Journal of Business and Psychology (2021) 36:573–588
https://doi.org/10.1007/s10869-020-09701-8

ORIGINAL PAPER

The Effects of Team Context on Peer Ratings of Task and Citizenship Performance

Joseph A. Schmidt[1] · Thomas A. O'Neill[2] · Patrick D. Dunlop[3]

Published online: 13 June 2020

Abstract

Recent trends indicate that organizations will continue their strategic pursuit of teamwork for the foreseeable future, which will create a need for accurate assessments of individuals' performance in teams. Although individual behaviors can be perceived and assessed by fellow team members (i.e., peers), the extent to which the team shapes perceivers' judgments versus the target's behavior is unclear. We conducted two studies to understand how and why team context influences peer ratings of individual performance. In study 1, we conducted cross-classified modeling on a sample of 7160 performance observations of 568 targets made by 567 perceivers, who were each members of four separate teams. Results indicated that team membership accounted for a substantially higher proportion of perceiver, relative to target, variance. In study 2, we conducted social relations modeling with a sample of 679 performance observations collected from 217 individuals nested in 46 teams to test the effects of psychological safety on perceiver, target, and team variance components. Perceptions of psychological safety accounted for proportionally larger perceiver, relative to target, variance in OCB, and task performance ratings. Altogether, team context appears to affect perceivers' judgments of behavior more than the target's behavior itself, implying that peer ratings sourced from different teams may not be comparable. We consider the implications for the collection and interpretation of peer performance ratings in teams and the potential implications for social cognitive theory, such that certain aspects of the team context, including psychological safety, may act as a cognitive heuristic by molding perceiver judgments of targets.

Keywords Peer rating · Teams · Task performance · Organizational citizenship behavior · Psychological safety

Collaboration in organizations is critical for adaptability, innovation, and learning, and recent trends suggest that top management teams plan to continue to leverage teamwork as a strategic organizational advantage (O'Neill & Salas, 2018). Assessing an individual's performance within a collaborative team environment is therefore critical for addressing human resource management (HRM) functions such as needs analysis, performance management, and training (Aguinis, 2013). Perhaps one of the most obvious potential sources of information about work behavior of individuals within teams is the perspectives of fellow team members (i.e., peers). Indeed, peers have substantial observation opportunities, interact regularly with team members, and are knowledgeable about others' behaviors (Dominick, Reilly, & McGourty, 1997). However, the leniency and severity of peer ratings appear to be affected by team membership (Loignon et al., 2017), potentially making cross-team comparisons of peer ratings inappropriate within a broader HRM system. Thus, although peers represent a potential source of information about individuals' effectiveness in teams (Ohland et al., 2012), we need to know more about what contributes to variability in the ratings from peers (cf. Bamberger, 2007).

While some research has shown that aspects of the rating context—including the type of task (Dierdorff & Surface, 2007), salience of performance relative to others (Goffin, Jelley, Powell, & Johnston, 2009), and the rating purpose (i.e., administrative versus developmental; Greguras, Robie, Schleicher, & Maynard, 2003; Jawahar & Williams, 1997)—can influence rating quality, there is much less knowledge

✉ Joseph A. Schmidt
jschmidt@edwards.usask.ca

[1] Edwards School of Business, University of Saskatchewan, 25 Campus Drive, Saskatoon, SK S7N5A7, Canada

[2] Department of Psychology, University of Calgary, 2500 University Drive N.W., Calgary, AB T2N 1N4, Canada

[3] Future of Work Institute, Faculty of Business and Law, Curtin University, Kent Street, Bentley, Western Australia 6102, Australia

regarding how the characteristics of a team may affect peer ratings of performance. The limited research that has been conducted has revealed that unit- or team-level variance in individual performance ratings is non-trivial (e.g., Ellington & Wilson, 2017; Waldman, Yammarino, & Avolio, 1990), yet little research has examined the sources of team variance in individual performance ratings. The "face-value" interpretation of peer ratings assumes that the ratings are equivalent, and therefore comparable, between different teams. If this assumption is invalid, however, then so too are inferences we draw from behavioral ratings for research, performance appraisal, or development collected from peers in nested structures (e.g., 360-degree assessments). It is therefore vital to develop more precise knowledge about the sources of team-level variance in peer ratings of behaviors. Moreover, investigating how teams influence individual ratings is particularly appropriate given that teams are a proximal and highly salient entity for most employees and team constructs are well-developed and highly predictive of team outcomes (e.g., LePine, Piccolo, Jackson, Mathieu, & Saul, 2008). The current research, therefore, may also generalize to most rating processes in organizations given that they often occur within a team context (e.g., supervisor performance ratings).

