heneman, Moore, & Wexley, 1987; Latham, 1986; Sulsky & Balzar, 1988). A third option for estimating true score performance (with greater application to field research) is to take the average rating of multiple raters who each evaluate a group of ratees on a consistent set of performance dimensions. This simple idea addresses numerous pitfalls of the other methods, and yet, as of 1995, had rarely been adopted in leniency research (Murphy & Cleveland, 1995). However, with the growing prevalence of multi-source performance ratings providing ample opportunity for analyses of this variety, the mean of multiple raters' scores for a given target employee has been frequently used as a true score performance estimate in recent years (e.g., Antonioni & Park, 2001; Kane et al., 1995).

Impact of Rater Error in Performance Ratings

There is general agreement in the field that our measures of job performance are deeply flawed. This conclusion is not new—researchers have wrestled with the task of

accurately defining and measuring the construct for over a century (Austin & Villanova, 1992). In fact, many authors have expressed doubt whether the process of performance appraisal is worth the effort—and some have argued that the practice of assigning ratings to employee performance should be abandoned entirely (Coen & Jenkins, 2000; Murphy & Cleveland, 1995).

What's wrong with ratings of job performance? Many empirical studies have shown, using a variety of methods, the shortcomings of job performance ratings in research and practice. First, job performance ratings are notoriously unreliable. Viswesvaran, Ones, and Schmidt (1996) demonstrated in a meta-analysis that the interrater reliability of supervisor ratings of overall job performance is .52. The findings were even more bleak for peer ratings of job performance, which Viswesvaran et al. reported was .42. Although this unreliability can be (and often is) corrected for in academic research, the implications for practice are stark. Individual performance ratings—the basis of compensation and other talent-related decision-making in the vast majority of organizations (Mercer, 2013)—are largely unreliable. Although some have argued that interrater reliability is not an appropriate index of the reliability of performance ratings (Murphy & Deshon, 2000; c.f., Schmidt, Viswesvaran, & Ones, 2000), other evidence also points to the questionable integrity of performance ratings. Bommer, Johnson, Rich, Podsakoff, and Mackenzie (1995) demonstrated meta-analytically that subjective ratings of performance correlate only .39 with objective measures of performance. While this is in part a reflection of the fallibility of objective measures of performance, this also speaks to the imprecision with which evaluations are able to appraise individual performance. In addition to the presence of measurement error

in performance ratings, other systematic non-performance-related factors have been shown to explain considerable amounts of variability in job performance ratings (Adler et al., 2016; Murphy, 2008; Smither, 2012). Of these systematic factors introducing error into ratings of job performance, rater effects and rater biases are among the most extensively studied and most severe (DeNisi & Murphy, 2017).

Much of the variance in ratings of job performance has more to do with the rater than the person being evaluated. Using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), Hoyt and Kerns (1999) were able to estimate the proportion of variance in observer ratings that is attributable to rater main effects and rater-target interactions. This meta-analysis of 79 studies found that, on overage, 37% of variance in observer ratings was attributable to either rater main effects or rater-target interactions. Moderator analysis revealed the proportion of variance attributable to raters was even greater when the ratings required raters to make inferences about a target, as is typical in ratings of performance. In these situations, nearly 50% of variance was attributable to rater effects. Other studies have examined the role of idiosyncratic rater effects specifically in the context of job performance ratings. One such study examined a large sample of multi-source ratings of manager performance using a multi-trait multi-method approach (Mount et al., 1998). These authors found that 72% of reliable variance in performance ratings was attributable to idiosyncratic rater effects. A more recent study conducted by O'Neill, McLarnon, and Carswell (2015) confirmed this finding in a large sample of multi-source ratings of manager performance, estimating that between 50% and 71% of variance in performance ratings were attributable to idiosyncratic rater error, compared to only 29% of variance which was actually attributable to the ratee. O'Neill et

al. were able to go one step further by estimating the relative impact of rater main effects and rater-ratee interactions, finding that each contributes about equally to the variance explained in ratings of managerial performance.

Research indicates that leniency-severity accounts for much of this systematic rater error (Borman & Hallam, 1991; Kane et al., 1995). Kane et al. (1995) analyzed manager ratings in three samples (police sergeants, patrol officers, and nurses) and concluded that over one quarter of the variance in performance ratings is attributable to leniency-severity. Nearly identical results were recently obtained by Dewberry, Davies-Muir, and Newell (2013) on a sample of medical doctors. The consistent finding that leniency-severity effects constitute a substantial source of systematic error has important ramifications for the interpretation and use of performance ratings. Taken as a whole, these results reveal that idiosyncratic rater effects constitute a substantial proportion of the reliable variance in ratings of job performance, even more so than the variance attributable to the actual target whose performance is supposed to be reflected in the performance ratings, and a rater's tendency to rate leniently or severely relative to other raters is a key contributor to this systematic error found in ratings of job performance.

The immediate and obvious implication of these findings is that performance ratings are severely impacted by error introduced by raters. Performance ratings, intended to reflect the behaviors of the target employee being assessed, are actually more a reflection of the individual rater or raters who are evaluating the employee. This has profound effects on the utility of these performance ratings in organizational decision-making. Rater leniency-severity obscures the true performance of the target employee in performance ratings. As a result, decisions made about individual employees based on

performance ratings (e.g., performance-based compensation decisions) are made on the basis of seriously flawed measures. When performance ratings assigned by different raters are used to compare multiple employees—for example, when making decisions about whom to promote—the differential rater leniency and severity changes the rank order or employees, and thus leads to inaccurate human capital decisions on the basis of performance ratings.

Causes and Correlates of Leniency Effect

The finding that rater leniency-severity has a substantial negative impact on performance ratings has led many researchers to devote considerable attention to understanding the underlying causes of rater leniency-severity. Researchers in the performance appraisal literature have identified a diverse set of factors which might contribute to rater error, including leniency-severity. These factors include characteristics of the measurement tool (e.g., the rating scale and instrument design), contextual factors (e.g., purpose of appraisal, social context), characteristics of the rater, and interactions between characteristics of the rater and ratee. In the following section, each of these will be discussed in turn, with particular attention paid to the literature on rater characteristics which are most pertinent to the present investigation.

Rating Instrument. Prior to the 1980s, the primary focus of performance appraisal research was on the impact of the measurement tool on the psychometric quality of the derived ratings. The underlying rationale was that different scale formats guide the rater to use different judgment processes in evaluating employee behavior, and as a result, identifying the optimal rating scale should improve the quality of the resulting performance ratings (Murphy & Constans, 1988). A number of rating instrument

characteristics were examined. For example, a substantial amount of energy was invested in determining the advantages and disadvantages of various types of scale anchors, including graphic rating scales and behaviorally anchored rating scales (e.g., Borman & Vallon, 1974; Burnaska & Hollman, 1974). Likewise, significant attention was devoted to determining whether forced choice methods of evaluation or mixed standard scales were superior to direct rating practices (e.g., Bernardin & Orban, 1990; Blanz & Ghizelli, 1972; Lepkowski, 1963; Taylor & Wherry, 1951). Yet another scale characteristic examined includes the number of response categories with which raters evaluate performance (e.g., Bernardin, La-Shells, Smith, & Alvares, 1976; Cicchetti, Shoinralter, & Tyrer, 1985; Lissitz & Green, 1975). This literature generated a handful of useful insights about the effect of rating instrument characteristics on leniency-severity. For example, behaviorally anchored rating scales were found to result in less leniency in ratings than graphic rating scales (Campbell, Dunnette, Arvey, & Hellervik, 1973; Sharon & Bartlett, 1969), and forced choice rating processes were found to result in less leniency bias than direct rating processes (Sharon & Bartlett, 1969; Taylor, Schneider, & Clay, 1954). Ultimately, however, this stream of research ended with pessimism about its value in improving the quality of ratings. Authors like Cozan (1959) and Borman and Dunnette (1975) noted that the slight advantages provided by methods such as behaviorally anchored rating scales or forced choice rating formats did not warrant the considerable investment of time and resources that were required to implement these methods. Landy and Farr (1980), in their summary of this literature, concluded that the effort to improve performance rating through scale optimization had not been particularly fruitful, and argued for a moratorium on these investigations. Although researchers continue to

occasionally revisit questions of rating scale format in the context of modern research on performance appraisal (e.g., Härtel, 1993), the literature has for the most part moved on to examine other opportunities to enhance the quality of performance ratings.

Appraisal Context. Following Landy and Farr's moratorium, the field shifted to a focus on the cognitive processes underlying rater evaluations of performance. A central focus of this literature was on the motivations of raters in a performance appraisal context, which had previously not been given much consideration in the research focused on measurement instruments and scale features. Several cognitive models were put forth which explicitly accounted for the impact of contextual factors in influencing rater motivations which, in turn, influenced the ratings themselves (e.g., DeNisi, Cafferty, & Meglino, 1984; Ilgen & Feldman, 1983). The role of contextual factors in the performance appraisal process has endured to this day, even after the heyday of cognitive theories of performance rating had ended. Central to this literature is the purpose of appraisal. In 1951, Taylor and Wherry hypothesized that ratings assigned for the purpose of administrative decision-making (e.g., compensation, promotion) would exhibit greater leniency than ratings used for non-administrative purposes (e.g., employee development, research). This notion was novel in that the rater was construed as an agentic contributor to the leniency bias, consciously inflating or not inflating ratings depending on the rater's goals and the expected consequences of the rating, both for the ratee as well as the rater him or herself. When ratings are used for administrative purposes, raters may be motivated to assign lenient ratings for a number or reasons. The rater may be resistant to assigning low ratings out of concern for negatively impacting an employee's compensation or career advancement (Murphy & Cleveland, 1991). Raters may also

assign inflated ratings because their direct reports' ratings could influence how the rater's performance is evaluated, in turn influencing the rater's compensation or career advancement. In appraisal contexts, raters may assume that other raters will be lenient, and so consciously inflate ratings so as to be consistent with the assumed behavior of others (Bernardin & Orban, 1990). Raters may be especially lenient in appraisal contexts in order to prevent negative reactions from the ratees (Fisher, 1989). Harris (1994) presents a thorough discussion of the role of rater motivation in performance appraisal contexts. These concerns and others led to a number of studies on the impact of purpose for appraisal on leniency. In a meta-analytic summary of 22 studies, Jawahar and Williams (1997) found that ratings in appraisal contexts were, on average, 1/3 of a standard deviation larger than ratings that were assigned in a development-only or research context.

A number of other contextual influences on leniency-severity were also investigated (Tziner, Murphy, & Cleveland, 2005). Murphy and Cleveland (1991) suggested that rater's attitudes toward an organization's appraisal process will influence their rating behavior. Tziner, Murphy, Cleveland, Beaudin, and Marchand (1998) tested this hypothesis and found that raters who believed that their ratings were meaningful and would impact important decisions demonstrated less leniency than raters who believed the performance appraisal was inconsequential. Tziner et al. also found that raters in organizational climates characterized by high work performance (Tziner & Dolan, 1984) were also less lenient in their ratings. Yun, Donahue, Dudley, and McFarland (2005) found that raters are more lenient when they know they will have to provide ratees with their performance feedback face-to-face. Benardin, Thomason, Buckley, and Kane (2016)

compared the leniency observed in contexts with high rater accountability (operationalized as when ratings are attributable to the rater and must be discussed between the rater and ratee) versus low rater accountability contexts. Bernardin et al. found that, in low accountability contexts, rater agreeableness was related to rating level bias; however, in high accountability contexts, this relationship was not observed. Appraisal accountability was also related to leniency-severity in a self-appraisal context (Farh & Werbel, 1986). These findings speak to the considerable impact that the performance appraisal context has on rating behavior and the quality of derived ratings. This body of research is ongoing, as researchers continue to investigate the social mechanisms affecting performance appraisal processes; see Levy and Williams (2004) and Pichler (2012) for a more thorough review of this broader literature.

Personality. Rater characteristics have been a central focus of research on the causes and correlates of rater leniency-severity, and rater personality has been a particular area of concentration. Landy and Farr's (1980) model of job performance ratings positioned personality one of the main factors influencing rater appraisal strategies, which in turn influence the leniency or severity of performance ratings. Since then, hundreds of studies have examined the impact of rater traits on performance ratings. Unfortunately, much of the early literature in this area was disjointed and fragmented; a broad range of traits had been examined with little coherent structure or framework organizing the investigations, including traits such as rater self-esteem (Wexley & Youtz, 1985), need for achievement (Kovacs & Kapel, 1976), and hostility (Phillips, 1960). However, as researchers increasingly organized their investigations around the five factor

model of personality, a more coherent picture of the relationship between rater personality and leniency-severity began to emerge.

Initially, published findings on the personality-leniency relationship demonstrated conflicting findings. Kane et al. (1995) were among the first to explicitly articulate the hypothesis that agreeableness should be positively associated with leniency, reflecting agreeable raters' desire to maintain positive relationships with the assessed employees, and conscientiousness should be negatively associated with leniency, as a result of conscientious raters' desire to be dependable and accurate in their ratings and tendency to have high standards of performance. These authors have published a number of studies demonstrating that rater agreeableness is a positive correlate of leniency and rater conscientiousness is as a negative predictor (e.g., Bernardin, Tyler, & Villanova, 2009; Bernardin, Cooke, & Villanova, 2000). However, other studies have found very small to no relationships (e.g., Dewberry et al., 2013; Yun et al., 2005). Although less often discussed as a predictor of leniency, extroversion has also been hypothesized to be positively associated with leniency, given that extroverts are more likely to have favorable relationships with colleagues (John, Naumann, & Soto, 2008), and relationship quality is positively associated with performance ratings (Duarte, Goodson, & Klich, 1994). Results have varied as to whether this relationship is borne out in the data (e.g., Bernardin et al., 2009; Bono, Hooper, & Yoon, 2012). The opposite rationale has been used to hypothesize a positive relationship between emotional stability and performance ratings—employees with low emotional stability (high neuroticism) are likely to have less positive relationships with colleagues, negatively affecting ratings assigned to those colleagues (Duarte et al., 1994). Openness to experience is generally not hypothesized to

relate to performance ratings, although there have been demonstrated associations with rater openness and ratings of transformational leadership (Bono et al., 2012).

In 2015, Harari et al. (2015) published a meta-analysis of 21 studies that examined the relationship between the five factors of personality and performance ratings. The authors noted that the existing literature demonstrated a number of conflicting findings, and argued that meta-analysis could help to provide clarity around the relationships between personality and leniency. Their results revealed a moderate relationship between agreeableness and performance ratings (ρ=.25) and smaller relationships for both extraversion and emotional stability (ρ=.12 for both). Notably, despite the prevalent assertion that conscientiousness should theoretically be negatively related to performance ratings, the direct relationship between conscientiousness and performance ratings was actually positive (ρ=.10), and relative weights analysis demonstrated that conscientiousness did not impact performance ratings after accounting for the other personality dimensions.

Cognitive Ability. Relative to the substantial body of research on the relationship between rater personality and leniency-severity, there has been limited investigation of rater cognitive ability. Furthermore, the limited research in this domain has not led to consistent findings. For example, both Hauenstein and Alexander (1991) and Bartels and Doverspike (1997) found that rater cognitive ability was related to rater leniency-severity; however, the observed relationships were in opposite directions. Hauenstein and Alexander found a non-linear relationship between intelligence and elevation accuracy, with moderately intelligent raters demonstrating the greatest elevation accuracy, while more intelligent raters were overly severe and less intelligent raters were overly lenient.

Bartels and Doverspike found the opposite to be true, with more intelligent raters demonstrating greater leniency than less intelligent raters. Smither and Reilly (1987) found that there was no significant direct relationship between intelligence and elevation accuracy overall, although in certain rating contexts—specifically, when the job components being assessed were highly intercorrelated—a relationship did emerge. Studies by Dewberry et al. (2013) and Borman and Hallam (1991) showed no relationship between cognitive ability and rater leniency-severity. Further research is necessary to more fully understand the relationship between rater cognitive ability and leniency-severity of ratings. However, the available evidence suggests that, if a relationship exists at all, it is likely not great in magnitude and variable by features of the rating context.

Demographics. In addition to trait-based rater characteristics, a number of demographic variables have also been investigated as correlates or performance ratings and rating leniency-severity. This section will review four demographic characteristics that have been the subject of investigation in the performance appraisal literature: gender, race, age, and culture. Within each section, general findings on differences in rater leniency-severity across rater demographic groups will be presented. Furthermore, leader-member exchange theory has been applied to performance appraisal research to posit that dyadic relationships are affected by in-group / out-group status, and as a result, demographic similarities or differences between raters and ratees could impact leniency-severity (e.g., Duarte et al., 1994; Kacmar, Witt, Zivnuska, & Gully, 2003; Levy & Williams, 2004; Vecchio & Gobdel, 1986). As such, literatures examining the interaction of rater and ratee demographic characteristics will be briefly reviewed.

