The Performance Appraisal Milieu: A Multilevel Analysis Of Context Effects In Performance Ratings

By: **J. Kemp Ellington** and Mark A. Wilson

Abstract

The purpose of this study was to take an inductive approach in examining the extent to which organizational contexts represent significant sources of variance in supervisor performance ratings, and to explore various factors that may explain contextual rating variability. Using archival field performance rating data from a large state law enforcement organization, we used a multilevel modeling approach to partition the variance in ratings due to ratees, raters, as well as rating contexts. Results suggest that much of what may often be interpreted as idiosyncratic rater variance, may actually reflect systematic rating variability across contexts. In addition, performance-related and non-performance factors including contextual rating tendencies accounted for significant rating variability. Supervisor ratings represent the most common approach for measuring job performance, and understanding the nature and sources of rating variability is important for research and practice. Given the many uses of performance rating data, our findings suggest that continuing to identify contextual sources of variability is particularly important for addressing criterion problems, and improving ratings as a form of performance measurement.

Ellington, J.K. & Wilson, M.A. J Bus Psychol (2017) 32: 87. https://doi.org/10.1007/s10869-016-9437-x. Publisher version of record available at: https://link.springer.com/article/10.1007/s10869-016-9437-x

The Performance Appraisal Milieu: A Multilevel Analysis of Context Effects in Performance Ratings

- J. Kemp Ellington
- Mark A. Wilson

Abstract

Purpose

The purpose of this study was to take an inductive approach in examining the extent to which organizational contexts represent significant sources of variance in supervisor performance ratings, and to explore various factors that may explain contextual rating variability.

Design/Methodology/Approach

Using archival field performance rating data from a large state law enforcement organization, we used a multilevel modeling approach to partition the variance in ratings due to ratees, raters, as well as rating contexts.

Findings

Results suggest that much of what may often be interpreted as idiosyncratic rater variance, may actually reflect systematic rating variability across contexts. In addition, performance-related and non-performance factors including contextual rating tendencies accounted for significant rating variability.

Implications

Supervisor ratings represent the most common approach for measuring job performance, and understanding the nature and sources of rating variability is important for research and practice. Given the many uses of performance rating data, our findings suggest that continuing to identify contextual sources of variability is particularly important for addressing criterion problems, and improving ratings as a form of performance measurement.

Originality/Value

Numerous performance appraisal models suggest the importance of context; however, previous research had not partitioned the variance in supervisor ratings due to omnibus context effects in organizational settings. The use of a multilevel modeling approach allowed the examination of contextual influences, while controlling for ratee and rater characteristics.

Keywords

Job performance Performance appraisal Performance ratings Multilevel Rater effects Context effects

Introduction

Job performance is considered one of the most important variables in organizational research and practice (Bennett et al. 2006; Borman 2004), yet performance "criterion problems" are well documented (Austin and Crespin 2006; Austin and Villanova 1992). The most common method for measuring performance is a supervisory rating, and an extensive literature documents the many issues associated with ratings of performance (e.g., Landy and Farr 1980; Murphy 2008; Woehr and Roch 2012). In particular, research suggests that although ratings reflect actual ratee performance to a degree, they also reflect systematic rater effects (as well as measurement error). For example, several studies have examined the structure of multisource performance ratings (MSPR), and found relatively large idiosyncratic rater effects (Hoffman et al. 2010; Mount et al. 1998; Scullen et al. 2000). With regard to supervisor ratings specifically, depending on the sample and methodology employed, estimates range from 43 % (Hoffman et al. 2010; Scullen et al. 2000) to as much as 58 % (O'Neill et al. 2012) of performance rating variance which is idiosyncratic to the rater.

Given this evidence regarding the presence of rather large rater effects in performance ratings, one implication is that a potential solution to the criterion problem is to identify the factors that drive these effects, so that steps may be taken to lessen their impact (Murphy 2008). Consequently, researchers have investigated and identified a variety of rater and situational characteristics that influence ratings

(for thorough reviews see, Landy and Farr 1980; Levy and Williams 2004; Murphy 2008; Murphy and Cleveland 1995). However, pertinent questions remain regarding the nature and sources of performance rating variance. More specifically, although both theory and existing research suggest that *context* plays a significant role in performance appraisal (Ilgen and Feldman 1983; Judge and Ferris 1993; Levy and Williams 2004; Murphy and Cleveland 1995), it is currently unclear as to the extent to which context may be a systematic source of variance in ratings. Raters are nested within rating contexts, and as noted by Murphy and DeShon (2000), what is often viewed as idiosyncratic rater variance is more likely, "a combination of the effects of rater characteristics and the effects of the context in which the rater operates" (p. 879).

In studying contextual effects in performance appraisal, it is important to first define what is meant by the term "context." Here, we conceptualize context similar to Johns (2006), as "situational opportunities and constraints that affect the occurrence and meaning of organizational behavior as well as functional relationships between variables" (p. 386). In addition, context can be defined broadly as *omnibus* context (i.e., who, where, when, and why), as well as in terms of the *discrete* contextual characteristics of the social, task, and physical environment (Hattrup and Jackson 1996; Johns 2006; Mowday and Sutton 1993). With regard to the context of performance appraisal specifically, Murphy and Cleveland (1995) call for research on "levels of context," and note that several intra-organizational units may be salient. Organizational units and work groups (e.g., divisions, departments, offices, stores, etc.) can be viewed as omnibus contexts, particularly when they are distinctive with respect to their discrete contextual features (i.e., social, task, and physical characteristics). If supervisors do indeed vary systematically in their ratings across these organizational contexts, this not only raises further concerns regarding the validity of supervisor ratings (Murphy 2008; O'Neill et al. 2012), but also has implications for performance appraisal research and practice.

With these issues in mind, the current research sought to determine the extent to which work contexts account for variance in supervisory task performance ratings, and to also explore characteristics that are potentially responsible for this variation. In order to address these goals, we take an inductive approach in investigating sources of performance rating variance, and incorporate a multilevel modeling methodology in analyzing archival field data from a large state law enforcement agency. In examining the influence of ratees, raters, and rating contexts, we first partition the variability in supervisory performance ratings due to each source, operationalizing omnibus rating contexts (Johns 2006) as distinct organizational units. Secondly, in order to more thoroughly study the nature of the rating variability associated with each source, we include both performance-related and non-performance factors as ratee- and rater-level control variables. This not only provides critical information regarding the degree to which these variables account for both rater and contextual variance, but also gives an estimate of the remaining variability associated with each source that is yet to be explained. Finally, we also investigate several discrete contextual characteristics as potential predictors of between-context rating variance.