In addition to limited empirical research, there has been a corresponding lack of theory development that explains why team variance might exist in peer ratings of individual behavior. This is important because teasing out the potential social-psychological processes responsible for the effects can advance our theoretical understanding of the phenomena. Accordingly, we developed and systematically tested two plausible theoretical propositions about how team-level factors may influence peer ratings of team members' behavior. The first is the situation strength perspective (e.g., Mischel, 1973), wherein teams influence the actual *behavior* of all members, thereby resulting in systematic raising or lowering of all team members' performance ratings. The second proposition, which we refer to as the social cognitive perspective, describes how teams systematically influence the *cognitions* of people providing ratings (henceforth, we refer to these entities as perceivers) such that the context could be affecting how team members comprehend their colleagues' behavior. In other words, individuals may apply their perceptions about the team to derive overall judgments about their fellow team members or make attributions about peers' behavior that may or may not be accurate. Thus, the team may provide cues that influence how perceivers assign their peers to social categories or prototypes (e.g., diligent vs. sloppy, helpful vs. selfish) to make overall attributions about others (see Srull & Wyer, 1989) and "fill in the gaps" for the unobserved behavior of their peers. Although prior research has examined ratings within the context of teams, units, or organizations, our theorizing and research design allows us to shed light on the extent to which ratings are due to the actual behavior of those receiving ratings (henceforth, we refer to these entities as targets) versus the perceiver's cognitions related to team membership and associated team-level constructs.

We conducted two studies to address these questions. Study 1 employed a unique cross-classified design (e.g., Putka & Hoffman, 2013), which provides a valuable opportunity to isolate the variance due to perceivers and targets *across multiple team memberships*, rather than only within a single team. Such a design avoids confounding team-related variance components with who is in the team, which is a limitation of nearly all studies on this topic. That is, in studies where participants are only members of one team, it is not possible to determine if observed between-team variance in individual ratings is due to differences between teams in the average ability of the teams' members, or if the processes and emergent states created by the team are influencing target behaviors or perceiver judgments. Further, the cross-classified design allows us to estimate the extent to which team membership accounts for variance *specific to the perceivers or targets*.

In study 2, we build on study 1 by conducting social relations modeling to partition the variance of individual ratings into the target, perceiver, rating dyad, and team components. Specifically, we examine how perceptions of psychological safety influence peer ratings of task performance and organizational citizenship behavior (OCB). Given that high levels of trust and acceptance are characteristic of psychological safety, we propose that psychological safety affects perceivers' cognitions and is used as a heuristic to make positive inferences about others' behaviors, whether or not the behaviors were directly observed. We included two behavioral rating criteria because there may be differences in the extent to which perceptions of task versus trait-oriented behaviors are shared between perceivers (e.g., Dierdorff & Morgeson, 2009) and that perceivers may make different attributions regarding the motivations for OCB of fellow team members (e.g., Halbesleben, Bowler, Bolino, & Turnley, 2010). We begin by reviewing the literature on sources of rating variance and then we explore the potential causes of team variance to advance the study hypotheses and research questions.

Sources of Rating Variance

Perceptions of others' behavior form as a function of the complex interplay between individual differences and interactions among the individuals within a social context (e.g., as described in the Lens Model of person perception: cf. Kuncel, Klieger, Connelly, & Ones, 2013). More specifically, judgments (i.e., ratings) about behaviors in a team (e.g., performance or citizenship) will be influenced by the degree to which target behavior is rated consistently by all perceivers (i.e., "target effect"), the perceiver's tendencies to provide similar ratings to all targets (i.e. "perceiver effect"), the quality of the relationship between the two individuals (i.e., "perceiver-target dyad effect"), and the team environment (Christensen & Kashy, 2012).

Substantial research has been dedicated to estimating sources of target and perceiver variance in the general rating literature, reporting that 8% (Loignon et al., 2017) to 54% (Ellington & Wilson, 2017) of model variance is attributable to consistent perceptions of target behavior (for other estimates that fall within this range see Dierdorff & Surface, 2007; Hoffman, Lance, Bynum, & Gentry, 2010; O'Neill, McLarnon, & Carswell, 2015). These findings raise some concerns about the accuracy of performance ratings in particular because a relatively low percentage of target variance implies that peer ratings might contain little performance-relevant information. Indeed, research has found that a large proportion of variance in peer ratings is attributed to perceivers, which is often interpreted as a form of rater bias (Ellington & Wilson, 2017; Hoffman et al., 2010; Scullen, Mount, & Goff, 2000). Moreover, Greguras, Robie, and Born (2001) reported that perceivers accounted for more variance in overall performance ratings than did the targets, implying that, far from being entirely objective, performance judgments can be in the eye of the beholder. Finally, the nature of the dyadic relationship between perceivers and targets can also influence ratings, although often to a relatively lesser extent (e.g., 11 to 12%; Loignon et al., 2017; O'Neill et al., 2015; O'Neill, Goffin, & Gellatly, 2012).