The role of gender in the performance appraisal context has received considerable attention. However, the majority of these studies have focused on the impact of ratee gender on rating outcomes (e.g., Eagly, Makhijani, & Klonsky; Millmore, Biggs, & Morse, 2007; Roth, Purvis, & Bobko, 2012). The impact of rater gender on leniency-severity, however, has also been subject to some investigation. A fairly consistent phenomenon emerging from gender research in performance evaluation is that males tend to overestimate their own performance, whereas females tend to provide lower and more accurate self-ratings of their performance (Fletcher, 1999; Ostroff, Atwater, & Feinberg, 2004). Somewhat surprisingly, however, the direct effects of rater gender on leniency-severity in ratings of *others* has not often been the central focus of studies. The available research in this domain suggests that female supervisors may provide slightly more severe ratings than male supervisors (Furnham & Stringfield, 2001; Ng, Koh, Ang, Kennedy, & Chan, 2011; Varma & Stroh, 2001). However, more research is needed in this area before reaching firm conclusions. An interesting domain of research examining the relationship between gender and leniency-severity focuses on the interaction of rater and ratee gender in influencing performance ratings. A study by Varma and Stroh (2001) demonstrated that, as previously noted, female supervisors provided more severe ratings than male supervisors. However, they found an interaction between rater and ratee gender, by which female supervisors rated male subordinates more severely than they did female subordinates. In contrast, the ratings of male supervisors did not differ significantly between male and female direct reports. LMX quality played an important role in this relationship, with same-sex supervisor-subordinate pairs having greater LMX quality than opposite-sex pairs, and female supervisors being especially influenced by

LMX quality when evaluating performance of direct reports. In whole, this research suggests that rater gender may have a minor impact on leniency-severity in performance evaluation processes. However, further investigation is needed to establish a more robust understanding of how rater and ratee gender impact rating outcomes.

A related line of research has examined the relationship between race and rater leniency-severity. Similar to the available research on gender, the race of the rater has rarely been the focal area of concern; instead, most studies have examined the impact of the ratee's race on performance ratings, or the interaction of rater and ratee race on ratings (e.g., Kraiger & Ford, 1985; Sackett & DuBois, 1991; Stauffer & Buckley, 2005). These studies have typically used theories of similarity-attraction and/or leader-member exchange as the basis of their hypotheses (Stark & Poppler, 2008). There has been considerable back-and-forth in this literature, with many conflicting findings. While Kraiger and Ford (1985) and Mount, Sytsma, Hazucha, and Holt (1997) found that raters tend to provide higher ratings to ratees of their own race, other authors have disputed this conclusion. Pulakos, White, Oppler, and Borman (1989), Oppler, Campbell, Pulakos, and Borman (1992), and Sackett and Dubois (1991) found that this pattern did not necessarily hold true, and that the relationship between rater race, ratee race, and performance ratings was more nuanced. In terms of direct effects of rater race on leniency, findings are also mixed. Sackett and DuBois (1991) find no substantial differences in ratings from white raters versus black raters overall, whereas Mount et al. (1997) found that black raters were more lenient for both white and black ratees. Many authors have also noted that the overall effect of racial demographics on performance ratings is negligible in magnitude (Brutus, Fleenor, & McCauley, 1999; Stark & Poppler, 2008; Pulakos et al., 1989). Since

its peak in the 1980s and 1990s, this line of research has tapered off, with some authors concluding that no consistent pattern of results has emerged (e.g., Ellis, Ilgen, & Hollenbeck, 2006).

A third demographic that has been researched in the rater bias literature is that of rater age. In keeping with the pattern of research focus on demographic characteristics and job performance, the focus has been on ratee characteristics rather than rater characteristics. The findings on ratee age and job performance has been notably inconsistent, with even meta-analytic evidence arriving at disparate conclusions. Meta-analyses on the relationship between ratee age and job performance have shown that older employees have moderately higher overall job performance than younger employees (Waldman & Avolio, 1986), younger employees have slightly higher job performance ratings than older employees (Gordon & Arvey, 2004), no relationship between age and job performance (McEvoy & Cascio, 1989), an inverted U-shaped relationship between age and performance (Sturman, 2003), and evidence that, although core task performance is largely unrelated to age, other dimensions of job performance do exhibit relationships (Ng & Feldman, 2008). Other studies have looked at the impact of age differences between supervisors and subordinates on job performance ratings; in general, these studies have found little evidence that age differences have a practically significant impact on performance ratings (e.g., Van der Deijden et al., 2010; Vecchio, 1993). While relatively rare, there has been some direct investigation of the impact of rater age on job performance ratings. Cleveland and Landy (1981) found that rater age did not have a substantial impact on performance ratings. In contrast, Griffith and Bedeian (1989) did find that younger raters are more severe than older raters, although

this effect was not large. Interestingly, Liden, Stilwell, and Ferris (1996) found that older supervisors and younger supervisors gave similar subjective performance ratings to their direct reports; however, the subordinates of the older supervisors outperformed subordinates of younger supervisors on objective measures of performance, which the authors interpreted as direct reports benefitting from the experience of their older supervisors. Although speculative, this could be interpreted as severity on the part of the older supervisors, as their subjective ratings of their direct reports did not reflect their subordinates' superior performance on objective indices of performance.

Finally, cultural differences in performance appraisal have recently emerged as an important domain of research. Given the increasing prevalence of multinational corporations and globally administered HR practices, it has been acknowledged for some time that cultural differences must be accounted for (e.g., Hofstede, 1983); however, until recently these differences have not been carefully examined in the context of performance evaluation. Recent research has investigated the impact of cultural differences on performance appraisal from two distinct perspectives: the comparison of performance ratings between groups of different national or regional origins, and the study of how specific cultural traits, largely focused on Hofstede's five dimensions of cultural difference (Hofstede, 1992), impact performance ratings. In a comparison of supervisor performance ratings in Canada, South Korea, and Spain, Ployhart, Wiechmann, Schmitt, Sacco, and Rogg (2003) demonstrated that, although job performance ratings are in many ways invariant across the three countries included in the study, there was clear evidence that implicit theories of performance *did* differ—for example, there were small to moderate differences in the extent to which age, tenure, and

opportunity to observe were associated with performance ratings. In a comparison of performance ratings between the U.S. and India, Varma, Pichler, and Srinivas (2005) found that supervisors' liking of subordinates was largely unrelated to performance ratings in the U.S. sample, but that Indian supervisors inflated ratings of well-liked low performers. Adsit, London, Crom, and Jones (1997) found that upward evaluations differed across regions, with U.S. and European raters evaluating their supervisors more favorably than Brazilian and Asian raters. DeVoe and Iyengar (2004) demonstrated that raters from North American, Asian, and Latin American countries differed in how their perceptions of employees' intrinsic versus extrinsic motivation was related to their evaluations of those employees' performance. Similarly, studies leveraging Hofstede's model of cultural differences (individualism-collectivism, power distance, masculinity-femininity, uncertainty avoidance, and long / short-term orientation) found that performance ratings were related to some of these dimensions. For example, Ng, Koh, Ang, Kennedy, and Chan (2011) showed that raters' power distance and individualism-collectivism were associated with leniency-severity of performance ratings. Hu, Hsu, Lee, and Chu (2007) and Zhou and Martocchio (2001) found that individualism-collectivism influenced which aspects of performance (e.g., maintaining relationsihps) were more heavily weighted in overall ratings of job performance and compensation decisions. Mishra and Roch (2013) found similar patterns in peer and subordinate ratings. This area of research has just recently been a major focus of empirical investigation, and several authors (e.g., Claus & Briscoe, 2009; Cho & Payne, 2016) have noted that there is still much work to be done before any definitive conclusions can be made. However,

available evidence does seem to suggest that cultural differences play a substantial role in how raters evaluate others in performance appraisal contexts.

To summarize the literature on demographic characteristics and rater behavior in performance evaluation situations, there are a great number of studies which have—and have not-- found meaningful relationships between demographic characteristics and performance rating patterns. In those cases where relationships have been found, these have tended to be modest in size, and dependent upon a number of complex situational and interpersonal factors (Arvey & Murphy, 1998). For most of these characteristics, further research is required. However, demographic variables are often controlled for in examinations of rater effects on performance evaluations, and the research available to date suggests that this practice is likely warranted. Although demographic characteristics are unlikely to explain a sizeable proportion of variance in rater leniency-severity on their own, controlling for these rater attributes may help clarify how various rater characteristics interact to impact rating patterns.

Perspective. The vast majority of the research on rater effects on performance evaluations has focused on supervisory ratings of subordinate performance. However, increasingly, with the proliferation of multisource performance ratings or 360 degree evaluations (Conway & Huffcutt, 1997; Tornow & London, 1998), research has focused not only on how supervisors evaluate the performance of their direct reports, but how colleagues from different rater sources (e.g., supervisor, peer, direct report, self) evaluate a target employee. In this section, the literature on this topic will be reviewed and patterns in leniency-severity between different rater sources will be examined.

For some time, there was some speculation as to whether rater source effects should be interpreted as unique and complementary perspectives on employee performance, or as a reflection of rater source bias primarily driven by idiosyncratic rater effects (Mount et al., 1998). After considerable investigation and discussion of the appropriate analytical approach to identifying the presence of systematic rater source effects in multi-source performance ratings, the resulting conclusion has been that, although rater source effects are not large, they do, in fact, represent distinct perspectives on an employee's performance (Bynum, Hoffman, Meade, & Gentry, 2013; Hoffman & Woehr, 2009; Hoffman, Lance, Bynum, & Gentry, 2010; Lance, Hoffman, Gentry, & Baranik, 2008). What has been less closely examined, however, is whether these unique rater sources demonstrate differential patterns in rater errors.

The most well-researched element of this line of inquiry has been on the unique characteristics of self-ratings as compared to ratings from other rater sources. In their 2009 meta-analysis of 115 studies, Heidemeier and Moser examined the relationship between self-ratings and supervisory ratings of performance. They found that self-ratings correlate weakly (r=.22, ρ=.34) with supervisory ratings. They also found that self-ratings were substantially more susceptible to leniency than were supervisory ratings (d=.32). Other research has also demonstrated that self-ratings of job performance have low convergent validity with other criteria, including non-supervisory others and objective performance criteria. (e.g., Conway & Huffcutt, 1997; Dunning, Heath, & Suls, 2004; Hoffman, Nathan, & Holden, 1991). Overall, self-ratings are plagued by high leniency, low accuracy, and poor convergent and criterion-related validity.

Ng et al. (2011) examined how different rater sources demonstrate differential levels of leniency. As they predicted, subordinate ratings demonstrated the highest levels of leniency. The authors attribute this to subordinates' reluctance to convey negative information to their immediate supervisors. Peers demonstrated greater leniency than supervisors, but less leniency than subordinates. Supervisors were the least lenient in their performance evaluations.

This information reflects the importance of considering rater source in examinations of rater leniency. Raters from different sources are differentially likely to be lenient in their ratings of others. As such, any examination of leniency effects should account for the perspective of raters providing the ratings, and should ensure that analyses of leniency leveraging multisource performance ratings should control for the proportion of ratings provided by raters from different sources.

Performance. The focus of the present research is in examining the relationship between a rater's performance and the leniency or severity with which they assign performance ratings to others. To date, this has been a gap in the literature on rater leniency-severity; rater performance has not often been among the characteristics examined in research on rater leniency-severity, and the very limited research that does exist has focused exclusively on supervisor ratings in small-sample studies. In a later section, the theory and rationale underlying a hypothesized relationship between rater performance and leniency-severity will be explicated in greater detail. However, in the present section, the available research on rater performance and its relationship with rater leniency will be reviewed.

An early study related to the focal topic of this paper was conducted by Schneider and Bayroff (1953), and is a rare exception to this literature's focus on supervisors as the rater of interest. In this study, officers who were students in the Army Command and General Staff College provided peer ratings of other officers with whom they were classmates. In total, 400 individuals provided ratings of their peers' on a number of dimensions, including overall value to the Army, aptitude, and achievement at the college. Raters were divided into three groups based on their overall scores across dimensions: high-performing, middle, and low-performing officers. Within each of these three groups, the validity of the officers' ratings in predicting several performance criteria, including final class standing, were examined. The study found that high-performing officers provided more valid ratings of performance for their peers than low-performing officers, suggesting a link between rater performance and the quality of ratings assigned to others.

One of the first studies to directly examine the impact of rater performance on leniency-severity was conducted by Kirchner and Reisberg (1962). In this study of supervisor ratings in a sample of technical employees, supervisors were categorized as either high or low in supervisory effectiveness based on both (a) overall job performance ratings of the supervisor, and (b) specific evaluations of supervisory effectiveness provided by the supervisors' managers. Supervisors who received mixed evaluations from these two sources of information were dropped from the analysis, resulting in a final sample of 17 "better" supervisors and 11 "less-effective" supervisors. Next, these supervisors provided ratings of overall job performance and technical performance for their direct reports. The authors found that there was greater variance in the ratings

provided by better supervisors than those given by less-effective supervisors, and that less-effective supervisors were more lenient overall in their ratings. Of course, little can be definitively concluded based on such a small (and dated) sample of supervisor ratings. However, these findings were replicated in a very similar study conducted by Levy and Stone (1963).

In a related, but slightly more tangential, investigation of the relationship between supervisor effectiveness and rating behavior, Gruenfeld & Weissenberg (1966) examined how supervisor effectiveness related to supervisors' attitudes towards performance appraisal. The authors measured supervisor effectiveness via a self-report instrument assessing several characteristics of the supervisor: intelligence, supervisory quality, initiative, self-assurance, occupational level, structure, and consideration. The criterion measure was an overall rating of supervisors' favorable (or unfavorable) attitudes toward performance appraisal. Example items include "the time spent in the preparation and review of performance appraisals is not generally time wasted," and "formal performance appraisal reviews of subordinates should be conducted at least once a year." In a sample of 72 male supervisors in a public sector finance department, the authors found that supervisory quality, initiative, self-assurance, consideration, and structure were all positive related to supervisors' attitudes towards performance appraisal. In their discussion, the authors suggest that these findings suggest that these more effective supervisors are likely to be more diligent and accurate in their evaluations of direct reports' performance.

In a review article on rater characteristics and performance appraisal validity, Bedeian (1976) summarized the available research on rater performance and rating

outcomes and concluded that "findings strongly suggest the existence of a relationship between supervisory ability and both the development of subordinates and the validity of performance appraisal measures" (p. 40). This conclusion was based largely on the studies described above, as well as a handful of other studies only tangentially related to the central question of rater performance and rating accuracy. Although the available evidence at that time did suggest a possible relationship, there was far from sufficient research to draw any firm conclusions. However, since the 1970s, rater performance has only infrequently been evaluated as a predictor of rating accuracy or leniency-severity, and has rarely been a central hypothesis of investigation.

More recent research has examined leniency-severity as a product of more specific rater competencies. Hauenstein (1992) introduced the notion of performance management competence (PMC)—a collection of knowledge, skills, abilities, and motivations related to managers' ability to successfully implement performance management practices, of which performance evaluation is one component. In a recent study, Bernardin et al. (2016) introduced a measure of performance management competence. They defined PMC as the ability to detect performance problems, take strategic corrective action, precisely define performance standards, and accurately observe and assess performance against these defined standards. In a sample of 125 associate managers in retail stores, Bernardin et al. found that elevation accuracy in evaluations of direct reports, peers, and supervisors was positively associated with the raters' performance management competence. This research suggests that behaviors associated with performance management, including performance evaluation, are a

central performance requirement for managers, and these behaviors are also associated with manager's leniency-severity in evaluating others.

In summary, the research on rater performance as a predictor of leniency-severity is due for more thorough consideration. Available evidence suggests that a rater's level of performance may have an impact on the validity of the ratings they assign to others. However, there has been very limited investigation of the direct effect of rater performance on leniency-severity. The few studies that have investigated this have largely focused on supervisor ratings in small, dated samples of employees. As will be discussed later, there are a number of reasons which would suggest rater performance might be expected to impact the leniency or severity with which they evaluate others. However, no clear picture has yet emerged of how rater performance, across rater perspectives, influences the leniency-severity of their rating behaviors.

Strategies to Reduce Leniency-Severity

As discussed in the previous section, available research has suggested that various factors, including rating format, context, and rater characteristics, have influence on the leniency or severity with which raters assign ratings. There is still much work to be done in order to fully understand these phenomena. However, our incomplete understanding of the causes of rater leniency-severity has not prevented researchers and practitioners from proposing and testing a number of interventions to mitigate the effects of leniency and severity in performance evaluations. The theoretical rationale for each of these interventions, and the relative efficacy of each in reducing leniency-severity effects in ratings, provides useful insight into the underlying mechanisms driving rater leniency-severity. These insights are useful in building a theoretical rationale for the hypothesis

that rater performance impacts leniency-severity. As such, this literature will be briefly reviewed before this paper turns to its central hypotheses.