Multilevel Model of Supervisory Rating Variance

The majority of research to date examining sources of variance in performance ratings has incorporated a confirmatory factor analytic (CFA) approach, and has focused on MSPRs (Hoffman et al. 2010; Mount et al. 1998; Scullen et al. 2000). However, linear mixed models (LMM), including the more specific case of multilevel random coefficient (MRC) models, have

also been proposed as an alternative approach for decomposing rating variance, which may offer certain advantages (LaHuis and Avis 2007; O'Neill et al. 2012; Putka et al. 2008). For example, O'Neill et al. (2012) recently applied the MRC modeling approach to partition the variance in performance ratings due to ratees, raters, and rater–ratee interactions, and found substantial rater effects (i.e., 58 %), as well as influential predictors such as familiarity with the ratee and the number of ratees evaluated. Despite the potential benefits of the approach, MRC models have not yet been extensively applied in the case of fully "nested" rating systems in field settings (for an exception see LaHuis and Avis 2007). As mentioned previously, supervisor performance ratings are the most common method for measuring job performance (Murphy 2008; Woehr and Roch 2012), and in most cases each supervisor evaluates a unique group of ratees (i.e., the employees within their span of control). Ratees can therefore be viewed as nested within raters, and both ratees and raters are often nested within work contexts such as organizational units, thereby creating a hierarchical or multilevel data structure.

Initial Partitioning of Rater and Contextual Variance

Before examining potential explanatory variables at each level of analysis, it is necessary to first partition the variability due to groups/clusters, to provide a preliminary estimate as to the rating variance associated with raters and work contexts. With regard to rater variance, numerous models of the performance appraisal process suggest that rater characteristics, tendencies, biases, goals, and/or intentions are likely to result in the presence of rater effects in job performance ratings (DeCotiis and Petit 1978; DeNisi et al. 1984; Ilgen and Feldman 1983; Judge and Ferris 1993; Landy and Farr 1980; Levy and Williams 2004; Murphy and Cleveland 1995; Spence and Keeping 2013; Wherry and Bartlett 1982). And, as discussed previously, empirical support for this proposition is well established, with several studies demonstrating relatively large rater effects in performance ratings (Hoffman et al. 2010; LaHuis and Avis 2007; Mount et al. 1998; O'Neill et al. 2012; Scullen et al. 2000).

In addition to variability between raters (within contexts), there are also reasons to expect systematic rating differences across work environments. From a theoretical standpoint, numerous scholars in performance appraisal have proposed that ratings must be considered in context, and that both proximal and distal contextual influences are likely to shape rating behaviors (Ilgen and Feldman 1983; Judge and Ferris 1993; Levy and Williams 2004; Murphy and Cleveland 1995). Furthermore, empirical research has identified several specific situational variables that are influential in performance appraisal (e.g., rating purpose, rater accountability, climate; Greguras et al. 2003; Jawahar and Williams 1997; Mero et al. 2003; Murphy et al. 2003). Although researchers to date have not attempted to partition the variance in supervisory ratings due to omnibus context effects while controlling for ratee and rater characteristics, several other related empirical findings also suggest the likely importance of work context sources of rating variability. For example, Dierdorff and Surface (2007) examined sources of variance in peer ratings, and found significant rating variability associated with the situations (i.e., defined as distinct training exercises) in which peers performed and evaluated one another. Although these contexts are different in many respects than organizational units, importantly the performance situations varied in terms of environmental cues, required tasks, as well as normative expectations (Dierdorff and Surface 2007).

With regard to the rating contexts created by organizational units, Waldman et al. (1990) examined supervisor performance ratings collected as part of a training needs analysis, and found significant rating differences between organizational "departments" for several dimensions of performance. An average of 10 % of variance was associated with departments (Waldman et al. 1990), providing some evidence that supervisors' rating behavior may be influenced by the intra-organizational contexts in which they work. Moreover, variability across organizational units may be even greater with appraisals that have been in use for longer durations, and that are formal components of organizations' performance management systems (e.g., as opposed to one-time appraisals for assessing training needs). For example, Ilgen and Feldman (1983) noted that rating norms are more likely to develop once an appraisal system has been in place for some time. In this study, we focus specifically on examining an appraisal that has been in place in a field setting for a considerable duration (i.e., over 10 years).

Accounting for Ratee- and Rater-Level Characteristics

Although MRC models are well suited to nested performance rating data structures, it should be noted that the approach assumes that ratees are comparable across raters or higher level groups such as contexts, and if this assumption is violated (as is likely the case with field data) then efforts should be made to control for ratee characteristics (LaHuis and Avis 2007). Moreover, LaHuis and Avis (2007, p. 98) note that, "a major advantage of MRC modeling is the ability to study how the attributes of raters influence their ratings *while controlling for ratee characteristics* [italics added]" (p. 98). In the case of multilevel models incorporating additional hierarchical levels such as organizational units, this also allows the ability to examine contextual influences while holding both ratee- and rater-level characteristics constant. In other words, it is possible that rater and contextual variance merely reflects that some supervisors and organizational units have better, more experienced subordinates. In addition, other ratee- and rater-level characteristics may also vary across supervisors and units, and thus, partially explain rating differences across raters and contexts.

With MRC models, variables entered at lower levels (e.g., ratees or raters) can explain variability at the level of entry, as well as at higher levels (e.g., raters and contexts), to the extent that the lower-level predictors vary systematically across the higher level groups (note that this is only true when the variables are either scaled in their raw metric or centered around their grand mean). If rater or context effects are primarily a function of differences in ratee objective performance outcomes, ratee job tenure, supervisor rating tendencies, or other ratee- and rater-level factors, this would be evidenced by a non-significant rater or context variance component after controlling for these variables. The inclusion of ratee- and rater-level variables therefore serves dual roles, to explain the rating variability within each level, and to more accurately estimate the total variance accounted for at higher levels. Accordingly, a sequence of models were estimated which included characteristics at each level, and at each stage the variance components were tested to determine whether significant rating variance due to raters and contexts remained.

Ratee-Level Variables

After the initial rating variance partitioning, ratee variables were first added to the model that are likely to be predictive of performance differences. In particular, at least to a degree, objective measures of ratee effectiveness or performance outcomes (Campbell et al. 1993) should reflect some actual performance variability, and hence are likely to be associated with supervisor ratings (Bommer et al. 1995; Deadrick and Gardner 1997; Heneman 1986; Reb and Cropanzano 2007; Reb and Greguras 2010). In addition, to the extent that it conveys information about experience, variability in job tenure is also likely to be related to ratee performance differences, and therefore should be associated with ratings (McDaniel et al. 1988; Schmidt and Hunter 1998). Importantly, the remaining variability can therefore be interpreted as rating variance due to raters and contexts (and ratees), after controlling for differences in ratee objective performance outcomes and job tenure.

In addition, contextual factors at the ratee level may not only explain rating variance across ratees, but may also account for variability across raters and/or organizational units. Numerous studies of performance appraisal indicate the importance of observation, suggesting that the extent to which raters have had opportunities to observe ratee performance is likely to impact their appraisals (Ilgen et al. 1993; Kingstrom and Mainstone 1985; Kozlowski et al. 1986; O'Neill et al. 2012). In addition, the performance appraisal "purpose effect" is also well documented, with ratees tending to receive higher ratings when there are administrative consequences, versus when ratings are assigned for developmental or research purposes (Jawahar and Williams 1997). In particular, the use or purpose of the appraisal is believed to influence rater intentions and goals (Murphy 2008; Murphy and Cleveland 1995; Spence and Keeping 2013). Although in many applied settings the rating purpose may be consistent across all ratees, in certain cases of mixed-use appraisals this contextual factor varies at the ratee level. For example, ratings can be largely developmental for some employees, while having administrative implications for others, such as those being considered for promotion. Therefore, we controlled for both the number of documented performance incidents (a proxy for the number of observations) as well as the rating purpose, in order to determine the extent to which these ratee-level situational variables are responsible for ratee, rater, and work context rating variability.