Statistical correction. The first strategy which has been employed to address issues of leniency-severity in performance ratings is the use of after-the-fact statistical corrections to adjust rating differences due to individual differences in rater leniency-severity. One of the first to suggest this method was Ford (1931). Ford argued that managers vary in their leniency-severity, and that efforts to encourage similar rating patterns through training are unlikely to be successful. Instead, Ford suggested a fairly straightforward practice of comparing the means and standard deviations in the distributions of ratings for each manager across ratees, and using a correction factor for each individual employee's ratings in order to artificially equalize the means and standard deviations for each rater's scores across ratees. Since this early study, statistical corrections for leniency-severity have grown more sophisticated and nuanced in consideration of various rating scenarios. Houston, Raymond, and Svec (1991) present a review and comparison of three of the more common methods for correcting individual ratings for leniency-severity effects: ordinary least squares, weighted least squares, and imputation of missing data (where the "missing data" are ratees' ratings from raters who did not evaluate that ratee). Houston et al. found that all three methods provided greater accuracy than uncorrected ratings, and that the imputation method offered greatest accuracy. Raymond et al. (2011) provided addition evidence of the improved reliability in performance ratings when statistically corrected, and even demonstrated that the same computational corrections could be applied to new sets of ratees (evaluated by the original set of raters) to reduce error. In a somewhat different vein, Hoyt (2000)

developed a method for correcting correlations between performance ratings and other variables to adjust for the biasing effects of rater inconsistency in leniency-severity. Hoyt's recommendations accounted for the rating scenario (e.g., fully crossed vs. nested ratings) as well as whether or not observations are linked. Hoyt's methodology leveraged generalizability theory and interrater reliability to provide guidance on correcting correlation coefficients for attenuation.

These and other investigations of post hoc corrections for error attributable to rater leniency-severity have been valuable for several reasons. First, they are often recommended and successfully used for the purposes of reducing the impact of error in ratings used in research settings. These methods can help to provide more accurate ratings in studies in which employee performance is a key variable. Moreover, these studies have demonstrated that rater leniency-severity is a relatively consistent attribute of raters that can generalize beyond a single set of evaluated employees (e.g., Raymond et al., 2011), confirming research from other sources which has reached similar conclusions (e.g., Kane et al., 1995). However, for a number of reasons, this method is limited in its practical application. Despite its adoption in many research settings, organizations have rarely been willing to adopt post-hoc statistical corrections of ratings in applied settings (c.f., Harik et al., 2009; McManus, Thompson, & Mullen, 2006). Additionally, many of these models rely on the assumption that any differences in the mean and distribution of ratings assigned by raters is erroneous, and does not reflect actual differences in the performance of employees in one group over another. This, despite the fact that ample research has demonstrated that the performance of direct reports is likely to differ between groups due to various attributes of the supervisor (e.g.,

Chen, Bian, & Hou, 2015; Judge & Piccolo, 2004; Sauer, 2011). Likewise, very few of these methods have been applied in settings where a limited number of raters assign ratings concurrently, or when each rater provides ratings for a small number of subjects (Raymond et al., 2011). For these reasons and others, there has been far more attention devoted to a priori interventions to reduce leniency-severity in ratings of performance, and these strategies have been much more extensively adopted in applied settings.

Rating Method. As discussed previously, there is a massive body of research on the impact of the rating instrument and rating accuracy. Landy and Farr (1980) provide a comprehensive review of this literature. For many years, research focused on the optimal design and process for performance rating instruments. Ultimately, this approach did not lead to conclusions which could adequately resolve the problems of performance rating. Bernardin (1977) summarized this sentiment nicely: "One conclusion that can be drawn from this research is that nothing conclusive can be said with regard to respective psychometric characteristics. Further research in this area would probably only substantiate this" (p. 426). This was followed by Landy and Farr's (1980) moratorium on this area of research, which was largely heeded.

However, the research cited above focused exclusively on direct (rather than comparative) methods of performance appraisal. Landy and Farr explicitly clarify this, noting that comparative methods represent "a qualitatively different discrimination process" (p. 73). Comparative methods of performance rating involve comparing the evaluated employee to other employees before assigning a rating, and can take a variety of forms, including rank ordering, peer to peer comparisons, relative percentile ratings, and others. The most straightforward comparative method of performance rating is rank

ordering, in which all employees are ranked in order from lowest to highest performers (or ranked according to a specific dimensions of performance), and then ratings are assigned based on the employee's rank relative to peers. In their meta-analysis, Nathan and Alexander (1988) showed that rank ordering results in higher validity than subjective ratings of performance. A related method, the paired comparison technique, involves comparing each employee to each other employee in a particular group and, with each comparison, the rater decides which of the two employees is superior. This results in an overall rating based on an index of how often each employee was superior to the other employees in the paired comparisons (Lawshe, Kephart, & McCormick, 1949). Despite their demonstrated incremental validity over direct rating methods, these comparative techniques have a number of widely acknowledged limitations—they are cumbersome and time-consuming to implement, are socially contentious, ratings are not easily comparable across raters, and ratings are dependent on who each employee is being assessed against (Jewell, 1998). For these reasons, comparative methods have been adopted in only 4% of published research studies, and likely even less often in applied settings (Goffin, Jelley, Powell, & Johnston, 2009). The relative percentile method represents a hybrid of comparative and direct subjective rating processes, whereby a rater places each of his or her evaluated employees on a continuum based on their relative percentile— what percent of employees within the organization, as subjectively evaluated by the rater, perform lower than the target employee. This method has been demonstrated to outperform direct evaluations in the validity of performance ratings (Goffin, Gellatly, Paunonen, Jackson, & Mayer, 1996; Goffin et al., 2009).

Despite the disappointing outcome of the research on rating instrument and methodology, these studies have important implications for the focal hypothesis of this paper. Specifically, the research on rank ordering, paired comparisons, and the relative percentile method of performance appraisal reveals that social comparisons play an important role in how raters evaluate subjects. Raters appraise others differently when comparing subjects to other employees, and as noted by Jewell (1998), these comparative evaluations are dependent on the comparison group against which an employee is being judged. This point will be resurfaced later as a central rationale underlying a hypothesized relationship between rater performance and leniency-severity.

Rater Training. Rater training is widely regarded as the most successful strategy for reducing rater errors (Roch et al., 2012). Since Landy and Farr's (1980) moratorium on scale design research, rater training has been the most actively researched intervention for improving rater accuracy. In broad strokes, rater training achieves two goals which serve to enhance the quality of ratings: increasing raters' knowledge and skill related to observing and evaluating employee behavior, and enhancing raters' motivation to provide quality ratings (McIntyre, Smith, & Hassett, 1984). Rater training has taken a variety of forms in research and practice. In their review and meta-analysis of this literature, Woehr and Huffcutt (1994) built on earlier work conducted by Smith (1986) to introduce a framework for categorizing these training methods, arriving at four general approaches: rater error training, performance dimension training, behavioral observation training, and frame of reference training. In this section, each of these approaches will be briefly reviewed, with particular attention to frame of reference training, the most widely adopted and successful method of rater training.

Rater error training (RET) was the earliest, and perhaps most obvious, of the rater training approaches. The rationale behind RET is that, in order to reduce halo error, leniency, central tendency, and contrast errors, raters should be trained on what these errors are and instructed to avoid committing them in the rating process (Pulakos, 1984). In some iterations of this approach, raters were presented with distributions of performance ratings which reflect these errors (e.g., negatively skewed distributions), and were guided to identify the errors in the distributions. However, Latham (1986) and other argued that this distribution-based method was flawed, in that it often resulted in raters forcing ratings to reflect "ideal" performance distributions, sometimes to the detriment of the actual accuracy of those ratings. The preferred method, according to Latham and others, is simply to instruct raters about the various forms of rating error and instruct them to avoid committing these errors. In Woehr and Huffcutt's (1994) meta-analysis, this avoidance-based method was demonstrated to positively impact several dependent measures of rating quality, including leniency (d=.21). However, rater error training has fallen out of favor since this early research, with only a few published studies adopting this method since Woehr and Huffcutt's paper was published (Roch et al., 2012).

As mentioned previously, performance appraisal research in the 1980s shifted to focus on the cognitive processes underlying the evaluation of performance (e.g., Feldman, 1981; Lance & Woehr, 1986). Born from this effort was the adoption of performance dimension training (PDT) as a means of reducing rating errors. A central premise of the cognitive approach to performance appraisal is that, in many ways, evaluations of a ratee's performance is guided by judgments that are made at the time a behavior is observed, rather than at the moment of performance evaluation (e.g.,

Feldman, 1981; Lichtenstein & Srull, 1987; Murphy, Philbin, & Adams, 1989). As a result, obtaining accurate ratings of performance relies heavily on a rater's ability to recall and consider only performance-related behaviors in the evaluation of performance, rather than relying on judgments of performance-irrelevant observations. Essentially, PDT is an effort to train raters on the dimensions of performance being assessed. This could be as simple as reviewing the rating scale with the raters, or by having raters actually participate in the development of the rating scale. The primary focus of this training method has been to improve rating accuracy, with a handful of studies showing small to moderate effects of PDT on rating accuracy. However, the limited research on the effectiveness of PDT for reducing leniency shows that this method is not particularly useful in this respect (Woehr & Huffcutt, 1994). This is not surprising, given that PDT focuses on what is being measured, *not* how to scale or differentiate between levels of effectiveness on specific dimensions of performance. Similarly to RET, performance dimension training is not often adopted in current research or practice—at least not independently of other rater training methodologies (Roch et al., 2012).

A third category of rater training is behavioral observation training (BOT). This approach deviated from others in that its emphasis is not on rater's judgment in the appraisal process, but rather in their ability to effectively observe behavior. This approach was first introduced by Thornton and Zorich (1980). In a BOT intervention, evaluators are guided to carefully observe behavior and record observations wherever possible. In most instances, this guidance is coupled with instructions on the critical dimensions of performance to be assessed and how to identify behaviors related to these dimensions (e.g., Hedge & Kavanagh, 1988). Unlike RET, which focuses primarily on the avoidance

of rater biases such as halo and leniency, and PDT, which is primarily concerned with rating accuracy, the main criterion in studies of BOT is *observational accuracy*—raters' ability to accurately recall specific behaviors pertinent to the evaluated dimensions of performance. Relative to other training programs, BOT has been rarely employed, and almost all investigations are in research settings rather than field settings. BOT has primarily been used to improve observational accuracy within a very limited time frame, rather than over the course of a year, as is typically the purpose of formal performance appraisal processes. For this reason, although BOT has been demonstrated to improve observational accuracy in experimental settings during a short time frame, there has been little investigation of its impact on rating accuracy or rater errors (Woehr & Huffcutt, 1994).

By far the most researched and most successful rater training method has been frame of reference (FOR) training. Although earlier researchers had adopted training methods akin to FOR (e.g., Latham et al., 1975; Wexley, Sanders, & Yukl, 1972), Bernardin and Buckley (1981) were the first to explicitly outline the underlying rationale for FOR training and present it as a novel approach to improving rater accuracy. As described by Bernardin and Buckley, the goal of FOR training is to establish a common conceptualization of performance (frame of reference) among raters. Borman (1987) demonstrated that individuals hold distinct theories of work performance; that is, they differ in their beliefs (either conscious or unconscious) about what defines effective performance. These discrepancies can take many forms. For example, raters may differ in which dimensions they use to define performance (Schleicher & Day, 1998), the behaviors used to evaluate each dimension (Ellett, Wren, Callender, Loup, & Liu, 1996),

and the behaviors which illustrate specific levels of effectiveness or ineffectiveness within a particular dimension of performance. Some individuals have more discrepant theories of work performance than others (Uggerslev & Sulsky, 2008). The purpose of FOR training is to establish a common frame of reference for raters within an organization by facilitating alignment among raters on which dimensions of performance are to be evaluated, what behaviors are associated with each dimension, and how to evaluate degrees of effectiveness or ineffectiveness in each behavior (Sulsky & Day, 1992). Essentially this is a process aimed at calibrating theories of work performance across a set of raters or an organization (Athey & McIntyre, 1987).

Although the exact protocol for conducting FOR training varies, there are a few process elements which are common to most interventions (Bernardin, Buckley, Tyler, & Wiese, 2000). The first stage of FOR training is similar to performance dimension training. In order to reduce idiosyncrasies in raters' implicit theories of performance, raters are trained on the dimensions of performance that are to be assessed. Next, the training differentiates between levels of performance within each dimension. This is typically done by providing behavioral examples of performance at each level of effectiveness within each performance dimension. Having been instructed on the dimensions and levels of effectiveness, trainees then practice by evaluating subjects on each dimension of performance. These "subjects" are typically prepared vignettes developed as part of the training. Finally, raters receive feedback on their ratings. Each rating is contrasted with a "true score" rating of the subject's performance on each dimension. Participants are often provided the opportunity to engage in several rounds of

practice and feedback until their theory of performance aligns with that which is endorsed by the organization.

When it was originally conceived, there was not a thorough theoretical justification for why or how FOR training improves the accuracy of performance ratings. However, subsequent work has offered substantial insight into the cognitive processes impact by FOR training. In particular, a number of studies have examined how FOR training impacts rater's recall and recognition memory of performance-relevant behaviors, and the categorization of these behaviors into performance dimensions. Sulsky and Day (1992) demonstrated that FOR training results in cognitive prototypes of varying levels of effectiveness in each performance dimension. These prototypes are then used by raters when processing observed behaviors. They showed that, in a recognition memory task, FOR training resulted in enhanced recognition memory for performance-relevant behaviors. Building on this research, Woehr (1994) found that raters' recall of behaviors was improved following FOR training, and that these memories were more closely related to the actual judgments or evaluations as reflected in the ratings of performance. These studies helped to clarify the mechanisms by which FOR impacts rating outcomes—the training results in the development of performance schemas which not only help in the recall of performance-relevant behaviors, but also help to categorize these observations into the appropriate dimensions of performance when assigning ratings.

Frame of reference training has been demonstrated to improve ratings across several criteria, and in diverse evaluative contexts. In the first meta-analysis examining the impact of frame of reference training, Woehr and Huffcutt (1994) used four criteria to

evaluate training effectiveness: halo, leniency, rating accuracy, and observational accuracy. As was typical at the time, halo was operationalized as intercorrelations across performance dimensions, leniency as the presence of negatively skewed distributions of performance ratings, rating accuracy as the absolute average deviation of subject ratings from subject true score performance, and observational accuracy as the deviation of rater's recall of behavioral incidents from the actual occurrence or presence of these incidents. For each of these indices, a meta-analytic *d*-value was calculated to contrast the experimental groups (trainees) with control groups (non-trainees) across 15 empirical studies of FOR training (although not all criteria were reported in each study). This meta-analysis found evidence that FOR training leads to the greatest increases in rating accuracy relative to other rater training techniques (*d*=.83, *k*=6), with small to moderate effects for halo (*d*=.13, *k*=4), leniency (*d*=.15, *k*=3), and observational accuracy (*d*=.37, *k*=2). In 2012, an updated meta-analysis (Roch et al., 2012) greatly expanded the number of studies included in this investigation. This updated meta-analysis found that FOR training was associated with positive rating outcomes in terms of elevation accuracy (*d*=.41, *k*=21), differential elevation (*d*=.45, *k*=20), stereotype accuracy (*d*=.49, *k*=18), differential accuracy (*d*=.44, *k*=28), and behavioral / observational accuracy (*d*=.88, *k*=9). Frame of reference training has also been demonstrated to result in performance schema or theories of performance which more closely align with those endorsed by the organization (Gorman & Rentsch, 2009). As evinced by the included studies in Roch et al.'s (2012) meta-analysis, it is clear that the primary application of FOR training has been to improve the quality of ratings in performance appraisal scenarios. However, FOR training has also been demonstrated to positively impact rating accuracy in interviews

(e.g., Melchers, Lienhardt, von Aarburg, & Kleinmann, 2011; Stevens, 1995), assessment centers (e.g., Jackson, Atkins, & Fletcher, 2005; Schleicher, Day, Mayes, & Riggio, 2002), and assorted other contexts (e.g., Angkaw, Tran, & Haaga, 2006; Dierdorff, Surface, & Brown, 2010; Lundstrom, 2008).

The research pertaining to frame of reference training has significant implications for a hypothesized relationship between rater performance and rater leniency-severity. The FOR literature is just one of several bodies of research which, as a whole, suggest that a rater's own level of performance might impact how they evaluate others. In the following section, these bodies of research will be summarized in order to provide the theoretical and empirical rationale underlying a hypothesized rater performance – rater leniency relationship.

Rationale for a Rater Performance – Rater Leniency Relationship

Raters hold distinct performance schema. Multiple streams of research have established that individuals develop cognitive structures in order to process social information, and that these cognitive structures influence how we perceive, encode, recall, and interpret behavior. Although the term *schema* has been in use across disciplines for centuries, Bartlett (1932) introduced the term to describe how, over time, people organize knowledge into cognitive structures which facilitate (often erroneously) our memory or judgments of a particular observation. This concept was built upon by social cognition researchers. Rosch (1978) and others introduced the concepts of prototypes—cognitive representations of what it looks like when a subject meets all of the fundamental criteria of a particular category. This was closely related to the notion of stereotypes, which is specifically focused on schema related to groups of people and the

characteristics attributed to those groups (Hamilton & Gifford, 1976). Of particular note in this early research was personal construct theory, emerging from clinical psychology (Kelly, 1955). Leveraging a method called the Kelly Repertory Grid, Kelly demonstrated that people develop disparate schema (or *personal constructs),* which in turn influence how an individual interprets the behaviors and characteristics of others and anticipates future behavior.

Borman (1987) was the first to synthesize the considerable research on schema and personal constructs and apply those insights to the evaluation of performance in a work setting. In this initial study, 25 U.S. military officers were asked to describe the key differentiators between effective and ineffective performance in the role of a non-commissioned officer, a first-line supervisory position which all of the officers were highly familiar with. Borman analyzed the responses of these military officers and came to the conclusion that, although some criteria for effective performance were widely shared across the military officer sample, there was considerable variability in the performance criteria included in any single officer's performance schemata, as well as the relative significance placed on each dimension of performance.

Borman's research focused primarily on the content of the performance schema. Further investigations (closely intertwined with the literature on frame of reference training) have not only corroborated Borman's initial findings, but also expanded our understanding of how individuals differ in their implicit theories of work performance. This body of research has led to the conclusion that individuals differ not only on which dimensions of performance they believe to be important for a particular role, but also which behaviors are associated with a particular dimension of performance, and what

behaviors represent a specific level of effectiveness or ineffectiveness (Ellen et al., 1996; Schleicher & Day, 1998; Sulsky & Day, 1992; Uggerslev & Sulsky, 2008). Importantly, employees develop distinct prototypes across levels of performance, (e.g., prototypes for high, average, and low performance) and these prototypes vary in the extent to which they coincide with espoused organizational standards (Sulsky & Day, 1992). These findings—and the last point in particular regarding levels of effectiveness—has direct implications for our understanding of leniency-severity. Available evidence suggests that a rater's leniency or severity with which they evaluate others might be impacted by their individual performance schema, particularly those prototypes related to degrees of effective or ineffective performance within or across performance dimensions.

People make self-referent comparisons when judging others. Having established that theoretical and empirical evidence suggests a link between raters' performance schema and leniency-severity, the next question is—where do these performance schema come from, particularly regarding raters' prototypes for high, average, and low performance? As described previously, the research on cognitive schema in performance appraisal asserts that these schema develop over time as employees gain experience with a particular role or function (e.g., Borman, 1987). However, this literature has not fully addressed *how* performance schema develop, or what factors influence the cognitive categorization of behaviors into prototypes of effectiveness. The social cognition literature and social comparison theory (e.g., Festinger, 1954) provides some insight which may help to address this gap.

In a recent article, Goffin et al. (2009) called on performance appraisal researchers to consider the implications of social cognition theory in their research. At a

high level, modern social comparison theory argues that, when objective criteria are unavailable, people employ comparative judgments to evaluate themselves and others (Kruglanski & Mayseless, 1990). Prior research had primarily focused on social comparison theory as an explanation for the development of self-concept and evaluation of one's own performance (e.g., Thornton & Arrowood, 1966); however, this emphasis on understanding of self has since broadened to encompass social comparisons more broadly. As a result of this broadened conceptualization, researchers have learned a great deal about how evaluations of others are informed by comparisons to the self.

Specifically, researchers in social psychology have found that, in making judgments or evaluations of others, people use the self as a "habitual referent" (Catrambone, Beike, & Niedenthal, 1996; Karylowski, 1990). That is, we are naturally inclined to evaluate others against our own self-concept. This expands upon Festinger's (1954) original hypotheses concerning how perception of others influences self-concept, and reveals this process of comparison and evaluation to be bi-directional in nature (Wood, 1989). Researchers have also examined the impact of self-esteem on evaluations of others. For example, Long and Spears (1998) investigated how participants' self-esteem in an experimental task influenced the interpersonal and inter-group (in-group vs. out-group) evaluations of others. They found that high personal self-esteem was associated with both how participants evaluated individuals as well as other groups.

Taken as a whole, this body of research implies that the self is an important reference point used in the evaluation of others. It follows, then, that in evaluations of performance, raters will instinctively compare the performance of others to their perception of their own performance level.

Self-referent evaluations are most likely with similar others. A central tenet to Festinger's (1954) social comparison theory is that the target of comparison is not selected arbitrarily. Festinger hypothesized that "the tendency to compare oneself with some other specific person decreases as the difference between his opinion or ability and one's own increases," leading to the resulting corollary that "given a range of possible persons for comparison, someone close to one's own ability or opinion will be chosen for comparison" (pp. 120–121). Subsequent research has demonstrated that the selection of a comparative referent is more complex; it depends on a number of factors, including individual differences, motivational variables, context, content of the information being compared, and more (Buunk & Gibbons, 2007; Kruglanski & Mayseless, 1990). However, while the selection of referent others is more nuanced than originally hypothesized by Festinger and can be influenced by many factors, Festinger's original hypothesis remains generally supported— similarity is a critical parameter in self-other comparisons (Buunk & Gibbons, 2007).

This is highly relevant in the organizational sciences. Employees selectively identify referents for self-other comparison in organizations, and years of research has examined how this process impacts individual and organizational outcomes. For example, numerous authors have investigated how referents in social comparisons impact distributive justice perceptions in pay (e.g., Goodman, 1974; Sweeney & McFarlin, 2005). Others have considered how social comparison referents influence employee perceptions, including how employees perceive the nature and desirability of the characteristics of their job (e.g., Oldham et al., 1982; Oldham, Kulik, Stepina, & Ambrose, 1986) or the quality of leader-member exchange with their supervisor (Hu &

Liden, 2013). Still other research has considered the ways in which organizational context influences how social comparison referents are selected, such as investigations of the impact of virtual work environments (Conner, 2003) and department size (Oldham et al., 1986) on referent others. In all of these investigations, the basic premise of Festinger's (1954) theory hold true—people tend to choose social comparison referents who are similar to themselves on salient criteria. Additional research continues to validate and extend this understanding of social comparison referent selection in organizations (Goodman & Haisley, 2007).

Of particular importance are several studies which have investigated the relationship between job level and the selection of social comparison referents. For example, in the context of pay comparisons, Major and Forcey (1985) found that people instinctively reference others whose jobs are most similar to their own—including the level of the job within the organizational hierarchy. Bamberger & Biron (2007) found that employees' norms and behaviors related to absenteeism were influenced more strongly by peer-level referents than by departmental or organizational norms. Of particular relevance, Shah (1998) conducted a network analysis to identify who employees choose as their referents when seeking various types of information. The findings revealed that, for job-related information (such as performance expectations), people tend to compare themselves to "structurally equivalent referents"—people who hold a similar job level, function, role, etc. within the organization.

What does this mean for the impact of rater performance on the relative severity or leniency with which they evaluate others? Given the self as a natural referent for social comparisons, and the particular salience of similar others when making self-other

comparisons, we might expect that raters will be more likely to compare their *peers'* performance to their own, versus the likelihood of doing so when evaluating less similar other such as their direct reports, supervisors, or business partners. Peers are more likely to occupy the same hierarchical level and role as the subject being evaluated. As such, the impact of the rater's performance on leniency-severity of ratings may be moderated by the raters' relationship to the subject, with peer ratings demonstrating greater rater performance effects on leniency-severity than ratings from other rater sources.

Employees view their own performance favorably. It has long been acknowledged that self-ratings of performance are far more lenient than ratings provided by other rater sources (e.g., Parker, Taylor, Barrett, & Martens, 1959; Prien & Liske, 1962; Thornton, 1968). This phenomenon is not unique to the context of performance appraisal; similar favorable biases towards the self have been observed in other settings, such as in health and educational contexts (Dunning et al., 2004). The difference between ratings of self and others is substantial; a meta-analysis of supervisory and self-ratings of performance by Heidemeier and Moser (2009) found the difference to be d=.32, Δ=.49. An individual's self-rating is largely unrelated to the ratings provided by supervisors (r=.22), peers (r=.19), and subordinates (r=.14) (Conway & Huffcutt, 1997).

An interesting finding throughout the literature on self-evaluation of job performance is that leniency toward the self is not uniform across the spectrum of low to high performers. Low performers often demonstrate the greatest leniency in their self-assessments of performance, far overestimating their own effectiveness (e.g., Atwater, Ostroff, Yammarino, & Fleenor, 1998; Eichinger & Lombardo, 2004; Shipper & Dillard, 2000). Conversely, the highest performers tend to have more accurate self-appraisals, and

oftentimes are overly severe in their self-ratings (e.g., Church, 1997; Eichinger & Lombardo, 2004). Although this could be attributed, in part, to ceiling effects of the performance rating scales, there is psychological justification for such a phenomenon. For example, the well-known Dunning-Kruger effect reflects the consistent finding that those who are most lacking in knowledge, skill, or experience are least likely to accurately gauge their own competency—and thus are more likely to inflate their own self-appraisal without consciously intending to do so (Kruger & Dunning, 1999). This finding is not only true in the workplace, but for self-evaluations of performance across a diverse range of appraisal purposes and content domains. The implication resulting from this body of research is that nearly all employees view their own performance as, at the very least, satisfactory. Low performing employees believe that their behaviors demonstrate adequate levels of performance, and high performers tend to rate themselves lower than other raters or objective indicators of performance.

Summary of theoretical rationale. Taken together, the preceding research suggests that people define effective performance differently—not just in terms of how performance is defined, but also what qualifies as effective versus ineffective levels of performance. Raters evaluate the performance of others based, in part, by comparing observed behaviors of the subject to their own. The vast majority of employees believe that their own performance is satisfactory. As such, it stands to reason that how a rater evaluates a target employee will be influenced, in part, by their own performance. This would result in greater leniency for low-performing raters, whose standards for effective performance (based on their understanding of their own performance as effective) will be lower than for high-performing raters. Similarly, one might expect greater severity from a

high-performing rater, whose standards for effective performance are informed by their own level of performance. This effect is likely to be particularly pronounced when raters evaluate their peers, with whom they share a similar hierarchical level in the organization. The present study represents the first empirical investigation of these hypotheses.

Hypotheses

The purpose of the present research is to understand how rater performance influences the leniency or severity with which raters evaluate others. As described previously, prior research has examined a number of rater factors which influence their rating patterns. A limited number of these studies have examined the role of supervisor performance as a predictor of the accuracy of supervisory performance ratings (e.g., Kirchner & Reisberg, 1962; Mandell, 1956). Theory suggests that there is cause to expect a negative relationship between rater performance and rater leniency. However, no studies have directly examined the impact of rater performance on leniency and severity in a large sample of raters. Thus, the primary hypothesis of the present study is as follows:

Hypothesis 1: Rater performance will be negatively associated with rater leniency.

Moreover, the present study will consider the impact that rater perspective has on the relationship between rater performance and leniency/severity. Recent research has demonstrated that raters with different relationships to the ratee (e.g., supervisor, peer, direct report) have unique perspectives on the ratee, and these rater perspectives impact rating behavior (Bynum et al., 2013; Hoffman et al., 2010). To date, no research

examining the impact of rater performance on ratings has considered how rater perspective might impact the relationship between rater performance and the leniency-severity of ratings they assign to others. There are a number of reasons to suspect that rater perspective might be relevant to this relationship. For example, if raters are using their own performance as a baseline for evaluating others, the degree to which the rater's job tasks and responsibilities align with those of the ratee may impact the saliency of the rater's performance when making judgments of the ratee's performance. As a result, the leniency or severity or ratings from peers may be more closely related to rater performance than for other types of raters. Moreover, social comparison theory suggests that individuals are more likely to draw comparisons between themselves and others based on their relationship within a social hierarchy. In particular, individuals draw comparisons with those who are most similar to them in order to make self-evaluations (Buunk & Gibbons, 2007; Festinger, 1954; Mumford, 1983; Shah, 1998). As a result, rater perspective is likely to influence the extent to which a rater compares their own performance to a ratee, which in turn would impact the strength of the relationship between rater performance and leniency/severity. This leads to the second hypothesis of the present study:

Hypothesis 2: The negative relationship between rater performance and rater leniency will be strongest for raters who are peers of their subjects.

Finally, as discussed earlier, researchers have examined the impact of gender on rater leniency/severity. Specifically, a relatively new subject of inquiry has been the interaction of rater and ratee gender on leniency/severity. Although several studies have suggested gender may play an important role, further research is required to corroborate

and clarify these initial findings—and no research has examined how gender similarity or dissimilarity may impact the rater performance – rater leniency relationship. As with rater perspective, it is possible that raters who are the same gender as their ratees may be more likely to draw comparisons between themselves and the ratee, thus strengthening the relationship between the rater's own performance and the leniency/severity with which he or she evaluates the ratee. As such, a third hypothesis will be addressed in the current study:

Hypothesis 3: The relationship between rater performance and rater leniency will be strongest when rater and subject are of the same gender.

The following section presents the procedures and analyses by which the above hypotheses will be tested.

Methods

Participants

The participants of the present study are full-time employees of a large, multi-national healthcare company who participated in a multi-rater (360) assessment between the years of 2014 and 2016. This includes employees who were the subject of the multi-rater assessment (hereafter referred to as "subjects") as well as employees who served as raters of the assessment (hereafter referred to as "raters"). The data includes 121,038 sets of ratings (not including self-ratings) evaluating 11,299 subjects. Many raters provided ratings on multiple subjects. 56.62% of raters provided ratings for only one subject, whereas 43.38% rated two or more subjects. The number of subjects rated by a single rater between 2014 and 2016 ranged from 1 to 29, with the mean number of subjects per rater being 2.1. Additionally, some participants were the subject of the 360 assessment multiple times between 2014 and 2016. During this time, 459 participants (4.06%) were the subject of the assessment two times, and three participants (<0.01%) participated as the subject of the assessment three times.

Subjects participated in the multi-rater assessment for a variety of reasons. Many of the subjects completed the multi-rater assessment as part of their participation in high potential development programs or manager training programs. Other ad hoc participants were administered the assessment as part of their individual development planning with their manager or as a self-initiated request for feedback.

Raters and subjects were diverse in gender, age, job level, function, and geographic location. Below, demographic characteristics of raters and subjects are reported. The rater sample, our primary sample of concern, is reported first, followed by

the subjects' demographic characteristics. The percentages below reflect the proportion of raters or subjects in each category for whom data was available; information regarding unknown demographic characteristics is included in Table 1.

The rater sample included 59,578 unique employees who served as raters. This represents approximately 47% of the total employee population of the organization. The sample was 46.2% female (*N*=19,674) and 53.8% male (*N*=22,897). Raters' ages ranged from 19 to 80, with a mean age of 41.4 years (SD=9.2). The sample of raters was 47.16% individual contributors (*N*=19,674), 34.58% managers (*N*=14,428), 13.5% directors (*N*=5,614), 2.04% vice presidents (*N*=849), and <0.01% executives (*N*=18). Raters worked in regions across the globe, including 22.92% from Asia Pacific (*N*=12,461), 33.62% from Europe, Middle East, and Africa (*N*=18,275), 8.53% from Latin America (*N*=4,635), and 34.93% from North America (*N*=18,986). Raters from EMEA were predominantly from European countries. Raters also represented all 16 functions within the organization. The number and percentage of subjects from each function are presented in Table 1.

The data set also included 11,299 unique subjects. This represents approximately 9% of the total employee population of the organization. The subjects of the multi-rater assessment were 44.91% female (*N*=4,317) and 55.09% male (*N*=5,296). Subjects' ages ranged from 22 to 71, and the mean subject age was 40.15 (SD=7.31). Subjects were 28.35% individual contributors (*N*=2,709), 54.02% managers (*N*=5,161), 15.44% directors (*N*=1,475), 1.82% vice presidents (*N*=206), and <0.01% executives (*N*=3). Geographically, 29.43% of subjects worked in Asia Pacific (*N*=2,555), 31.27% were from Europe, Middle East, and Africa (*N*=3,521), 8.86% from Latin America (*N*=998),

and 30.43% from North America (*N*=3,426). As with the rater sample, subjects from EMEA were overwhelmingly from European countries. The number and percentage of subjects from each function are presented in Table 2.

Measures

The data for the present study were obtained from a large, multinational health care company as well as a consulting firm working in partnership with the healthcare company. These data include the results of a multi-rater assessment as well as employee data obtained from the organization's HRIS. Below, key measures and variables included in the data set are described.

Multi-rater assessment. Data were obtained from a multi-rater developmental assessment administered between 2014 and 2016. The assessment was administered electronically. Subjects selected a number of raters to provide feedback on their performance. Raters included subjects' managers, peers, direct reports, business partners, and others. Each rater evaluated the subject on five dimensions of performance: (1) *relationships*, (2) *driving innovation*, (3) *learning and development*, (4) *leading others*, and (5) *ethical behavior*. These dimensions—aligned with the organization's competencies—are defined in Table 3, along with their underlying behaviors. All scales included 6 to 7 items. Although the dimensions measured were consistent across all participants and raters, some of the items varied according to the subject's level. Participants were assessed using either the individual leader form of the assessment (9.27%), the mid-level leader form (87.92%), or the senior leader form (2.81%). Items asked raters to identify the extent to which the subject demonstrated a specific behavior, with response options ranging from 1 ("Not at All") to 5 ("To a Very Great Extent").