Rater-Level Variables

Numerous supervisor characteristics may also explain rater variance in performance ratings, and potentially contextual variation as well. One of the most commonly suggested sources of rater effects includes differences in the idiosyncratic rating tendencies of the raters. More specifically, rater tendencies for leniency/severity are believed to be pervasive concerns in applied settings (Hauenstein 1992; Murphy and Cleveland 1995; Scullen et al. 2000). Furthermore, there is also evidence that leniency is a relatively stable rater tendency (Kane et al. 1995). In addition, raters also have varying degrees of supervisory job tenure, which is indicative of within-organization appraisal experience. Although research on rater experience is somewhat mixed, there is empirical evidence suggesting that experience affects the quality of rating data (Landy and Farr 1980; Zalesny and Highhouse 1992). In addition, recent research found that raters with more experience in conducting performance appraisals gave lower ratings than those with less experience (Spence and Keeping 2010). Finally, raters also differ in terms of the number of ratees they supervise and evaluate (i.e., span of control), and research by O'Neill et al. (2012) indicates that the number of ratees evaluated explains significant rating variance.

Consequently, in order to establish the degree to which rater-level characteristics account for both rater and contextual rating variance, we controlled for supervisor rating tendencies, supervisory job tenure, and span of control.

Research Question 1 (RQ1): Is there significant work context variability in supervisor task performance ratings after controlling for ratee- and rater-level characteristics?

Context-Level Characteristics

In order to potentially explain the remaining rating variability across organizational units, we also explored several context-level characteristics. As suggested previously, intra-organizational units often represent distinct social, task, and physical environments (Hattrup and Jackson 1996; Johns 2006; Mowday and Sutton 1993), and variables associated with each of these dimensions of discrete context may be influential in shaping rating behaviors. First of all, much like individual supervisors have rating tendencies, such distributional tendencies may also exist at the work context level. If a tendency exists in a given organizational unit for higher/lower ratings, then it could be expected that subsequent appraisals in that context would also display corresponding higher/lower mean ratings, even with a distinct set of ratees. In other words, if there are contextual tendencies for rating behavior, then the mean unit/context ratings for ratees from previous performance cycles may explain between-context variability in subsequent appraisals for a different group of ratees. Given the archival nature of our study, we cannot provide a definitive theoretical interpretation of contextual performance rating means; however, we can speculate as to the potential meaning of such a variable. For example, contextual rating tendencies may represent an aspect of the social context (Hattrup and Jackson 1996; Johns 2006; Mowday and Sutton 1993), which could be a function of norms, expectations, or standards for acceptable ratee behavior as well as performance ratings.

Discrete characteristics of the task context may also explain variability across organizational units. Even when employees hold the same job title, it is possible that due to their particular work context, some task activities may be performed more/less often than others. For example, previous research indicates that elements of the task context shape work role requirements (Dierdorff et al. 2009). In the law enforcement organization under study here, units are geographically dispersed, and the frequency of certain work activities varies based on the geographic region. More specifically, work contexts differ with respect to the number of accidents investigated, the number of cases made (i.e., the number of cases brought to court), and the number of calls for service. Therefore, supervisors in a context in which the level of these work activities is higher may potentially weight and evaluate task performance differently than supervisors in contexts in which these activities occur less frequently.

The impact of the physical work context on organizational behavior has been generally understudied (Johns 2006), and has also rarely been examined in performance appraisal research (Murphy and Cleveland 1995). However, previous scholars have concluded that, “physical environments play a major role in facilitating and constraining organizational action” (Elsbach and Pratt 2008, p. 182). For instance, the physical work context has been shown to impact technical role requirements (Dierdorff et al. 2009), and has been suggested to constrain task performance (Peters and O’Connor 1980). In addition to occupying distinct physical spaces geographically, contexts within a state law enforcement agency differ

in terms of the physical presence of an interstate highway. Consequently, work contexts that include an interstate may differ from contexts consisting of only rural state roads, in terms of the types of circumstances often encountered (e.g., contact with out-of-state travelers) and/or the frequency or importance of specific task performance dimensions. In other words, as discrete context dimensions are not orthogonal (Dierdorff et al. 2009), this physical context characteristic may shape ratings via its influence on the task context. Therefore, in order to further explore potential predictors of between-unit rating variation, we examined several work context characteristics.

Research Question 2 (RQ2): Do contextual rating tendencies, work activities, and/or the presence of an interstate highway explain work context variability in supervisor task performance ratings, after controlling for ratee- and rater-level characteristics?

Methods

Participants

Performance ratings were collected from members of a large state law enforcement agency as part of the organization's annual performance management process, and these archival performance records provided the data for this study. Although all ratees had the same basic job title, in order to ensure the most comparable ratee sample possible, ratees were excluded if their primary job responsibilities were not typical patrol activities (e.g., training, aviation, etc.). In addition, in order to calculate supervisor and contextual rating tendencies, raters and contexts (and their corresponding ratees) were excluded if they did not have sufficient data from previous performance cycles. Complete data were available for a sample of 804 ratees, 119 supervisors/raters, from 58 organizational units/contexts (i.e., "districts"). Ratees were nested within raters, and both ratees and raters were nested within units/contexts, which were geographically distinct and consisted of their own respective unit offices. The number of ratees per supervisor ranged from 1 to 16 (M = 6.76, SD = 2.71), and the number of supervisors per unit/context ranged from 1 to 4 (M = 2.05, SD = .78). The majority of the sample was male (97.5 %) as well as Caucasian (84.7 %).

Procedure

The organization's performance management process stipulates that supervisors provide performance ratings annually for all of their respective subordinates. Furthermore, policy dictates that supervisors document behavioral observations of their subordinates' performance throughout the course of each performance cycle (i.e., 1 year). All supervisors are provided rater training on how to record these observations, in addition to frame-of-reference training in performance ratings (Bernardin and Buckley 1981; Roch et al. 2012). Furthermore, refresher training was provided annually to all supervisors. The organization also maintains ongoing records regarding objective performance data (e.g., the number of accidents investigated, the number of cases made, etc.) by year and unit/context. Finally, in any given year, approximately 10 % of the ratees participate in the organization's annual promotion process. For those participants, the performance ratings have a stronger administrative impact, as they must receive a rating of average or above across specific performance dimensions in order to be eligible to

remain in consideration for promotion. The ratings for the majority are more developmental in nature, in that they have no direct impact on promotions or raises.

Ratee-Level Measures

Task Performance Ratings

A job analysis conducted in the organization of interest identified 10 task performance dimensions. These dimensions are incorporated into the performance management process, with supervisors documenting behavioral observations on these dimensions of performance, and providing ratings in the organization's annual evaluation process. Examples of dimensions rated include "collision investigation," "preventative patrol," and "arrest procedures." Performance dimensions are rated on a scale ranging from 1 = *excellent*, to 7 = *well below average*, which was reverse coded in order to ease the interpretation of results.

Objective Performance Outcomes

Based on organizational records regarding the number of "cases made" per ratee, as well as the number of hours on patrol, a variable was created reflecting the number of cases per hour (i.e., by dividing the total of number cases by the hours worked). Cases made consisted of a variety of objective indicators common in law enforcement settings (e.g., speeding, seatbelt, driving while impaired, and drug violations). It should be noted that this is an objective measure of performance "quantity," which does not capture performance "quality." In addition, to examine the potential for a curvilinear association, a quadratic term was also calculated for objective performance outcomes.

Ratee Job Tenure

Ratee job tenure was operationalized based on the number of months the individual worked for the organization. This operationalization can therefore be interpreted as a time-based measure (Tesluk and Jacobs 1998). As we are unable to determine how much previous law enforcement experience participants may have had with other organizations, our measure provides an indication of the within-organization ratee experience.

Number of Documented Incidents

As described previously, the performance management process required supervisors to document behavioral incidents of performance (corresponding with the performance dimensions) throughout the performance cycle. The number of incidents was therefore operationalized based on the sum total of documented performance incidents per ratee over the performance management cycle.

Rating Purpose

In order to differentiate those with a more administrative versus developmental appraisal purpose, a dummy-coded variable was created. Ratees who were participating in the organization's annual promotion process in the same year that the criterion performance ratings

were collected were coded “1” (i.e., a stronger administrative purpose), and all other employees were coded “0” (i.e., a more developmental purpose).

Rater-Level Measures

Supervisor Rating Tendency

Based on the archival performance rating records, a mean task performance rating was calculated for all supervisors who had provided ratings in any of the previous five performance cycles. It is important to note that, similar to the approach employed by Kane et al. (1995), any ratees who were evaluated by a rater in the performance management cycle used to operationalize our criterion performance data were excluded from the calculation of that rater’s mean tendency. This ensured that the mean value (i.e., predictor) did not overlap in terms of the ratees who were evaluated (i.e., criterion) by a given rater. Furthermore, any given ratee was only included once in the calculation of a rater’s mean tendency. The number of previous rating cycles used to calculate supervisor mean rating tendencies ranged from 1 to 5 (M = 1.94, SD = 1.02), and the number of previous ratees evaluated ranged from 3 to 31 (M = 10.64, SD = 5.90). The reliability of the rater means was calculated using an intraclass correlation coefficient (ICC; Bartko 1976; Bliese 2000). The ICC(2) for the rater mean tendencies was .73.

Supervisor Job Tenure

Rater supervisory experience was operationalized as the number of months the rater worked as a supervisor in the organization. This measure therefore represents a time-based measure (Tesluk and Jacobs 1998), and should be interpreted as an indication of within-organization supervisory experience (i.e., we cannot determine previous supervisory experience from other organizations).

Span of Control

Similar to other studies examining the number of ratees (LaHuis and Avis 2007; O’Neill et al. 2012), the span of control was defined as the number of subordinates/ratees supervised and evaluated by each supervisor/rater.

Context-Level Measures

Contextual Rating Tendency

Contextual rating tendencies were operationalized in a manner similar to our measure of supervisor rating tendencies, as the mean task performance rating in each unit/context from the previous five performance cycles. Again, any ratees who were evaluated in a given context during the cycle used to operationalize our criterion data were excluded from the calculation of the contextual rating tendency (i.e., mean) for that unit. Consequently, the mean value (i.e., predictor) did not overlap in terms of the ratees who were evaluated (i.e., criterion) in a given context. Furthermore, any given ratee was only included once in the calculation of a context’s mean tendency. The majority (89 %) of units/contexts included rating data from two or more of the five previous performance cycles, however several units were recently formed, and thus only included data from the previous year. Therefore, the number of previous rating cycles used to calculate the contextual rating tendencies ranged from 1 to 5 (M = 4.25, SD = 1.33), and the

number of previous ratees evaluated ranged from 3 to 43 (M = 16.03, SD = 9.39). The ICC(2) for the contextual rating tendencies was .72, indicating the reliability of the group means.

Work Activity

Organizational data were obtained for three indicators of contextual work activity: the number of accidents investigated (M = 1714.86, SD = 956.90), the number of cases made (M = 17,703, SD = 7834.24), and the number of calls for service (M = 6798.95, SD = 2758.78). In order to combine the indicators into a single work activity variable, each indicator was first standardized, and an average was calculated across the three standardized values.

Interstate Highway

Each unit/context covers a different geographic region, approximately 55 % of which contain an interstate highway. Therefore, a dummy-coded variable was created. Contexts that included an interstate highway were coded "1," and those that did not were coded "0."

Analytical Approach

A multilevel modeling approach was incorporated to address our research questions, as this method is appropriate for nested or hierarchical data, and allows for the simultaneous modeling of both within- and between-group variance (Raudenbush and Bryk 2002). This approach allows intercepts (means) to vary as a function of nested groups (i.e., raters and work contexts), and therefore allows the partitioning of rating variance due to ratees (within-rater), raters (between-rater, within-unit), and contexts (between-unit). A staged modeling approach was incorporated, with the first stage including the estimation of an unconditional or "null model" with no predictors, in order provide the initial partitioning of variance in ratings. More specifically, the null model results allow the calculation of ICC(1), which indicates the proportion of total variance explained by group membership (Bliese 2000; Raudenbush and Bryk 2002). For comparison purposes, a preliminary two-level, null model was first estimated with ratees (level-1) nested within raters (level-2), in order to determine the proportion of variance assigned to the rater when ignoring context. All subsequent analyses included three-level models, with ratees comprising level-1, raters as level-2, and contexts as level-3. The null model was followed by a series of random intercept and fixed slope models (RIFSM; Aguinis et al. 2013; Raudenbush and Bryk 2002), which included entering our various predictors from each level in stages. Predictors were centered around their grand mean, as our research questions were consistent with an "incremental" perspective (Hofmann and Gavin 1998). All multilevel modeling was conducted using HLM 7 software (Raudenbush et al. 2011).

Results

Table 1 presents the descriptive statistics and zero-order correlations for all study variables. An examination of the correlations among the ratee-level (level-1) variables indicates that task performance ratings were positively correlated with objective performance outcomes (r = .16, p < .01), job tenure (r = .18, p < .01), number of documented performance incidents (r = .08, p < .05), and an administrative rating purpose (r = .16, p < .01). In addition, rater-level (level-2) correlations suggest that raters' previous rating tendencies were negatively associated with supervisory experience (r = −.23,

$p < .01$). The context-level (level-3) correlations indicated that previous contextual rating tendencies were negatively related to work activity ($r = -.42$, $p < .01$), and contexts with interstate highways had higher levels of work activity ($r = .36$, $p < .01$).