The *relationships* dimension measures the extent to which employees build and maintain effective relationships with customers and coworkers. This scale includes 6 items. An example item from the *relationships* scale is "Promotes a culture of collaboration and teamwork across organizational boundaries." The *driving innovation* dimensions assesses how well subjects support or implement changes that add value to the organization. This scale includes 7 items. An example *driving innovation* item is "Encourages others to take appropriate risks, and helps them deal with the failure of well-reasoned ventures." The *learning and development* scale includes 7 items which measure the extent to which the employee engages in activities which contribute to his or her own development or the development of others. An example item from this scale is "Provides feedback, coaching, and guidance where appropriate to enhance others' skill development." The *leading others* scale has 7 items which assess the degree to which employees engage in effective management and leadership of others. An example item from *leading others* is "Delegates assignments to the lowest appropriate level." Finally, the *ethical behavior* scale includes 7 items targeting the extent to which employees behave according to established ethical guidelines and encourage others to do the same. An example item from the *ethical behavior* scale is "Demonstrates the courage to do what is right despite personal risk or discomfort."

Scores for each participant included an overall assessment score, scale scores, and individual item scores. Scale scores reflect the mean rating on all items within that scale, and overall assessment scores reflect the mean rating on all items in the assessment. For each of these (overall, scale, and individual item ratings), scores were reported to the participant as averages across all raters as well as within-rater group averages (e.g.,

supervisor, peer, direct report). In addition, data were also available from each individual rater for overall ratings, scale scores, item ratings, although these individual rater scores were not made available to the subject, with the exception of the subject's manager's ratings.

The ratings assigned by raters in the multi-rater assessment were anonymous; although subjects knew the identity of the raters and which rater groups they were assigned to, individual ratings were not attributable to specific raters. The sole exception to this is the subject's primary manager, whose ratings were identifiable.

Organizational Performance Ratings. For each rater and subject, official performance ratings were obtained from the organization's HRIS. Performance ratings at this organization range from 2-8, and are composed of two sub-scores. The first sub-score, ranging from 1-4, is the direct manager's rating of how well the employee achieved specific performance goals defined at the beginning of the annual performance management cycle. The second sub-score, also ranging from 1-4, is the direct manager's evaluation of how well the employee performed a set of performance behaviors defined by the organization's competency model, including the same scales defined above in the multi-rater assessment. Each rater and subject's overall performance rating is the sum of these two sub-scores. Organizational performance ratings were available for all three years (2014-2016), and as a result performance ratings were averaged across years where possible.

360 Performance Ratings. The sample includes 9,305 raters who were also *subjects* of the multi-rater assessment. 16.16% of all raters were also subjects of their own 360 assessment at some point between 2014 and 2016. On this subset of raters, two

additional performance metrics were used in analyses. First, the overall score on the rater's 360 was used as a secondary measure of rater performance. Secondly, the rater's *self-rating* on the 360 was used as a measure of the rater's self-perceived performance. Possible 360 performance ratings ranged from 1 to 5.

Rater Perspective. Each rater was categorized into a particular rater group for each subject they evaluated. These rater groups included primary manager, other manager, peer, direct report, business partner, and other. Of the 121,038 ratings in the data set (not including self-ratings), 11,604 (9.59%) were primary manager ratings, 8,286 (6.85%) were other manager ratings, 38,227 (31.58%) were peer ratings, 28,437 (23.49%) were direct report ratings, 25,936 (21.43%) were business partner ratings, and 8,548 (7.06%) were categorized as other raters. The primary manager refers to the subject's formal supervisor, whereas the other manager category refers to a "dotted line" reporting relationship which may or may not be formally established. Peers include team members or coworkers at a similar level to the subject. Direct reports include employees who are officially supervised by the subject. Business partners are typically employees of the organization who interact with the subject as an internal service provider—for example, an employee who acts as a subject's designated procurement partner. Other includes employees who have had opportunity to observe the employee's performance but did not fall into any of the other categories. Raters were assigned into categories by the subjects, who grouped the raters into categories when selecting raters. Participants were required to select their manager and peers as raters. As a result, 98.7% of participants received ratings from their primary manager and 97.3% received ratings from peers. The other rater perspectives were optional. 68.7% of participants received

evaluations from direct reports, 61.6% from business partners, 48.1% from secondary managers, and 25.8% from other raters. The total number of raters providing ratings from each perspective are presented in Table 4.

Gender and Gender Similarity. The gender of both raters and subjects were available via the organization's HR record system. Males were coded as 0 and females were coded as 1. Following the example of previous research, gender similarity / dissimilarity was coded using the absolute difference method, such that a value of 0 indicates gender similarity between rater and subject and a value of 1 indicates gender dissimilarity (Adeel & Pengcheng, 2016; Bauer & Green, 1996; Liden, Wayne, & Stilwell, 1993).

Demographic variables. As with gender, other demographic characteristics of each rater and subject were accessed through the organization's HRIS. These demographic variables include region, job profile grade, rater job level, function, organizational tenure, and age. Region includes North America, EMEA (Europe, Middle East, and Africa), Asia Pacific, and Latin America. Job profile grade corresponds to the employee's pay grade, which was used to code each rater and subject into one of four job levels: individual contributor, manager, director, and executive. Function includes 16 organizational units, including Sales, Research and Development, Operations, Marketing, Finance, Quality, Engineering, Information Technology, Regulatory Affairs, Public Affairs, Legal, General Administration, Strategic Planning, Facilities, Human Resources, and General Management.

Rater Leniency / Severity. Leniency / severity was measured by calculating the difference between each rater's mean rating of a participant across all items to the

estimated true score of that employee's performance, based on the overall ratings from all *other* raters (Sulsky & Balzer, 1988). The estimated true score has, in the past, been calculated by simply averaging the scores from other raters (e.g., Antonioni & Park, 2001; Kane et al., 1995). However, this method does not account for the finding that raters from different perspectives tend to vary in the relatively leniency/severity with which they evaluate others (Ng et al., 2011), and participants in this study varied in the number of raters from each perspective who provided ratings. For example, some participants had only two ratings from direct reports, whereas others had many more ratings from direct reports; given that direct reports, on average, provide higher ratings than other rater groups, this would erroneously inflate the estimated true score of performance for the latter participant. To control for the different composition of rater perspectives for each participant, ratings from each rater group (e.g., peer, direct report) were first averaged within the rater group, and then these rater group averages were once again averaged to form the estimated true score of performance. In this way, differences in the number of raters from each rater group did not impact the estimated true score.

Having established the participant's estimated true score, this estimated true score was then subtracted from each rater's average rating of the participant across all items to form a measure of that rater's leniency or severity. A resulting score of 0 indicates that the rater was neither more lenient nor more severe than the other raters. Positive leniency/severity scores indicate that a rater was more lenient than the other raters—the rater's evaluation of the participant was higher than the estimated true score. Negative leniency/severity scores indicate that a rater was more severe.

Analysis

Preliminary analyses. The first step of the analysis was to create the calculated variables—specifically, the creation of each rater's leniency / severity score. The procedures for this calculation were described in the previous section. Next, descriptive statistics for key variables were calculated. This includes the mean and standard deviation for all continuous variables, as well as a correlation matrix of all variables.

Next, analyses were conducted to address the first hypothesis regarding the relationship between rater performance and rater leniency / severity. Separate analyses were conducted using three measures of rater performance. The first set of analyses used the HRIS rating of rater performance (ranging from 2-8) as the primary predictor, averaged across all years between 2014 and 2016 for which performance ratings are available for each employee. These analyses were conducted on the full sample of 59,578 raters. A second and third set of analyses were conducted on raters who were also *subjects* of the multi-rater assessment. These analyses were identical, with the exception being that two alternate measures of rater's performance were used which are derived from the rater's 360 ratings. The second rater performance metric—the first to be obtained from the rater's 360 assessment—was the rater's *overall* score on the multi-rater assessment (average rating from all non-self raters). The final rater performance metric was the rater's *self-rating* from the 360. This metric was unique in that it isolated the rater's self-perception of performance rather than his or her "true" performance as indicated by ratings from other evaluators. These final two analyses were conducted on a subset of 11,299 raters.

The initial assumption was that hierarchical linear modeling would be used to examine the hypotheses. A mixed effects model was considered to be likely necessary because nearly half of the raters included in this data set had evaluated multiple subjects, with some raters evaluating as many as 29 different subjects. This violated the independence of observations assumption in linear multiple regression, which increased the likelihood of Type 1 error (Nezlek, 2008). As a result, it would be important to separately partition within-rater variance. A disaggregation approach (ignoring common raters and treating each individual rater-by-ratee case as an independent observation) would ignore the violation of independence, likely inflating Type 1 error—an assumption that would be tested prior to analysis. An aggregation approach (averaging effects for all subjects for each rater, and treating each individual as an individual case) would make it impossible to examine the effect of specific ratee characteristics on rater severity/leniency behavior, and would also reduce sample size. As such, hierarchical linear modeling was initially considered to be the most appropriate solution (Hofmann, 1997).

However, many researchers have expressed concern about using multilevel modeling techniques when sample size is small (e.g., Hox, 1998), and particularly when the number of Level 1 observations per Level 2 cluster is low. This was clearly the case with the data set used in this study— more than half (57%) of raters evaluated only one subject, and as a result were the only subject in their Level 2 cluster ("singletons"); the mean number of subjects per rater was just 2.1. Other research suggests this should not be a major concern for the present study. Maas and Hox (2005) found that, although standard errors may become biased when Level 2 sample size is low, there was no impact

of small cluster size on regression coefficients, variance components, or standard errors on data sets with large samples at Level 2. Similar results were found by Clarke & Wheaton (2007), who generally cautioned against the use of HLM with small cluster sizes, but found that the accuracy of estimated parameters was generally sufficient when number of groups exceeded 200, even with small cluster size and high proportions of singleton clusters. Both Bell, Ferron, and Kromrey (2008) and Theall et al. (2011) found that even data sets with very large proportions of singleton clusters (70-90%) can accurately estimate parameters with large samples at both Level 1 and Level 2.

Given the competing concerns related to repeated measures and small cluster size, the first step was to determine the optimal analytic approach for these data. Specifically, a test was conducted to determine whether or not hierarchical linear modeling was warranted and necessary given these data. If, in fact, Level 2 clusters (rater) were not accounting for substantial variation, multiple linear regression would be a suitable analytic technique in lieu of multilevel modeling. To test for the necessity of HLM, two statistics were calculated: the intra-class correlation (ICC) and the design effect. These were derived by conducting a random effects ANOVA, or null-model HLM, which partitioned variability in leniency/severity into within-group (Level 1) and between-group (Level 2) components. The variance estimates resulting from this analysis were then used to calculate an ICC. Likewise, these results were also used to calculate a design effect, reflecting the effect of independence violations on the standard error estimates. A non-zero ICC and a design effect greater than 2.0 were decided as evidence that multilevel modeling was necessary (Muthén & Satorra, 1995; Peugh, 2010). If either of these criteria were not met, a standard multiple regression analytic approach would be adopted.

In the event that the Level 2 rater variable accounted for a meaningful amount of variation in rater leniency/severity, a multilevel analytic technique would be adopted. Before specifying the model that would be tested, an estimation method must be selected. Given the large sample size, the difference between full information maximum likelihood estimation and restricted maximum likelihood estimation is likely to be negligible; as such, it was decided that REML would be used in order to ensure that degrees of freedom are appropriately allocated to compute variance estimates (Singer & Willett, 2003).

The model would be specified such that the outcome being predicted is rater leniency/severity. Because observations are nested within rater, Level 1 would focus on the rater-by-subject observations, whereas Level 2 would focus at the rater level across observations of multiple subjects. Level 1 variables would include our dependent variable—rater leniency/severity for a particular subject—as well as rater perspective (the rater's relationship to a particular subject). Level 2 variables would include rater performance, as well as the control variables: rater age, gender, organizational tenure, and region. As discussed previously, prior research has demonstrated that rater age (Griffeth & Bedeian, 1989), organizational tenure (Smither, Walker, & Yap, 2004), and culture (Adsit, London, Crom, & Jones, 1997; Li & Karakowsky, 2001; Ployhart, Weichmann, Schmitt, Sacco, & Rogg, 2003; Varma, Pichler, & Srinivas, 2005) have been found to be related to performance ratings assigned to others. The evidence for gender is more mixed, with numerous studies suggesting that rater gender—either through direct or interactive relationships—may also have an impact on leniency / severity (e.g., Binning, Adorno, & Williams, 1995; Pulakos et al., 1989). As a result, rater gender and ratee gender would also be included in the model. All Level 2 variables would be grand mean centered, and

all level 1 variables would be person-mean centered by rater (Woltman, Feldstain, MacKay, & Rocchi, 2012). Hypothesis 1 would be tested by examining the cross-level main effect of rater performance (Level 2) on rater leniency/severity (Level 1). Hypothesis 2 would be addressed by testing the moderating influence of rater perspective (Level 1) on the cross-level main of rater performance on leniency-severity. Hypothesis 3 would examine how Rater Gender (Level 2) and Ratee Gender (Level 1) independently and interactively influence the cross-level main effort of rater performance on leniency-severity. As described earlier, these analyses would be conducted with three separate measures of rater performance— HRIS performance ratings, raters' 360 scores, and raters' self-evaluations of performance. Rater 360 scores and rater self-evaluations of performance would leverage a subset of the data that includes only raters who have *also* participated as subjects in the 360.

If the ICC and design effect did not meet the threshold indicating that multilevel modeling was necessary, however, a multiple regression approach was determined to be the appropriate analytic method. In this event, three multiple regression models would be tested for each of the rater performance metrics. The first would test Hypothesis 1 by regressing rater leniency on rater performance and the control variables. The second model would expand on the first by adding rater perspective and the interaction between rater perspective and rater performance to the model, addressing Hypothesis 2. The final regression model would add gender similarity and the interaction of gender similarity and rater performance to the first model, addressing Hypothesis 3. These three regression analyses would be repeated for each of the three rater performance metrics—annual

performance ratings, 360 overall ratings, and 360 self-ratings—for a total of nine multiple regression analyses.

Results

Descriptive Statistics

Descriptive statistics, including mean, standard deviation, and intercorrelations for all non-categorical variables are presented in Table 5. The three rater performance metrics— annual performance ratings (HR), rater 360 ratings, and rater self-ratings— all demonstrated weak to moderate positive correlations with each other. Raters' annual performance ratings were correlated .27 with raters' 360 ratings and .03 with raters' self-ratings. Raters' 360 ratings were correlated .34 with raters' self-ratings. Rater performance was not strongly related with the raters' demographic characteristics; the largest relationship between rater performance and rater demographics was between raters' 360 ratings and rater age (r = -.12).

Rater leniency demonstrated no relationship or weak relationships with rater and subject demographic variables. Rater leniency demonstrated no relationship (r = .00) with subject demographic variables, including subject gender, subject age, and subject tenure. The correlation between rater leniency was also very low with rater demographic variables, including rater gender (r= 00), rater age (r = -.05), and rater tenure (r = -.04). There were also near-zero mean differences in leniency between rater-subject diads where males raters females, female rated males, and rater and ratee were the same gender.

Although zero-order correlations between rater leniency and rater performance were also small, the magnitude of this relationship varied by performance metric. Rater leniency was correlated .02 with raters' annual HR performance ratings, .08 with raters' 360 raters, and .17 with raters' self-ratings from the 360. Rater leniency's relationship

with rater self-ratings was the strongest relationship observed between rater leniency and any of the other studied variables. This relationship was in the opposite direction of what was hypothesized in Hypothesis 1, with raters with higher self-ratings of performance providing more lenient performance ratings to others.

Test for Multilevel Modeling

Next, a null model or unconditional model (random intercept only) was tested to examine the need for multilevel modeling. This model was defined as:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

where

Y_{ij} is the rater leniency for each rater's rating of a particular subject;

β_{0j} is the mean leniency for rater *j*;

r_{ij} is the Level 1 random error term;

γ_{00} is the average of raters' mean leniency scores;

u_{0j} is the random effect for rater *j*.

As described previously, both a non-zero ICC for same-rater leniency and a design effect greater than 2.0 would demonstrate the necessity for HLM. The results of the unconditional model yielded an ICC of 0.322 and a design effect of 1.35. As a result, the criteria for multilevel modeling was not met, and a multiple regression approach would be used to address the present hypotheses.

Three separate multiple regression models were tested, each identical except for the measure of rater performance—Annual HR Ratings, 360 Ratings, and Self-Ratings. These models will be referred to as Model 1, 2, and 3, respectively. For each of the three

models, the first analysis (1a/2a/3a) will test Hypothesis 1 by regressing rater leniency on rater performance and the rater control variables: rater age, rater region, and rater gender. The second analysis (1b/2b/3b) will test Hypothesis 2 by adding the direct effect of rater perspective as well as the interaction between rater performance and rater perspective to the model. Finally, the third analysis (1c/2c/3c) will test Hypothesis 3 by adding rater-subject gender agreement and the interaction between gender agreement and rater performance to the rater performance and control variables included in analyses 1a/2a/3a. For all three models, the incremental prediction of analyses b (rater perspective) and c (gender agreement) beyond analysis a (rater performance and controls) will be assessed to determine the value of these additions to the model.

Regression Diagnostics

Prior to conducting the regression analysis, diagnostics were run to ensure the data met the necessary assumptions for multiple linear regression. Diagnostic tests were conducted separately for all three regression models. Given the large sample size, graphical methods to test assumptions were adopted rather than the more conservative statistical tests (Das & Imon, 2016). For each of the three models, residuals were plotted against fitted values to ensure the relationships between predictors and outcome variables were approximately linear; in all three cases there was no cause for concern regarding the assumption of linearity. Histograms and Q-Q plots were examined for each model to confirm the normality of residuals. For all three models the distribution of residuals was approximately normal, although there was some indication of a slight degree of negative skewness (Figures 1-3). A scale-location plot was used to identify the presence of heteroskedasticity in the models; residuals appeared to spread consistently across the

range of fitted values, providing evidence of approximate homoskedasticity (Figures 1-3). In order to test for extreme outliers, a Bonferonni test was conducted to identify the most extreme outliers. A very small number of suspect entries were removed because the data seemed anomalous or illogical given the range of possible values. In nearly all cases, outliers and high leverage points were retained. Given the high correlation between rater age and rater tenure, tenure was dropped from the analysis. After dropping rater tenure, Durbin-Watson tests on all three models were non-significant, suggesting no major problems related to autocorrelation. Overall, diagnostic plots and other tests indicated no serious concerns regarding the assumptions underlying multiple linear regression.

Regression Results

Model 1: Annual HR Performance Ratings. Model 1 used raters' annual performance review ratings as the metric for rater performance. This analysis leveraged the full sample of 93,919 raters. Results for Models 1a (Hypothesis 1: Rater Performance), 1b (Hypothesis 2: Rater Perspective), and 1c (Hypothesis 3: Rater-Subject Gender Agreement) are presented in Table 6. In Model 1a, rater leniency was regressed on raters' annual performance ratings and the rater control variables: rater age, rater region, and rater gender. The R^2 for this model was just .002. In Model 1b, rater perspective and the interaction between rater perspective and rater annual performance ratings was added to the model. Although this led to a statistically significant improvement in variance explained above Model 1a (R^2=.009, ΔR^2=.006, F=100, p<.001), this model still accounted for a negligible amount of variance in rater leniency. The incremental prediction provided by Model 1b was primarily driven by the direct effect of rater perspective rather than the interaction of rater perspective and rater

performance. Model 1c included the same variables as 1a as well as the addition of rater-subject gender agreement and the interaction between rater-ratee gender agreement and rater annual performance ratings. This model did not improve the prediction of rater leniency beyond the variance explained by Model 1a (R^2=.003, ΔR^2<.001, F=1.05, p=.35). Overall, no model leveraging annual HR ratings of rater performance explained more than 1% of variance in rater leniency, and the inclusion of rater perspective and rater-subject gender agreement as direct predictors and moderators of the rater performance-rater leniency relationship did little to improve the model.

Model 2: Rater 360 Ratings. Model 2 included only those raters who were also subjects of their own 360 assessments, and used raters' 360 performance ratings as the metric for rater performance. This subset of the sample included 28,501 raters. Results for Models 2a (Hypothesis 1: Rater Performance), 2b (Hypothesis 2: Rater Perspective), and 2c (Hypothesis 3: Rater-Subject Gender Agreement) are presented in Table 7. In Model 2a, rater leniency was regressed on raters' 360 performance ratings and the rater control variables (age, region, and gender). The R^2 for this model was .008, marginally higher than Model 1a. In Model 2b, rater perspective and the interaction between rater perspective and raters' 360 ratings was added to the model. This led to a slight improvement in variance explained above Model 2a (R^2=.015, ΔR^2=.007, F=31.9, p<.001). As with Model 1b, the incremental prediction provided by Model 2b was primarily driven by the direct effect of rater perspective rather than the interaction between rater perspective and rater performance. Model 2c included the same variables as 2a as well as the addition of rater-subject gender agreement and the interaction between rater-ratee gender agreement and raters' 360 assessment scores. This model did

not improve the prediction of rater leniency beyond the variance explained by Model 2a (R^2=.009, Δ R^2=.001, F=2.13, p=.12).

Model 3: Rater Self-Ratings. Model 3 used raters' self-ratings from the 360. This analysis again leveraged the rater sub-sample including 28,504 raters who were also subjects of their own 360 performance assessment. Results for Models 3a (Hypothesis 1: Rater Performance), 3b (Hypothesis 2: Rater Perspective), and 3c (Hypothesis 3: Rater-Subject Gender Agreement) are presented in Table 8. In Model 3a, rater leniency was regressed on raters' self- ratings as well as rater age, region, and gender. The R^2 for this model was .032, substantially higher than Model 1a or 2a. This was largely driven by raters' self-ratings of performance (β=.168); rater age, region, and gender contributed little to the model. In Model 3b, rater perspective and the interaction between rater perspective and rater self-ratings were added to the model. This model had an R^2 of .039, an improvement of .007 beyond Model 1a. As with previous models (1b and 2b), the direct effect of rater perspective contributed most substantially to this incremental prediction, with the interaction between rater perspective and rater self-ratings playing a very minor role. Model 1c included the same variables as 3a as well as rater-subject gender agreement and the interaction between rater-subject gender agreement and rater self-ratings. This model improved the prediction of rater leniency above and beyond Model 1a by just .002 (R^2=.034, Δ R^2=.002, F=3.19, p=.041). Although still low, Models 3a-c (using rater self-ratings of performance as the metric for rater performance) were by far the most predictive models. This was largely driven by the direct effect of rater self-ratings of performance on rater leniency, with the interaction of rater perspective and

rater-subject gender agreement on the rater performance-rater leniency relationship playing a negligible role.

Discussion

The present study sought to investigate whether a rater's performance is associated with the leniency or severity with which that rater evaluates others. The hypothesis was that rater performance and rater leniency would be negatively associated, such that high performing raters would evaluate others more severely, whereas low performing raters would be more lenient in their evaluations of others. The findings did not support this hypothesis. Rater performance as judged by others (either via formal managerial performance ratings or through a multi-rater assessment) was largely unrelated to a rater's leniency or severity. Interestingly, when a rater evaluates his or her *own* performance, there does appear to be a very modest relationship between rater self-ratings of performance and rater leniency—in the opposite direction predicted. That is to say, raters who evaluated their own performance positively were more likely to be lenient in their appraisal of others, while raters who evaluated themselves more severely were also more severe in their ratings of others. This study also examined whether the relationship between rater performance and rater leniency is moderated by the rater's degree of similarity to the subject—both in terms of their relationship to the subject (peers vs. other rater groups) as well as gender (same vs. different gender between rater and subject). Neither of these moderators appeared to play a significant role.

An interesting finding from the present study is that raters' *self-ratings* of performance, but not raters' *other-ratings* of performance, were associated with rater leniency/severity. It was hypothesized that rater performance impacts rater leniency because raters will use their own performance as a reference point against which to appraise the effectiveness or ineffectiveness of others. The critical factor, then, is not a

rater's objective performance, but rather the rater's perception of his or her own performance. It is not a new or surprising finding that self- and other-ratings are not closely related to one another and demonstrate divergent relationships with other variables (e.g., Harris & Schaubroeck, 1988; Heidemeier & Moser, 2009). However, the fact that rater self-ratings of performance were associated with rater leniency/severity—whereas other-ratings were not— supports the notion that a rater's self-perception plays an important role in how a rater evaluates others. That said, there are additional plausible explanations for this relationship that need to be accounted for, as detailed later in this section.

A second unexpected outcome of this study is that the relationship between rater self-ratings of performance and rater leniency was in the opposite direction of what was hypothesized. Raters who appraised their own performance highly were more—not less—lenient in their evaluations of others relative to other raters. One plausible explanation for this finding concerns rater goals. Researchers have documented that rater goals influence the leniency/severity with which raters evaluate others (e.g., Murphy, Cleveland, Skattebo, & Kinney, 2004). It is possible—though not explicitly tested—that high performing employees' goals in a rating context may differ systematically and consistently from low performing raters' goals in such a way that high performing employees are more lenient. For example, high performers may be more concerned than low performers with promoting harmony among their colleagues and motivating coworkers to improve, both of which are associated with inflated performance ratings (Wang, Wong, & Kwong, 2010). Conversely, low performers may be motivated to be more severe in their ratings as a self-defensive tactic to maintain parity with their

colleagues. Further research might seek to confirm the positive relationship between self-ratings of performance and rater leniency and understand the underlying mechanism contributing to this relationship.

Limitations of the present study, however, might provide alternative explanations for the observed relationship between rater self-ratings and performance. One such limitation is that individual differences not accounted for in the present study may explain the positive relationship between rater performance and rater leniency. For example, extraversion is correlated (albeit weakly) with both rater leniency and job performance (Barrick & Mount, 1991; Harari et al., 2015). It is possible that the small relationship observed in the present study between rater performance and rater leniency may reflect shared variance with rater traits rather than any causal mechanism between rater self-perceptions of performance and rater leniency/severity. In order to parse the meaning of this positive relationship between rater performance and rater leniency, additional research is needed to control for individual difference variables and test possible mechanisms by which rater performance might be positively associated with rater leniency.

In addition to possible third variables, another limitation that must be considered is that common method bias may be contributing to the observed relationship between rater self-ratings and rater leniency (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In Model 3— the regression model in which the positive relationship between rater performance and rater leniency was detected—the rater him or herself is the source of both the rater performance metric as well as the leniency/severity variable, both of which were derived from the same multi-rater assessment tool. Therefore, Model 3 is the only

model in which it is possible that the observed relationship between rater performance and rater leniency could be attributed to a rater's general pattern of rating behavior rather than a relationship between performance and leniency. In other words, some raters may be consistently more or less lenient than others in their 360 ratings regardless of whether they are evaluating themselves or others (Kane et al., 1995). This potential inflation of the relationship between rater self-ratings and rater leniency was not of concern in testing the original hypothesis that the two variables would be negatively related. However, this does become problematic in drawing any conclusions about a *positive* relationship between rater self-ratings and rater leniency. Future studies might adopt designs or employ control variables that can disentangle variance attributable to the general rating pattern of the rater.

The primary conclusion drawn from the present research is that, despite relevant theory that suggested raters may consciously or subconsciously use their own performance as a point of reference when evaluating others, the results of this study did not provide support for the hypothesis that this leads high performing employees to be more severe or low performing employees to be more lenient when rating others. Variability in rater performance does not appear to be a major contributor to the stable patterns of rater leniency or severity found in performance ratings within organizations. Continued research is required to more fully understand what other rater characteristics might lead to the consistent patterns of leniency or severity demonstrated by raters—just one component of the broader ongoing effort to understand and address the flaws which persistently remain in organizational performance ratings.

References

Adeel, A., & Pengcheng, Z. (2016). Gender similarity and individual creativity as moderators of the relationship between informal leadership and leader-member-exchange: A longitudinal study. *European Journal of Business and Management, 8*(2), 90–102.

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology: Perspectives of Science and Practice, 9*(2), 219–252.

Adsit, D. J., London, M., Crom, S., & Jones, D. (1997). Cross-cultural differences in upward ratings in a multinational company. *International Journal of Human Resource Management, 8*(4), 385–401.

Aguinis, H. (2004). *Regression analysis for statistical moderators*. Guilford: New York, NY.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94–107.

Angkaw, A. C., Tran, G. Q., & Haaga, D. A. F. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behavior Research and Therapy, 44,* 533–544.

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management, 27*(4), 479–495.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluations in work settings. *Annual Review of Psychology, 49*, 141–168.

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.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other rating agreement: Does it really matter? *Personnel Psychology, 51,* 576–597.

Austin, J. T, & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*(6), 836–874.

Bamberger, P., & Biron, M. (2007). Group norms and excessive absenteeism: The role of peer referent others. *Organizational Behavior and Human Decision Processes, 103, 179*–196.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26.

Bartels, L. K., & Doverspike, D. (1997). Effects of disaggregation on managerial assessment center validity. *Journal of Business and Pyschology, 12*(1), 45–53.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Bauer, T. N., & Green, S. G. (1996). Development of leader-member exchange: A longitudinal test. *Academy of Management Journal, 39*(6), 1538–1567.

Beck, J. W., Beatty, A. S., & Sackett, P. R. (2014). On the distribution of job performance: The role of measurement characteristics in observed departures from normality. *Personnel Psychology, 67*(3)*,* 531–566.

Bedeian, A. G. (1976). Rater characteristics affecting the validity of performance appraisals. *Journal of Management, 2*(1), 37–45.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM Proceedings,* 1122–1129.

Bernardin, H. J. (1977). Behavioral expectation scales versus summates scales: A fairer comparison. *Journal of Applied Psychology, 62*(4), 422–427.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205–212.

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*(2), 197–211.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology, 61*(5), 564–570.

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A reconsideration of strategies for rater training. In G. L. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 18, pp. 221–274). Greenwich, CT: JAI Press.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology, 85*(2), 232–234.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. (1976). Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology, 61*(1), 75–79.

Bernardin, H. J., Thomason, S., Buckley, R. M., & Kane, J. S. (2016). Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management, 55*(2), 321–340.

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment, 17*(3), 300–310.

Binning, J. F., Adorno, A. J., & Williams, K. B. (1995, May). *Gender and race effects in an operational assessment center.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25,* 185–199.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48,* 587–605.

Bono, J. E., Hooper, A. C., & Yoon, D. J. (2012). Impact of rater personality on transformational and transactional leadership ratings. *The Leadership Quarterly, 23*(1), 132–145.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*(2), 238–252.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate performance: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*(3), 307–322.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. *Handbook of industrial and organizational psychology, Vol. 2*. (pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette and L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 2* (pp. 271–326). Palo Alta, CA: Consulting Psychologists Press.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology, 60*(5), 561–565.

Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-difference correlates. *Journal of Applied Psychology, 76*(1), 1–18.

Borman, W. C., & Vallon, W. R. (1974). A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology, 59,* 197–201.

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development, 18*(5), 417–435.

Burnaska, R. F., & Hollmann, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology, 59*(3), 307–312.

Buunk, A. P., & Gibbons, F. X. (2007). Social comparison: The end of a theory and the emergence of a field. *Organizational Behavior and Human Decision Processes, 102,* 3–21.

Bynum, B. H., Hoffman, B. J., Meade, A. W., & Gentry, W. A. (2013). Reconsidering the equivalence of multisource performance ratings: Evidence for the importance and meaning of rater factors. *Journal of Business and Psychology, 28*(2), 203–219.

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the twenty-first century. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology, Volume 1.* New York, NY: Oxford University Press.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57*(1), 15–22.

Catrambone, R., Beike, D., & Niedenthal, P. (1996). Is the self-concept a habitual referent in judgments of similarity? *Psychological Science, 7*(3), 158–163.

Chen, A. S., Bian, M., & Hou, Y. (2015). Impact of transformational leadership on subordinate's EI and work performance. *Personnel Review, 44*(4), 438–453.

Cho, I., & Payne, S. C. (2016). Other important questions: When, how, and why do cultural values influence performance management? *Industrial and Organizational Psychology: Perspectives of Science and Practice, 9,* 343–350.

Church, A. H. (1997). Managerial self-awareness in high performing individuals in organizations. *Journal of Applied Psychology, 82,* 281–292.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation. *Applied Psychological Measurement, 9,* 31–36.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological Methods and Research, 35*(3), 311–351.

Claus, L., & Briscoe, D. (2009). Employee performance management across borders: A review of relevant academic literature. *International Journal of Management Reviews, 11*(2), 175–196.

Cleveland, J. N., & Landy, F. J. (1981). The influence of rater and rate age on two performance judgments. *Personnel Psychology, 34*(1), 19–29.

Coen, T., & Jenkins, M. (2000). *Abolishing performance appraisals: Why they backfire and what to do instead.* New York: Berrett-Koehler.

Conner, D. S. (2003). Social comparison in virtual work environments: An examination of contemporary referent selection. *Journal of Occupational and Organizational Psychology, 76,* 133–147.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*(4), 331–360.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218–244.

Cozan, L. W. (1959). Forced choice: Better than other rating methods? *Personnel Psychology, 36,* 80–83.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*(3), 177–193.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Riley.

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5–12.

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology, 102*(3), 421–433.

Denisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*(3), 360–396.

DeVoe, S. E., & Iyengar, S. S. (2004). Managers' theories of subordinates: A cross-cultural examination of manager perceptions of motivation and appraisal of performance. *Organizational Behavior and Human Decision Processes, 93,* 47–61.

Dewberry, C., Davies-Muier, A., & Newell, S. (2013). Impact and causes of rater severity/leniency in appraisals without postevaluation communication between raters and ratees. *International Journal of Selection and Assessment, 21*(3), 286–293.

Dierdorff, E. C., Surface, E. A., & Brown, K. G. (2010). Frame-of-reference training effectiveness: Effects of goal orientation and self-efficacy on affective, cognitive, skill-based, and transfer outcomes. *Journal of Applied Psychology, 95,* 1181–1191.

Duarte, N. T., & Goodson, J. R., & Klich, J. R. (1994). Effects of dyadic quality and duration on performance appraisal. *The Academy of Management Journal, 37*(3), 499–521.

Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., & Ferris, G. R. (2012). A meta-analysis of antecedents and consequences of leader-member exchange: Integrating the past with an eye toward the future. *Journal of Management, 38*(6), 1715–1759.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111*(1), 3–22.

Eichinger, R. W., & Lombardo, M. M. (2004). Patterns of rater accuracy in 360-degree feedback. *Human Resource Planning, 27*(4), 23.