Table 1

Descriptive statistics and zero-order correlations

	M	SD	1	2	3	4
Ratee-level (L1) variables						
1. Objective performance outcomes	1.17	.60				
2. Ratee job tenure	100.43	78.89	−.15**			
3. Number of documented incidents	16.83	6.25	.18**	−.04		
4. Rating purpose	.14	.35	−.11**	.21**	−.02	
5. Task performance ratings	4.70	.41	.16**	.18**	.08*	.16**
Rater-level (L2) variables						
1. Supervisor rating tendency	5.02	.30				
2. Supervisor job tenure	55.70	36.75	−.28**			
3. Span of control	7.31	2.81	.17	−.16		
Context-level (L3) variables						
1. Contextual rating tendency	4.91	.22				
2. Work activity	.02	.93	−.42**			
3. Interstate highway	.55	.50	−.11	.36**		

L1 level 1, *L2 level 2*, *L3* level 3, L1 $N = 804$, L2 $N = 119$, and L3 $N = 58$

* $p < .05$; ** $p < .01$

With regard to the multilevel model results, the preliminary two-level, null model showed significant rater variance ($\tau_{00} = .08$, $df = 118$, $\chi^2 = 774.34$, $p < .001$), and suggested that 47 % of the rating variability would be attributed to the rater when ignoring context. The three-level model results are presented in Table 2. The null model estimating both rater and context effects indicated that raters ($\tau_{\pi 0} = .03$, $df = 61$, $\chi^2 = 189.81$, $p < .001$) as well as contexts ($\tau_{\beta 00} = .05$, $df = 57$, $\chi^2 = 187.19$, $p < .001$) accounted for significant variability in supervisor task performance ratings. More specifically, 17 % of the variance was attributable to raters, and 28 % was associated with work contexts.

Table 2

Multilevel modeling results

Level and variable	Model				
	Null	**RIFSM 1**	**RIFSM 2**	**RIFSM 3**	**RIFSM 4**
Ratee level (L1)					
Intercept (γ_{000})	4.692** (.035)	4.691** (.035)	4.689** (.036)	4.689** (.035)	4.685** (.033)
Objective performance outcomes (γ_{100})		.318** (.044)	.282** (.043)	.289** (.044)	.312** (.044)
Objective performance outcomes quadratic (γ_{200})		−.036** (.007)	−.035** (.007)	−.036** (.007)	−.040** (.008)
Ratee job tenure (γ_{300})		.001** (.000)	.001** (.000)	.001** (.000)	.001** (.000)
Number of documented incidents (γ_{400})			.013** (.004)	.013** (.004)	.013** (.004)
Rating purpose (γ_{500})			.096** (.027)	.098** (.027)	.095** (.027)
Rater level (L2)					
Supervisor rating tendency (γ_{010})				.299** (.077)	.320** (.078)
Supervisor job tenure (γ_{020})				.001 (.001)	.001 (.001)
Span of control (γ_{030})				−.002 (.008)	.001 (.007)
Context level (L3)					
Contextual rating tendency (γ_{001})					.484** (.180)
Work activity (γ_{002})					-.006 (.037)
Interstate highway (γ_{003})					.007 (.073)
Variance components					
Within-rater (L1) variance (σ^2)	.094	.081	.076	.076	.076
Between-rater within-context (L2) variance ($\tau_{\pi 0}$)	.030**	.031**	.028**	.023**	.022**
Between-context (L3) variance ($\tau_{\beta 00}$)	.048**	.048**	.053**	.049**	.040**
Additional information					
Rater (L2) ICC(1)	.174				

Level and variable	Model				
	Null	RIFSM 1	RIFSM 2	RIFSM 3	RIFSM 4
Context (L3) ICC(1)	.278				
Ratee-level (L1) pseudo R^2	–	.133	.185	.185	.185
Rater-level (L2) pseudo R^2	–	.000	.048	.234	.268
Context-level (L3) pseudo R^2	–	.000	.000	.000	.165

Values in parentheses are robust standard errors; *t* statistics were computed as the ratio of each regression coefficient divided by its standard error

RIFSM random intercept and fixed slope model, *ICC* intraclass correlation coefficient, *L1* level 1, *L2* level 2, *L3* level 3, L1 $N = 804$, L2 $N = 119$, and L3 $N = 58$

* $p < .05$; ** $p < .01$

The initial RIFSM results indicated that both objective performance outcomes ($\gamma_{100} = .32$, $p < .01$) and ratee job tenure ($\gamma_{300} = .00$, $p < .01$) predicted rating variability. In addition, the quadratic term for objective performance outcomes was also significant ($\gamma_{200} = -.04$, $p < .01$), suggesting that the initial positive linear association between objective performance outcomes and ratings diminishes at higher levels of cases per hour. These predictors explained 13 % of the ratee variance, but did not explain rater or contextual variability, with significant rater ($\tau_{\pi 0} = .03$, $df = 61$, $\chi^2 = 214.26$, $p < .001$) and context ($\tau_{\beta 00} = .05$, $df = 57$, $\chi^2 = 191.92$, $p < .001$) variance remaining. The second model introduced ratee-level contextual variables, and found that both the number of documented performance incidents ($\gamma_{400} = .01$, $p < .01$) and rating purpose ($\gamma_{500} = .10$, $p < .01$) were significant predictors. The combination of all level-1 variables explained 19 % of the ratee variability, and 5 % of the rater variance in ratings. Again, significant variability remained between raters ($\tau_{\pi 0} = .03$, $df = 61$, $\chi^2 = 209.29$, $p < .001$) and contexts ($\tau_{\beta 00} = .05$, $df = 57$, $\chi^2 = 219.75$, $p < .001$). The third model added rater-level variables, and the results indicated that the supervisor's rating tendency ($\gamma_{010} = .30$, $p < .01$) was a significant predictor. However, neither supervisory tenure ($\gamma_{020} = .00$, $p > .05$) nor span of control ($\gamma_{030} = -.00$, $p > .05$) were significant. The addition of the rater characteristics explained a total of 23 % of the rater variance. This model also addressed research question 1 (RQ1), in that significant variability remained across raters ($\tau_{\pi 0} = .02$, $df = 58$, $\chi^2 = 181.36$, $p < .001$) and contexts ($\tau_{\beta 00} = .05$, $df = 57$, $\chi^2 = 232.19$, $p < .001$) after controlling for ratee- and rater-level characteristics.

The final model indicated that contextual rating tendencies were positively associated with performance ratings ($\gamma_{001} = .48$, $p < .01$), but neither contextual work activity ($\gamma_{002} = -.01$, $p > .05$) nor the presence of an interstate ($\gamma_{003} = .01$, $p > .05$) were significant, addressing research question 2 (RQ2). With regard to the variance explained across the three levels in the final model, the respective predictors explained 19 % of the ratee variability, 27 % of the rater variability, and 17 % of the contextual variability in ratings. After all variables were included, significant variation remained between raters ($\tau_{\pi 0} = .02$, $df = 58$, $\chi^2 = 180.08$, $p < .001$) and contexts ($\tau_{\beta 00} = .04$, $df = 54$, $\chi^2 = 204.19$, $p < .001$).