Ellett, C. D., Wren, C. Y., Callender, K. E., Loup, K. S., & Liu, X. (1996). Looking backwards with the Personnel Evaluation Standards: An analysis of the development and implementation of a statewide teacher assessment program. *Studies in Educational Evaluation, 22,* 79–113.

Ellis, A. P. J., Ilgen, D. R., & Hollenbeck, J. R. (2006). The effects of team leader race on performance evaluations, *37*(3), 295–322.

Farh, J.-L., & Werbel, J. D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. *Journal of Applied Psychology, 71*(3), 527–529.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127–148.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

Fisher, C. D. (1989). Current and recurrent challenges in HRM. *Journal of Management, 1*(5), 157–180.

Fletcher, C. (1999). The implications of research on gender differences in self-assessment and 360 degree appraisal. *Human Resource Management Journal, 9*(1), 39–46.

Ford, A. (1931). Neutralizing inequalities in rating. *Personnel Journal, 9,* 466–469.

Furnham, A., & Stringfield, P. (1998). Congruence in job-performance ratings: A study of 360 feedback examining self, manager, peers, and consultant ratings. *Human Relations, 51*(4), 517–530.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral

observation scale and the relative percentile method. *Journal of Business and Psychology, 11*(1), 23–33.

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*(2), 251–268.

Goodman, P. S. (1974). An examination of referents used in the evaluation of pay. *Organizational Behavior and Human Performance, 12,* 170–195.

Goodman, P. S., & Haisley, E. (2007). Social comparison processes in an organizational context: New directions. *Organizational Behavior and Human Decision Processes, 102,* 109–125.

Gordon, R. A., & Arvey, R. D. (2004). Age bias in laboratory and field setting: A meta-analytic investigation. *Journal of Applied Social Psychology, 34*(3), 468–492.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336–1344.

Griffeth, R. W., & Bedeian, A. G. (1989). Employee performance evaluations: Effects of ratee age, rater age, and ratee gender. *Journal of Organizational Behavior, 10,* 83–90.

Gruenfeld, L. W., & Weissenberg, P. (1966). Supervisory characteristics and attitudes toward performance appraisals. *Personnel Psychology, 19*(2), 143–151.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw Hill.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392–407.

Harari, M. B., Rudolph, C. W., & Laginess, A. J. (2015). Does rater personality matter? A meta-analysis of Big Five-performance rating relationships. *Journal of Occupational and Organizational Psychology, 88*(2), 387–414.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*(4), 737–756.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43–62.

Härtel, C. E. J. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. *Journal of Applied Psychology, 78*(2), 212–217.

Hauenstein, N. M. A. (1992). An information-processing approach to leniency in performance judgments. *Journal of Applied Psychology, 77*(4), 485–493.

Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes, 50,* 300–323.

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*(1), 68–73.

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*(2), 353–370.

Heneman, R. L., Moore, M. L., & Wexley, K. N. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research, 15*(5), 431–448.

Hoffman, B. J., & Woehr, D. J. (2009). Disentangling the meaning of multisource performance rating source and dimension factors. *Personnel Psychology, 62*(4), 735–765.

Hoffman, B. J., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*(1), 119–151.

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. *Personnel Psychology, 44*(3), 601–618.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management, 23,* 723–744.

Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of International Business Studies, 14,* 75–89.

Hofstede, G. (1992). Cultural dimensions in people management: The socialization perspective. In V. Pucik, N. M. Tichy, & C. K. Barnett (Eds.), *Globalizing*

management: Creating and leading the competitive organization (pp. 139–158). New York, NY: Wiley.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology, 63*(5), 579–588.

Houston, W. M., Raymond, M. R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15*(4), 409–421.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York: Springer Verlag.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64–86.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403–424.

Hu, J., & Liden, R. C. (2013). Relative leader-member exchange within team contexts: How and when social comparison impacts individual effectiveness. *Personnel Psychology, 66,* 127–172.

Hu, H. H., Hsu, C. T., Lee, W. R., & Chu, C. M. (2007). A policy-capturing approach to comparing the reward allocation decisions of Taiwanese and U.S. managers. *Social Behavior and Personality, 35,* 1235–1250.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. *Research in Organizational Behavior, 5,* 141–197.

Jackson, D. J. R., Atkins, S. G., & Fletcher, R. B. (2005). Frame of reference training for assessment centers: Effects on interrater reliability when rating behavior and ability traits. *Public Personnel Management, 34,* 17–30.

Jawahar, I. M., & Stone, T. H. (1997). Influence of raters' self-consciousness and appraisal purpose on leniency and accuracy of performance ratings. *Psychological Reports, 80*(1), 323–336.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*(4), 905–925.

Jewell, L. N. (1998). *Contemporary Industrial/Organizational Psychology* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research.* New York, NY: Guilford Press.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology, 89*(5), 755–768.

Kacmar, K. M., Witt, L. A., Zivnuska, S., & Gully, S. M. (2003). The interactive effect of leader-member exchange and communication frequency on performance ratings. *Journal of Applied Psychology, 88*(4), 764–772.

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38*(4), 1036–1051.

Karylowski, J. J. (1990). Social reference points and accessibility of trait-related information in self-other similarity judgments. *Journal of Personality and Social Psychology, 58,* 975–983.

Kelly, G. A. (1955). *The psychology of personal constructs.* New York: Norton.

Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1,* 377–383.

Kirchner, W. K., & Reisberg, D. J. (1962). Differences between better and less-effective supervisors in appraisal of subordinates. *Personnel Psychology, 15*(3), 295–302.

Kneeland, N. (1929). That lenient tendency in rating. *Personnel Journal, 7,* 356–366.

Kovacs, R., & Kapel, D. E. (1976). Personality correlates of faculty and course evaluations. *Research in Higher Education, 5*(4), 335–344.

Kozlowski, S. W. J., & Kirsch, M. P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology, 72*(2), 252–261.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70*(1), 56–65.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kruglanski, A. W., & Mayseless, O. (1990). Classic and current social comparison research: Expanding the perspective. *Psychological Bulletin, 108,* 195–208.

Lance, C. E., & Woehr, D. J. (1986). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal of Applied Psychology, 71*(4), 679–685.

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*(4), 223–232.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72–107.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 61, 750–758.

Latham, G. P. (1986). Job performance and appraisal. In C. L. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organizational Psychology.* London: Wiley.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*(5), 550–555.

Lawshe, C. H., Kephart, N. C., & McCormick, E. J. (1949). The paired comparison technique for rating performance of industrial employees. *Journal of Applied Psychology, 33*(1), 69–77.

Lepkowski, J. R. (1963). Development of a forced-choice rating scale for engineer evaluation. *Journal of Applied Psychology, 47*(2), 87–88.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*(6), 881–905.

Levy, S., & Stone, D. M. (1963). *Process and content of managerial ratings of subordinates*. Paper presented at meeting of Eastern Psychological Association: New York, NY.

Li, J., & Karakowsky, L. (2001). Do we see eye-to-eye? Implications of cultural differences for cross-cultural management research and practice. *Journal of Psychology, 135*(5), 501–517.

Lichtenstein, M., & Srull, T. K. (1987). Processing objectives as a determinant of the relationship between recall and judgment. *Journal of Experimental Social Psychology, 23,* 93–118.

Liden, R. C., Stilwell, D., & Ferris, G. R. (1996). The effects of supervisor and subordinate age on objective performance and subjective performance ratings. *Human Relations, 49*(3), 327–347.

Liden, R. C., Wayne, S. J., & Stilwell, D. (1993). A longitudinal study on the early development of leader-member exchanges. *Journal of Applied Psychology, 78*(4), 662–674.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10–13.

Long, K. M., & Spears, R. (1998). Opposing effects of personal and collective self-esteem on interpersonal and intergroup comparisons. *European Journal of Social Psychology, 28*(6), 913–930.

Lundstrom, J. T. (2008). *A new use of frame-of-reference training: Improving reviewers' inferences from biodata information.* Unpublished doctoral dissertation). Kansas State University, Manhattan, KS.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

Major, B., & Forcey, B. (1985). Social comparisons and pay evaluations: Preferences for same-sex and same-job wage comparisons. *Journal of Experimental Social Psychology, 21,* 393–405.

Mandell, M. M. (1956). Supervisory characteristics and ratings: A summary of recent research. *Personnel, 32,* 435–440.

McEvoy, G. M., & Cascio, W. F. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology, 74*(1), 11–17.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and purpose of rating. *Journal of Applied Psychology, 69,* 147–156.

McManus, I. C., Thomposn, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-faces Rasch modeling. *BMC Medical Education, 6*(42), 1272–1294.

Melchers, K. G., Lienhardt, N., von Aarburg, M., & Kleinmann, M. (2011). Is more structure always better? An evaluation of the effects of rater training and

descriptively anchored rating scales on rating accuracy in a structured interview. *Personnel Psychology, 64,* 53–87.

Mercer (2013). Global performance management survey report: Executive summary.

Millmore, M., Biggs, D., & Morse, L. (2007). Gender differences within 360-degree managerial performance appraisals. *Women in Management Review, 22*(7), 536–551.

Mishra, V., & Roch, S. G. (2013). Cultural values and performance appraisal: Assessing the effects of rater self-construal on performance ratings. *Journal of Psychology, 147,* 325–344.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51*(3), 557–576.

Mount, M. K., Sytsma, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater-ratee race effects in developmental performance ratings of managers. *Personnel Psychology, 50,* 51–69.

Mumford, M. D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. *Personnel Psychology, 36*(4), 867–881.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives of Science and Practice, 1,* 148–160.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619–624.

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

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

Murphy, K. R., & Constans, J. I. (1988). Psychological issues in scale format research: Behavioral anchors as a source of bias in rating. In R. L. Cardy, S. M. Puffer, & J. M. Newman (Eds.), *Advances in information processing in organizations* (Vol. 3, pp. 135–153). Greenwich, CT: JAI Press.

Murphy, K. R., & Deshon, R. (2000). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873–900.

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*(1), 158–164.

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

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology, 41,* 517–535.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*(2), 842–860.

Ng, T. W. H., & Feldman, D. C. (2008). The relationship of age to ten dimensions of job performance. *Journal of Applied Psychology, 93*(2), 392–423.

Ng., K.-Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K.-Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology, 96*(5), 1033–1044.

O'Neill, T. A., McLarnon, M. J. W., & Carswell, J. J. (2015). Variance components of job performance ratings. *Human Performance, 28*(1), 66–91.

Oldham, G. R., Kulik, C. T., Stepina, L. P., & Ambrose, M. L. (1986). Relations between situational factors and the comparative referents used by employees. *Academy of Management Journal, 29*(3), 599–608.

Oldham, G. R., Nottenburg, G., Kassner, M. W., Ferris, G., Fedor, D., & Masters, M. (1982). The selection and consequences of job comparisons. *Organizational Behavior and Human Performance, 29*, 84–111.

Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology, 1992, 77*(2), 201–217.

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology, 57,* 333–375.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. (1959). Rating scale content: III. Relationship between supervisory- and self-ratings. *Personnel Psychology, 12,* 49–63.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48,* 85–112.

Phillips, B. N. (1960). Authoritarian, hostile, and anxious students' ratings of an instructor. *California Journal of Educational Research, 11*(1), 19–23.

Pichler, S. (2012). The social context of performance appraisal and appraisal reactions: A meta-analysis. *Human Resource Management, 51*(5), 709–732.

Ployhart, R. E., Weichmann, D., Schmitt, N., Sacco, J. M., & Rogg, K. (2003). The cross-cultural equivalence of job performance ratings. *Human Performance, 16,* 49–79.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.

Prien, E. P., & Liske, R. E. (1962). Assessment of high-level personnel: III. Rating criteria: A comparative analysis of supervisory ratings and incumbent self-ratings of job performance. *Personnel Psychology, 15,* 187–194.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*(4), 581–588.

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*(5), 770–780.

Rand, T. M., & Wexley, K. N. (1975). Demonstration of the effect, "similar to me," in simulated employment interviews. *Psychological Reports, 36*(2), 535–544.

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. *Applied Psychological Measurement, 35*(3), 235–246.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85,* 370–395.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloys (Eds.), *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*, 537–560.

Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management, 38*(2), 719–739.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428.

Sackett, P. R., & DuBois, C. L. Z. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76*(6), 873–877.

Sauer, S. J. (2011). Taking the reins: The effects of new leader status and leadership style on team performance. *Journal of Applied Psychology, 96*(3), 574–587.

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73,* 76–91.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame of reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87,* 735–746.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901–912.

Schneider, D. E., & Bayroff, A. G. (1953). The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology, 37*(4), 278–280.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956–970.

Shah, P. P. (1998). Who are employees' social referents? Using a network perspective to determine referent others. *Academy of Management Journal, 41*(3), 249–268.

Sharon, A. T., & Bartlett, C. J. (1969). Effects of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 22, 251–263.

Shipper, F., & Dillard, J. (2000). A study of impending derailment and recovery of middle managers across career stages. *Human Resource Management, 39*(3), 331–347.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*(1), 22–40.

Smither, J. W. (2012). Performance management. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology, Volume 1.* New York, NY: Oxford University Press.

Smither, J. W., & Reilly, R. R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 30,* 369–391.

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An examination of the equivalence of web-baed versus paper-and-pencil upward feedback ratings: Rater- and ratee-level analyses. *Educational and Psychological Measurement, 64*(1), 40–61.

Stark, E., & Poppler, P. (2008). Leadership, performance evaluations, and all the usual suspects. *Personnel Review, 38*(3), 320–338.

Stauffer, J. M., & Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology, 90*(3), 586–591.

Stevens, S. E. (1995). *The effect of frame-of-reference rater training on idiosyncratic and normative raters.* (Unpublished doctoral dissertation). University of Akron, Akron, OH.

Sturman, M. C. (2003). Searching for the inverted u-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29*(5), 609–640.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*(3), 497–506.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*(4), 501–510.

Sweeney, P. D., & McFarlin, D. B. (2005). Wage comparisons with similar and dissimilar others. *Journal of Occupational and Organizational Psychology, 78,* 113–131.

Taylor, E. K., & Hastman, R. (1956). Relation of format and administration to the characteristics of graphic rating scales. *Personnel Psychology, 9*(2), 181–206.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4*(1), 39–47.

Taylor, E. K., Schneider, D. E., & Clay, H. (1954). Short forced-choice ratings work. *Personnel Psychology, 7*(2), 245–252.

Theall, K. P., Scribner, R., Broyles, S., Yu, Q., Chotalia, J., Simonsen, N., Schonlau, M., & Carlin, B. P. (2011). Impact of small group size on neighborhood influences in multilevel models. *Journal of Epidemiology and Community Health, 27,* 688–695.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Thornton, D. A., & Arrowood, A. J. (1966). Self-evaluation, self-enhancement, and the locus of social comparison. *Journal of Experimental Social Psychology Supplement, 1,* 40–48.

Thornton, G. C. (1968). The relationship between supervisory and self-appraisals of executive performance. *Personnel Psychology, 21,* 441–455.

Thornton, G. C., & Zorich, S. (1980) Training to improve observer accuracy. *Journal of Applied Psychology, 65*(3), 351–354.

Tornow, W. W., & London, M. (1998). *Maximizing the value of 360-degree feedback: A process for successful individual and organizational development*. San Francisco, CA: Jossey-Bass Publishers.

Tziner, A., & Dolan, S. (1984). The relationship of two sociodemographic variables and perceived climate dimensions to performance. *Canadian Journal of Administrative Sciences, 1,* 272–287.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group & Organization Management, 30*(1), 89–98.

Tziner, A., Murphy, K. R., Cleveland, J. N., Beaudin, G., & Marchand, S. (1998). Impact of rater beliefs regarding performance appraisal and its organizational context on appraisal quality. *Journal of Business and Psychology, 12*(4), 457–467.

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*(3), 711–719.

Van der Deijden, B. I. J. M., Scholarios, D., Van der Schoot, E., Jedrzejowicz, P., Bozionelos, N., Epitropaki, O., Knauth, P., Marzec, I., Mikkelsen, A., Van der Heijde, C. (2010). Supervisor-subordinate age dissimilarity and performance ratings: The buffering effects of supervisory relationship and practice. *International Journal of Aging and Human Development, 71*(3), 231–258.

Varma, A., & Stroh, L. K. (2001). The impact of same-sex LMX dyads on performance evaluations. *Human Resource Management, 40*(4), 309–320.

Varma, A., Pichler, S., & Srinivas, E. S. (2005). The role of interpersonal affect in performance appraisal: Evidence from two samples—the US and India. *International Journal of Human Resource Management, 11,* 2029–2044.

Vecchio, R. P. (1993). The impact of differences in subordinate and supervisor age on attitudes and performance. *Psychology and Aging, 8*(1), 112–129.

Vecchio, R. P., & Gobdel, B. (1984). The vertical dyad linkage model of leadership: Problems and prospects. *Organizational Behavior and Human Performance, 34,* 5–20.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108–131.

Waldman, D. A., & Avolio, B. J. (1986). A meta-analysis of age differences in job performance. *Journal of Applied Psychology, 71*(1), 33–38.

Wang, G., Oh, I.-S., Courtright, S. H., & Colbert, A. E. (2011). Transformational leadership and performance across criteria and levels: A meta-analytic review of 25 years of research. *Group & Organization Management, 36*(2), 223–270.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95*(3), 546–561.