Discussion

The goal of this research was to contribute to the existing literature on sources of variance in job performance ratings, by partitioning rating variability due to several sources. In particular, our study adds additional evidence to the proposition that contexts can play an important role in shaping rating behavior (Levy and Williams 2004; Murphy and Cleveland 1995). Although disentangling rater and contextual rating variance is difficult (Murphy and DeShon 2000), partitioning variance due to omnibus contexts in terms of distinct units within an organization provides a potentially useful approach for separating additional sources of variance. Although the findings here should be replicated in other settings/samples to ensure generalizability, our results suggest that much of what is often considered to be rater variance may be systematic contextual variability. More specifically, when estimating a two-level model using our data (i.e., ignoring context), 47 % of the rating variability would be interpreted as rater variance. It should be noted that this estimate is very similar to those found in other studies (i.e., 43–58 %) examining supervisory rating variance (Hoffman et al. 2010; O'Neill et al. 2012; Scullen et al. 2000). However, when the rating context is modeled, the data suggest that 28 % represents contextual variation, and 17 % reflects rater variance (within context). Although rater variance still represents a large portion of rating variability, our findings indicate that an even greater proportion may be due to aspects of the work environment that influence the rating behavior of the supervisors in those contexts. Importantly, this contextual variation remained even after accounting for several ratee- and rater-level characteristics.

We also examined a diverse set of predictor variables across levels, in order to determine the extent to which these commonly cited factors in performance appraisal research explain the rating variance associated with ratees, raters, and contexts. First of all, the results suggest that ratee differences in objective performance outcomes (i.e., quantity) and job tenure explained variance at the ratee level (13 %), but did not account for variance across supervisors or work contexts, suggesting that a large portion of the rater and context variance may be due to other factors. Although our measure of objective performance is certainly an imperfect one, and result-based measures of performance often do not correlate strongly with performance ratings (Bommer et al. 1995; Heneman 1986), incorporating this variable (and job tenure) allowed at least some degree of control over potential true performance differences across raters and contexts (LaHuis and Avis 2007). It is also interesting to note that we found evidence of a curvilinear association between objective performance outcomes and ratings, which to our knowledge had not been examined in previous investigations of the relationship between objective and subjective measures (Bommer et al. 1995; Heneman 1986). The positive linear association between outcomes and ratings plateaued and then appeared to become negative. However, this only occurred at very high levels of cases per hour (i.e., over 4 SDs above the mean), therefore caution should be taken in interpreting this finding. With the inclusion of the ratee-level contextual characteristics (i.e., number of documented incidents and rating purpose), a total of 19 % of the level-1 rating variability was accounted for, and these variables explained a small portion of the variability across raters (5 %).

With regard to rater characteristics, idiosyncratic tendencies for leniency are often cited as a ubiquitous concern in performance appraisal, and a likely driver of rating differences across supervisors (Hauenstein 1992; Murphy and Cleveland 1995; Scullen et al. 2000). In addition, as

noted previously, research has suggested that tendencies for leniency are a relatively stable rater characteristic over time (Kane et al. 1995). Following the approach employed by Kane et al. (1995), we were able to estimate the extent to which supervisors' rating tendencies (i.e., mean task performance ratings) from the past, were associated with their subsequent ratings of a different set of ratees. The results indicated that these previous rating tendencies explained an additional 19 % of the between-rater variability, but did not explain contextual variation. This is noteworthy, in that a supervisor's mean tendency seems to account for about one-fifth of the rater effect, after controlling for ratee-level characteristics. Although several previous studies suggested an association between performance ratings and supervisory experience (Landy and Farr 1980; Spence and Keeping 2010; Zalesny and Highhouse 1992), as well as span of control (O'Neill et al. 2012), our data did not support a link between these characteristics and rating behavior. However, findings regarding these particular rater variables have been mixed, as other researchers also did not find a significant relationship (Judge and Ferris 1993; Klores 1966; LaHuis and Avis 2007). The mixed results across studies suggest that other factors may moderate the extent to which supervisory experience and span of control predict ratings. For example, the previously cited study which found a relationship between the number of ratees evaluated and supervisor ratings (O'Neill et al. 2012) used a "relative" appraisal approach (Goffin et al. 2009), therefore it may be that span of control is only influential when explicitly making comparisons among ratees. In addition, our measure of supervisory tenure was a within-organization, time-based operationalization, so supervisory experience may be more meaningful when considering other definitions of experience (e.g., amount or density; Tesluk and Jacobs 1998).

One of the primary objectives of this research was to not only estimate the amount of rating variability due to rating contexts, but to also attempt to explain this variability. The collection of ratee- and rater-level characteristics above did not account for significant variance across contexts, suggesting that other factors are driving the rating differences across organizational units. We proposed that rating tendencies may also exist at the work context level, and thus could be influential in explaining between-unit differences in performance ratings. Our results suggest that the rating distributional tendencies (i.e., means) of organizational units show some level of consistency over time, with previous rating tendencies predicting subsequent ratings of a distinct group of ratees. Given the limitations of our data, we are unable to determine the mechanism driving these mean tendencies; however, we previously offered conjecture that one possible explanation is that these tendencies reflect an aspect of the social context, and are the result of contextual norms or standards for performance and rating behavior. If this were the case, it would be consistent with prior theory (DeCotiis and Petit 1978; Ilgen and Feldman 1983; Murphy and Cleveland 1995; Spence and Keeping 2013) as well as previous lab-based research (Shore and Tashchian 2002; Spence and Keeping 2010) suggesting the importance of rating norms in performance appraisal. We also explored characteristics of the task and physical work contexts; however, these variables were not predictive of between-context rating variability (discussed further in Future Research). The inclusion of contextual rating tendencies accounted for 17 % of the context effect, and it is important to note again that this relationship was demonstrated while holding all of the previously discussed ratee and rater characteristics constant, and in an organization in which supervisors are provided frame-of-reference training, along with annual refresher training (Bernardin and Buckley 1981; Roch et al. 2012). Although many others have cited the potential importance of the rating context in performance appraisal

(e.g., Murphy 2008; Murphy and Cleveland 1995), our findings add valuable empirical evidence as to the extent to which this may be the case in field settings.

Study Limitations

The findings presented here should be considered in light of several study limitations. First, given that the data analyzed here were from a single organization (which was a predominantly Caucasian, male sample from a law enforcement organization), this may limit the generalizability of our results. The similarity of our estimate of rater variability (i.e., if ignoring context) to those found in other studies does suggest that our results are comparable to previous research; however, future studies should seek to partition rating variance due to context in other settings, and with more diverse samples. Second, our within-context (i.e., raters per context) sample size was relatively low, which may have impacted our results. Although the organization here was a fairly large organization, even larger samples may be needed to examine contexts with more supervisors per unit.