Wells, F. L. (1907). *A statistical study of literary merit: With remarks on some new phases of the method* (Vol. 13). Science Press.

Wexley, K. N., & Youtz, M. A. (1985). Rater beliefs about others: Their effects on rating errors and rating accuracy. *Journal of Occupational Psychology, 58*(4), 265–275.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1972). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 45–48.

Wildman, J. L., Bedwell, W. L., Salas, E., & Smith-Jentsch, K. A. (2010). Performance measurement: Individual, team, and organizational strategies. In S. Zedeck (Ed.), *APA Handbook of Industrial and Organizational Psychology, Vol. 1: Building and developing the organization* (pp. 303–341). Washington, DC: American Psychological Association.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79*(4), 525–534.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189–205.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52–69.

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin, 106*(2), 231–248.

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97–107.

Zhou, J., & Martocchio, J. J. (2001). Chinese and American managers' compensation award decisions: A policy-capturing comparative study. *Personnel Psychology, 54,* 115–145.

Tables

Table 1.

Rater Demographic Characteristics

Variable	#	%
Total Unique Raters	59,578	100.00%
Gender		
Female	19,674	32.21%
Male	22,897	36.61%
Unknown	17,007	31.18%
Level		
Individual Contributor	20,810	35.19%
Manager	14,428	22.43%
Director	5,614	8.50%
Vice President	849	1.29%
Executive	18	0.03%
Unknown	17,859	32.57%
Region		
Asia Pacific	12,461	20.69%
Europe, Middle East and Africa	18,275	29.99%
Latin America	4,635	8.05%
North America	18,986	29.88%
Unknown	5,221	11.39%
Function		
Engineering	2,598	4.15%
Facilities	133	0.21%
Finance	2,943	4.88%
General Administration	1,282	2.15%
General Management	173	0.26%
Human Resources	172	0.35%
Information Technology	2,132	3.38%
Legal	605	0.94%
Marketing	3,696	6.07%
Operations	6,377	10.09%
Public Affairs	799	1.26%
Quality	2,866	4.56%
Research and Development	6,818	11.14%
Regulatory Affairs	1,102	1.77%
Sales	10,442	16.90%
Strategic Planning	354	0.55%
Unknown	17,086	31.33%

Table 2

Participant Demographic Characteristics

Variable	#	%
Total Unique Participants	11,299	100.00%
Gender		
Female	4,317	38.21%
Male	5,296	46.87%
Unknown	1,686	14.92%
Level		
Individual Contributor	2,709	23.98%
Manager	5,161	45.68%
Director	1,475	13.05%
Vice President	206	1.82%
Executive	3	<0.01%
Unknown	11,299	15.44%
Region		
Asia Pacific	3,314	29.33%
Europe, Middle East and Africa	3,521	31.16%
Latin America	998	8.83%
North America	3,426	30.32%
Unknown	40	<0.01%
Function		
Engineering	544	4.81%
Facilities	36	0.32%
Finance	749	6.63%
General Administration	135	1.19%
General Management	52	0.46%
Human Resources	33	0.29%
Information Technology	521	4.61%
Legal	142	1.26%
Marketing	1,086	9.61%
Operations	1,311	11.60%
Public Affairs	202	1.79%
Quality	595	5.27%
Research and Development	1,562	13.82%
Regulatory Affairs	253	2.24%
Sales	2,291	20.28%
Strategic Planning	103	0.91%
Unknown	1,684	14.90%

Table 3

Competency definitions and behaviors

Label	Definition	Behaviors
Ethical Behavior	Demonstrate and inspire the behaviors that reinforce our values	• Serve as a role model for making value-based decisions • Create a trusting, collaborative, and ethical work environment • Maintain the highest standards of quality, compliance and accountability • Champion programs and initiatives that support our environment and communities
Relationships	Develop deep insights into the needs of our patients, customers, markets and communities	• Cultivate external relationships and partnerships • Be insight-driven to uncover unmet needs • Forge internal collaboration across all levels of the enterprise
Driving Innovation	Drive innovation; anticipate and shape industry and market changes to advance health care globally	• Translate insights into viable products and solutions that create value • Challenge the status quo; lead and adapt to change • Take and manage risks
Leading Others	Deliver results by inspiring and mobilizing people and teams	• Empower people to act with speed, agility, and accountability • Demonstrate a global and enterprise-wide mindset • Balance short and long-term strategic choices
Learning and Development	Create an environment where leadership and talent development is top priority	• Take ownership for talent acquisition, performance and development of self and others • Maximize the power of diversity and inclusion • Engage in transparent and constructive conversations

Note. Competency labels were modified at the organization's request.

Table 4

Rater Perspectives

Rater Category	#
All	132,802
Self	11,764
Primary Manager	11,604
Other Manager	8,286
Direct Report	28,437
Peer	38,227
Business Partner	25,936
Other	8,548

Table 5

Descriptive statistics and intercorrelations for full rater sample

	N	X	SD	1	2	3	4	5	6	7	8	9	10
Rater Leniency / Performance													
1. Rater Leniency	121,015	.00	.50	1.00	.02	.17	.08	-.05	-.04	.00	.00	.00	.00
2. Rater Performance (HR)	93,974	6.34	.71	.02	1.00	.03	.27	-.09	-.02	.02	.00	.01	.02
3. Rater Self-Rating (360)	32,225	3.95	.47	.17	.03	1.00	.34	-.01	-.05	.01	-.06	-.07	.03
4. Rater Overall (360)	32,222	4.06	.31	.08	.27	.34	1.00	-.12	-.01	.06	-.06	-.03	.05
Rater Characteristics													
5. Rater Age	97,969	42.21	9.05	-.05	-.09	-.01	-.12	1.00	.57	-.11	.28	.17	-.03
6. Rater Tenure	97,969	10.86	7.60	-.04	-.02	-.05	-.01	.57	1.00	-.08	.18	1.00	-.03
7. Rater Gender (1=F)	97,583			.00	.02	.01	.06	-.11	-.08	1.00	-.03	-.04	.16
Subject Characteristics													
8. Subject Age	103,928	40.43	7.31	.00	.00	-.06	-.06	.28	.18	-.03	1.00	.47	-.07
9. Subject Tenure	103,928	9.44	6.27	.00	.01	-.07	-.03	.17	.23	-.04	.47	1.00	-.03
10. Subject Gender (1=F)	103,851			.00	.02	.03	.05	-.03	-.03	.16	-.07	-.03	1.00

Table 6

Multiple regression results for Model 1 (Annual Performance Ratings)

	1a: Rater Performance		1b: Rater-Subject Relationship		1c: Gender Agreement	
Variable	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	.214***	.017	.264***	.018	.217***	.018
Rater Performance (Annual HR)	.016***	.003	.046***	.006	.013**	.004
Rater Age	-.005***	<.001	-.004***	<.001	-.005***	<.001
Rater Region (vs. ASPAC)						
EMEA	-.022*	.009	-.024**	.009	-.020*	.009
LATAM	-.009	.013	-.011	.013	-.012	.013
NORAM	.000	.009	-.010	.009	.001	.009
Rater Gender (1=F)	-.015*	.006	-.025***	.007	-.009	.007
Rater Perspective (vs. Direct Report)						
Manager			-.216***	.011		
Peer			-.152***	.009		
Others			-.046***	.009		
Rater Perspective*Rater Performance						
Manager*Self-Rating			-.022*	011		
Peer*Self-Rating			-.044***	.008		
Other*Self-Rating			-.026**	.009		
Gender Agreement (1=diff. gender)					-.024***	.007
Gender Agreement*Self-Rating					.005	.007
R^2		.002***		.009***		.003***
ΔR^2 vs. 1a				.006***		<.001

Note. *N*=93,919.
*p<.05 **p<.01 ***p<.001

Table 7

Multiple regression results for Model 2 (360 Performance Ratings)

	2a: Rater Performance		2b: Rater-Subject Relationship		2c: Gender Agreement	
Variable	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	.023	.035	.101**	.038	.001	.037
Rater Performance (360)	.081***	.006	.126***	.015	.083***	.008
Rater Age	-.003**	.001	-.001	.001	-.002*	.001
Rater Region (vs. ASPAC)						
EMEA	.009	.014	.009	.014	.013	.015
LATAM	-.010	.020	-.017	.020	-.021	.021
NORAM	-.034*	.015	-.045**	.015	-.044**	.016
Rater Gender (1=F)	-.012	.011	-.018	.011	-.018	.012
Rater Perspective (vs. Direct Report)						
Manager			-.200***	.020		
Peer			-.157***	.018		
Others			-.072***	.019		
Rater Perspective*Rater Performance						
Manager*Self-Rating			-.112***	.020		
Peer*Self-Rating			-.034	.018		
Other*Self-Rating			-.014	.019		
Gender Agreement (1=diff. gender)					-.026*	.012
Gender Agreement*Self-Rating					.011	.012
R^2		.008***		.015***		.009***
ΔR^2 vs. 3a				.007***		.001

Note. *N*=28,501.
*p<.05 **p<.01 ***p<.001

Table 8

Multiple regression results for Model 3 (Self Performance Ratings)

	3a: Rater Performance		3b: Rater-Subject Relationship		3c: Gender Agreement	
Variable	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	.081*	.035	0.160***	.037	.072	.037
Rater Performance (Self)	.168***	.006	0.200***	.015	.161***	.007
Rater Age	-.004***	.001	-.002**	<.001	-.003***	.001
Rater Region (vs. ASPAC)						
EMEA	.003	.014	.003	.014	.005	.015
LATAM	-.060**	.194	-.066***	.019	-.072***	.021
NORAM	-.038*	.015	-.046**	.015	-.048**	.016
Rater Gender (1=F)	-.006	.011	-.012	.011	-.011	.012
Rater Perspective (vs. Direct Report)						
Manager			-.207***	.020		
Peer			-.153***	.018		
Others			-.073***	.019		
Rater Perspective*Rater Performance						
Manager*Self-Rating			-.114***	.019		
Peer*Self-Rating			-.006	.017		
Other*Self-Rating			0.018	.019		
Gender Agreement (1=diff. gender)					-.032**	.012
Gender Agreement*Self-Rating					.023	.012
R^2		.032		.039		.034
ΔR^2 vs. 3a				.007***		.002*

Note. *N*=28,504.
*p<.05 **p<.01 ***p<.001

Figures

Figure 1. Residual plots checking regression assumptions for Model 1

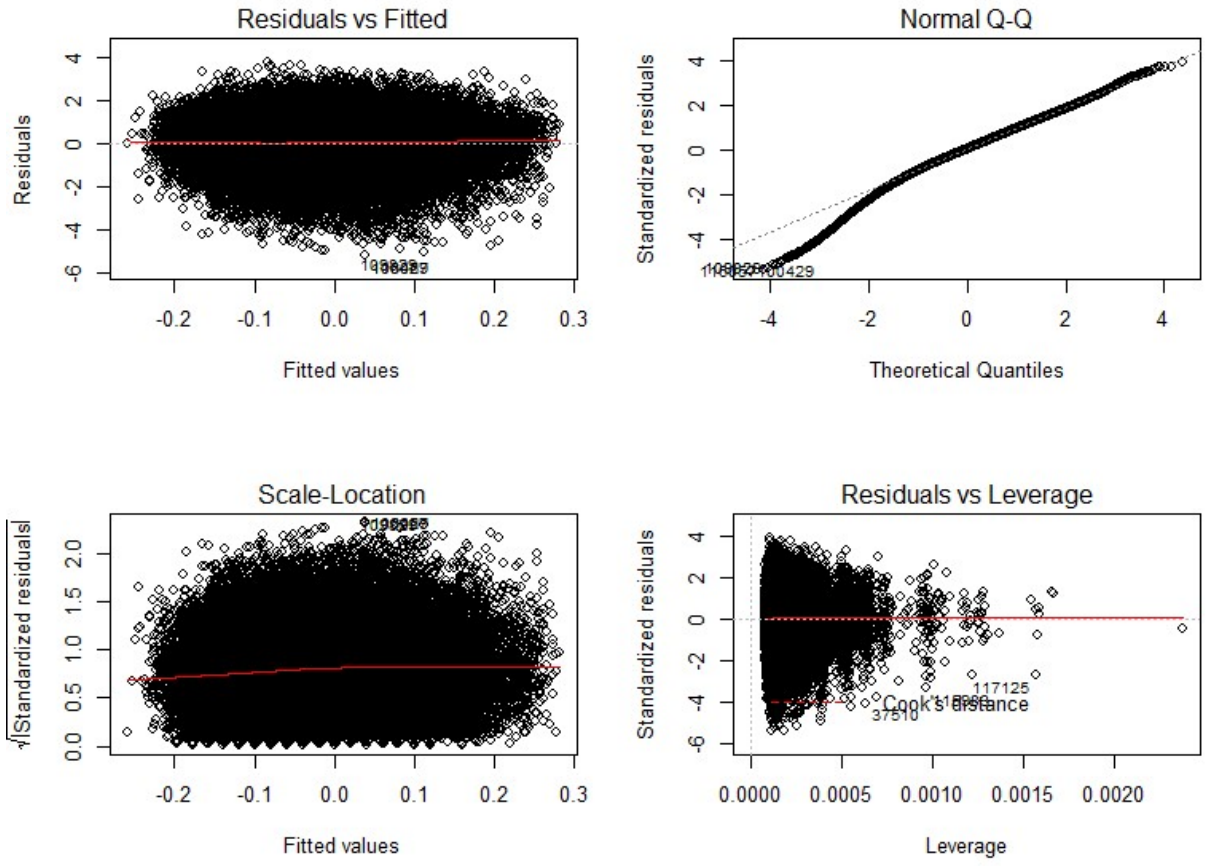

Figure 2. Residual plots checking regression assumptions for Model 2

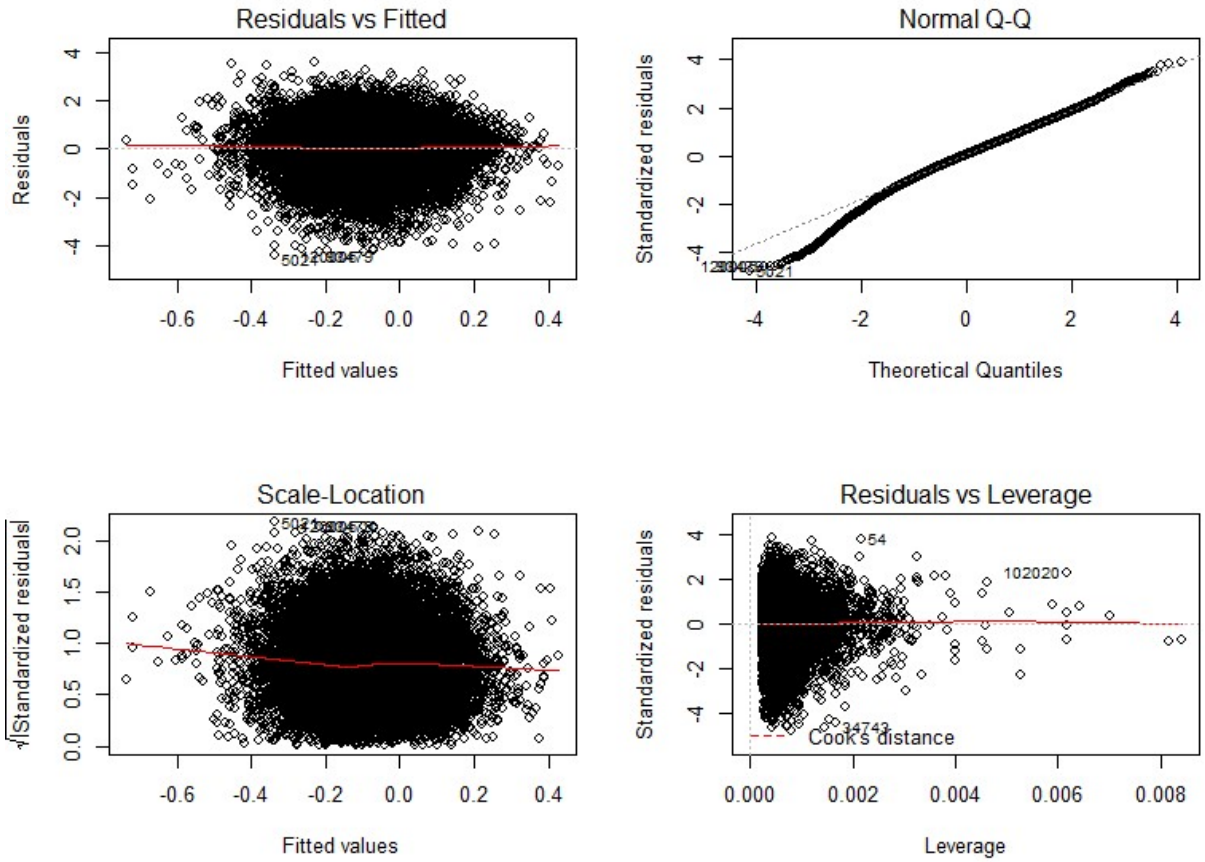

Figure 3. Residual plots checking regression assumptions for Model 3