Third, several potential issues with the measures incorporated here deserve mention. For example, though raters were excluded who had only completed one or two previous appraisals, rater and contextual rating tendencies were in some cases based on relatively low numbers of previous evaluations, and thus the tendencies in those cases may represent less stable estimates. Nonetheless, overall the rater/context means were relatively reliable, and were predictive of both rater and contextual variability. In addition, as mentioned previously, our measure of objective performance was a results-based operationalization that did not capture performance "quality," and may not have been an adequate control for true ratee performance differences. Therefore, some degree of the remaining rater and context variability likely still reflects actual performance differences across groups. In addition, the nature of our data prevents us from drawing definitive conclusions regarding the extent to which all variables represent valid or biasing sources of variability. Of the significant predictors in our study, we believe there are plausible reasons to expect that objective performance outcomes (quantity) and job tenure likely reflect at least some degree of valid rating variability, and that rating purpose and rater/contextual distributional tendencies likely reflect bias; however, this is less clear for documented incidents of performance. In our study, the number of incidents were positively associated with ratings, but we cannot determine if the rating variance explained represents true ratee performance differences, or bias based on supervisor familiarity (or lack thereof) with the ratee. Furthermore, though we described the number of documented incidents as a proxy measure for the opportunity to observe performance, other factors may in fact have systematically impacted the number of incidents recorded. For example, rater motivation or beliefs about the importance of documenting performance may have more to do with the number of incidents recorded than actual opportunity to observe performance (Harris 1994). However, again, this variable was nevertheless associated with ratee performance variability.

Practical Implications and Future Research

Given the inductive nature of our study, we believe caution should be taken in making recommendations for practice; however, there are a few important practical implications of our research. First off, the presence of relatively large rater and context effects in supervisor ratings

suggests that practitioners (or researchers) utilizing ratings as criteria when validating selection instruments should incorporate analytic approaches which account for this nested data structure (e.g., MRC models). Previous research demonstrates that ignoring the nonindependence in criterion data can have the effect of attenuating statistical power, particularly with higher levels of between-group variance (Bliese and Hanges 2004). In other words, if doing validation with performance ratings as the criterion using a method that does not account for the hierarchical nature of the data (e.g., ordinary least squares regression), one may erroneously conclude that an individual-level predictor (e.g., selection test) is not significantly related to performance. In addition, our results suggest that caution should be taken when using supervisor ratings to make ratee comparisons across supervisors and/or contexts for administrative purposes, as inconsistent expectations and standards may exist (Ilgen and Feldman 1983), even when attempts have been made to impart a common frame-of-reference among supervisors (Bernardin and Buckley 1981; Roch et al. 2012). Furthermore, this issue may be even more pronounced in multinational organizations, where work contexts potentially differ more drastically in terms of their social, task, and physical characteristics.

With regard to avenues for future research, this study demonstrates the potential utility of the MRC modeling approach in better understanding sources of performance rating variability, and in evaluating performance appraisal interventions. Previous research has incorporated this approach in examining rater effects (LaHuis and Avis 2007; O'Neill et al. 2012), and we believe the benefits extend to the study of additional levels/variables such as contexts. Studying sources of variability has been proposed as a useful approach for studying the quality of rating data in field research (Hoffman et al. 2012), as opposed to utilizing direct measures of rating accuracy, which is typically confined to lab settings (Murphy and Cleveland 1995). For example, despite a historical moratorium on rating scale research (Landy and Farr 1980), more recently scholars have proposed innovative new performance measurement methods (e.g., Borman et al. 2001; Hoffman et al. 2012), and additional field research investigating these approaches is warranted (Landy 2010). The MRC modeling approach could for instance be used to examine the effect of scale design interventions on systematic rater and contextual rating variability.

However, although we are certainly not the first to call for research on the context of performance appraisal (e.g., Levy and Williams 2004; Murphy and Cleveland 1995), our findings regarding significant differences associated with organizational units suggest a particular need for future research on contextual sources of rating variance. As described previously, intra-organizational units may differ in terms of discrete social, task, and physical characteristics (Hattrup and Jackson 1996; Johns 2006; Mowday and Sutton 1993), and additional research on all three of these contextual dimensions may be beneficial. For example, if future research can confirm that social norms or climates for performance management are a mechanism driving contextual variation, this would be in accordance with previous suggestions that, "the interventions most likely to improve the quality of performance appraisals in organizations are likely to look more like organizational development than scale development" (Murphy 2008, p. 158). In other words, an additional approach to the criterion problem may be to continue to identify social contextual influences (Levy and Williams 2004), and to potentially attempt to take steps to avoid contextual norms for rating behavior that may not support accurate evaluations of job performance or employee development. For example, many have called for research on rater goals and intentions (Murphy 2008; Murphy and Cleveland 1995; Spence and

Keeping 2011), and MRC models seem particularly relevant for examining the degree to which these goals are primarily a ratee- or rater-level phenomenon, or a function of goals which result from contextual influences.

Building on the points above, research is needed to better understand the nature and development of norms for rating behavior. In this study, we operationalized rating tendencies based on actual distributional rating characteristics (i.e., means), however, future research should directly collect field data regarding perceived normative influences from supervisors or peers. For example, Spence and Keeping (2013) propose a framework based on the theory of planned behavior (Ajzen 1991; Ajzen and Fishbein 2005), which includes several propositions regarding the role of subjective norms in influencing rater intentions and behavior. To the extent that we can better understand these influences, researchers and/or practitioners may be able to develop and test new interventions to potentially shape the development of norms across organizational units/contexts.

Furthermore, although in our study features of the task and physical context did not explain rating variance, additional research should examine these factors in order to identify the circumstances under which these aspects of the work context may be of more/less influence. We examined the frequency of certain work activities, but other factors such as contextual differences in task importance or difficulty may shape the standards used to evaluate performance. In addition, previous research indicates that supervisors develop "folk theories" of task performance (Borman 1987), and these views may be in part a function of the task context in which the supervisor works. Research also indicates that other features of the task context influence role requirements (e.g., accountability, autonomy, routinization; Dierdorff et al. 2009) and performance ratings (e.g., accountability; Mero et al. 2003), and these factors may explain contextual rating differences to the extent that they vary systematically across intra-organizational work contexts. Furthermore, although the physical presence of an interstate highway did not seem to influence ratings in our study, other physical context characteristics may be relevant depending on the organizational setting. For instance, physical work space characteristics such as the number of enclosures have been shown to influence performance (Oldham et al. 1991), and differences in these physical characteristics may influence the nature/frequency of employee performance-related interactions or the importance of work tasks.

Conclusion

Employee performance ratings are at least in part dependent on the supervisor/rater who produces them, as well as the work context in which they are produced. Given the many issues associated with ratings, there is currently a debate as to whether ratings should be abandoned altogether, or if continued efforts should be made to improve upon them as a component of performance management (Adler et al., in press). Going forward, it remains to be seen as to whether the latter goal can be achieved; however, if future efforts are to be made toward improving ratings, we believe that continuing to identify contextual influences in appraisal is a worthy endeavor. Although many questions remain regarding the nature of contextual rating variability, this line of research (among others) may help to better understand and hence improve ratings as a form of performance measurement.

References

1. Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K. R., Ollander-Krane, R., et al. (in press). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice*.
2. Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*, 1490–1528. doi: 10.1177/0149206313478188.
3. Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211. doi: 10.1016/0749-5978(91)90020-T.
4. Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Lawrence Erlbaum Associates.
5. Austin, J. T., & Crespin, T. R. (2006). Problems of criteria in industrial and organizational psychology: Progress, pitfalls, and prospects. In W. Bennett Jr, C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 9–48). Mahwah, NJ: Lawrence Erlbaum Associates.
6. Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836–874.
7. Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*, 762–765.
8. Bennett, W, Jr, Lance, C. E., & Woehr, D. J. (2006). Introduction. In W. Bennett Jr, C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 1–5). Mahwah, NJ: Lawrence Erlbaum Associates.
9. Bernardin, H. J., & Buckley, R. B. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205–212.
10. Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.
11. Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7*, 400–417. doi: 10.1177/1094428104268542.
12. Bommer, W. H., Johnson, J., 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. doi: 10.1111/j.1744-6570.1995.tb01772.x.
13. Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307–322.
14. Borman, W. C. (2004). The concept of organizational citizenship. *Current Directions in Psychological Science, 13*, 238–241.
15. Borman, W. C., Buck, D. E., Motowildo, S. J., Hanson, M. A., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965–973. doi: 10.1037//0021-9010.86.5.965.

16. Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.
17. Deadrick, D. L., & Gardner, D. G. (1997). Distributional ratings of performance levels and variability. *Group and Organization Management, 22*, 317–342.
18. DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review, 3*, 635–646.
19. 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 & Human Performance, 33*, 360–396.
20. Dierdorff, E. C., Rubin, R. S., & Morgeson, F. P. (2009). The milieu of managerial work: an integrative framework linking work context to role requirements. *Journal of Applied Psychology, 94*, 972.
21. Dierdorff, E. C., & Surface, E. A. (2007). Placing peer ratings in context: Systematic influences beyond ratee performance. *Personnel Psychology, 60*, 93–126. doi: 10.1111/j.1744-6570.2007.00066.x.
22. Elsbach, K. D., & Pratt, M. G. (2008). The physical environment in organizations. In J. P. Walsh & A. P. Brief (Eds.), *The academy of management annals* (Vol. 1, pp. 181–224). New York: Taylor & Francis Group/Lawrence Erlbaum Associates.
23. Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*, 251–268. doi: 10.1002/hrm.20278.
24. Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, M, I. I. I. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1–21.
25. Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737–756.
26. Hattrup, K., & Jackson, S. (1996). Learning about individual differences by taking situations seriously. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 507–547). San-Francisco: Jossey-Bass.
27. Hauenstein, N. M. A. (1992). An information-processing approach to leniency in performance judgments. *Journal of Applied Psychology, 77*, 485.
28. Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811–826.
29. Hoffman, B. J., Gorman, C. A., Blair, C. A., Meriac, J. P., Overstreet, B., & Atchley, E. K. (2012). Evidence for the effectiveness of an alternative multisource performance rating methodology. *Personnel Psychology, 65*, 531–563. doi: 10.1111/j.1744-6570.2012.01252.x.
30. Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119–151. doi: 10.1111/j.1744-6570.2009.01164.x.
31. Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*, 623–641.
32. Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54*, 321–368.
33. Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141–197). Greenwich, CT: JAI Press.
34. Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905–925.

35. Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review, 31*, 386–408.
36. Judge, T. A., & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal, 36*, 80–105.
37. Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38*, 1036–1051.
38. Kingstrom, P. O., & Mainstone, L. E. (1985). An investigation of the rater–ratee acquaintance and rater bias. *Academy of Management Journal, 28*, 641–653. doi: 10.2307/256119.
39. Klores, M. S. (1966). Rater bias in forced-distribution performance ratings. *Personnel Psychology, 19*, 411–421.
40. Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: An exploration. *Journal of Applied Psychology, 71*, 45–49.
41. LaHuis, D. M., & Avis, J. M. (2007). Using multilevel random coefficient modeling to investigate rater effects in performance ratings. *Organizational Research Methods, 10*, 97–107.
42. Landy, F. (2010). Performance ratings: Then and now. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 227–248). New York: Routledge/Taylor & Francis Group.
43. Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.
44. Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*, 881–905.
45. McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology, 73*, 327–330.
46. Mero, N. P., Motowidlo, S. J., & Anna, A. L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology, 33*, 2493–2514.
47. 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*, 557–576.
48. Mowday, R. T., & Sutton, R. I. (1993). Organizational behavior: Linking individuals and groups to organizational contexts. *Annual Review of Psychology, 44*, 195–229. doi: 10.1146/annurev.ps.44.020193.001211.
49. Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148–160. doi: 10.1111/j.1754-9434.2008.00030.x.
50. Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage Publications.
51. Murphy, K. R., Cleveland, J. N., Kinney, T. B., Skattebo, A. L., Newman, D. A., & Sin, H. P. (2003). Unit climate, rater goals and performance ratings in an instructional setting. *Irish Journal of Management, 24*, 48.
52. Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.
53. O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods, 15*, 436–462. doi: 10.1177/1094428112438699.
54. Oldham, G. R., Kulik, C. T., & Stepina, L. P. (1991). Physical environments and employee reactions: Effects of stimulus-screening skills and job complexity. *Academy of Management Journal, 34*, 929–938. doi: 10.2307/256397.
55. Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. *Academy of Management Review, 5*, 391–398. doi: 10.5465/AMR.1980.4288856.

56. Putka, D. J., Ingerick, M., & McCloy, R. A. (2008). Integrating traditional perspectives on error in ratings: Capitalizing on advances in mixed-effects modeling. *Industrial & Organizational Psychology, 1*, 167–173. doi: 10.1111/j.1754-9434.2008.00032.x.
57. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
58. Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7*. Lincolnwood, IL: Scientific Software International.
59. Reb, J., & Cropanzano, R. (2007). Evaluating dynamic performance: The influence of salient gestalt characteristics on performance ratings. *Journal of Applied Psychology, 92*, 490–499.
60. Reb, J., & Greguras, G. J. (2010). Understanding performance ratings: Dynamic performance, attributions, and rating purpose. *Journal of Applied Psychology, 95*, 213–220.
61. 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 & Organizational Psychology, 85*, 370–395. doi: 10.1111/j.2044-8325.2011.02045.x.
62. Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. doi: 10.1037/0033-2909.124.2.262.
63. Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956–970.
64. Shore, T. H., & Tashchian, A. (2002). Accountability forces in performance appraisal: Effects of self-appraisal information, normative information, and task performance. *Journal of Business and Psychology, 17*, 261–274. doi: 10.1023/A:1019689616654.
65. Spence, J. R., & Keeping, L. M. (2010). The impact of non-performance information on ratings of job performance: A policy-capturing approach. *Journal of Organizational Behavior, 31*, 587–608.
66. Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review, 21*, 85–95. doi: 10.1016/j.hrmr.2010.09.013.
67. Spence, J. R., & Keeping, L. M. (2013). The road to performance ratings is paved with intentions: A framework for understanding managers' intentions when rating employee performance. *Organizational Psychology Review, 3*, 360–383. doi: 10.1177/2041386613485969.
68. Tesluk, P. E., & Jacobs, R. R. (1998). Toward an integrated model of work experience. *Personnel Psychology, 51*, 321–355.
69. Waldman, D. A., Yammarino, F. J., & Avolio, B. J. (1990). A multiple level investigation of personnel ratings. *Personnel Psychology, 43*, 811–835.
70. Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521–551.
71. Woehr, D. J., & Roch, S. (2012). Supervisory performance ratings. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 517–531). New York: Oxford University Press.
72. Zalesny, M. D., & Highhouse, S. (1992). Accuracy in performance evaluations. *Organizational Behavior and Human Decision Processes, 51*, 22–50.