A key focus of the current research is on team-level variance in peer ratings of team members' behavior, although as we mentioned, our theorizing may generalize to any rating context that occurs within a team (e.g., supervisor performance ratings). Given the plethora of applications of ratings in organizations, it is vital to understand how peer ratings can be compared across teams and to enhance our understanding of the processes that influence team variance in individual ratings. Loignon et al. (2017) is the only study to our knowledge that has examined the impact of team context on peer ratings of performance, and they found that context accounted for 16% of the variance in peer performance ratings of ad hoc student work teams. They also found that team-level variance was reduced to 10% in teams that received frame-of-reference rating training, suggesting that the team context influenced how peer raters framed performance behavior. Evidence that the team context can shape performance ratings can be found in several other studies that did not use peers to rate performance. As discussed earlier, the unit-level context accounted for 28% of the variance in supervisor performance ratings of police officers (Ellington & Wilson, 2017), whereas the classroom context (climate in this case) accounted for 24% of the variance in student ratings of professor performance (Murphy, Cleveland, Kinney, Skattebo, & Newman, 2003). In both studies, the hypotheses regarding mediators and moderators that could explain the influences on team variance were not supported, indicating that more research is required to understand the causes of team variance and the viability of using peer ratings for administrative and developmental purposes.

Possible Causes of Team Rating Variance

As we noted earlier, it is important to understand why teams contribute to individuals' rating variance. If we can better understand the behavioral and social-psychological processes responsible for the effects of team contexts, we can develop a deeper theoretical understanding of ratings in teams and potentially improve the accuracy of those ratings through targeted interventions. The effects of team characteristics on evaluations of individual behavior may be explained by how the team facilitates or constrains the expression of behavior. In essence, this explanation falls in line with Mischel's (1973) classic "situation strength" concept, where situations are "strong" to the extent that behavioral expectations are clear and perceived consistently across individuals; there are sufficient incentives to align behavior with expectations; and individuals possess the skills or are provided with the skills necessary to produce the desired behaviors. Similar, but more recent, theorizing has argued that situations are strong when expectations are clear, different sources provide consistent cues about expected behaviors, there are significant constraints restricting individual discretion, and decisions or actions have important consequences (Meyer, Dalal, & Hermida, 2010). If teams create such situations, they should produce relatively uniform behavior among team members and individual differences, such as personality and ability, may contribute relatively less to observable behavior. Thus, individuals will largely demonstrate behavior that is relevant to team expectations, whether adaptive or maladaptive, because the team has created powerful incentives and selected or endowed team members with the skills to meet expectations, ultimately leading to within-team homogeneity. In such situations, the team influences individual behavior and peer performance rating variance across teams reflect the target's actual behavior rather than idiosyncratic perceiver judgments.

Meta-analytic evidence has indicated that many of the studies invoking situation strength explanations have not conducted direct tests of the theory, which require examining how stronger versus weaker situations restrict variance in the dependent variable (Keeler, Kong, Dalal, & Cortina, 2019). According to those authors, of the studies that have tested or reported differences in the variance of the dependent variable across situations, only a minority have supported situation strength theory. As such, it may be valuable to investigate alternative explanations for how teams influence performance ratings and theories of social cognition appear to be particularly relevant; to that end, we also consider the role of social-cognitive processes.

Specifically, individuals tend to view other team members as a collective category (Savitsky, Van Boven, Epley, & Wight, 2005) and may impose their interpretation of team characteristics on individual team members. As such, perceivers may be less prone to detect and utilize behavioral cues of individual team members and instead rely on their evaluation of team characteristics to derive probabilistic/likely judgments about the target's behavior. This line of reasoning is consistent with theories of social categorization, which suggest that judges conserve cognitive resources by quickly assigning targets to general conceptual categories based on early observations of small samples of behavior (Srull & Wyer, 1989). Future behavior is interpreted with respect to category membership and unknown or unobserved information is assumed based on behaviors that are prototypical for that category (Favero & Ilgen, 1989).

If the social cognition explanation is accurate, then teams may influence how perceivers assign targets to conceptual categories, which then influences how subsequent target behavior is recalled and interpreted by perceivers. For example, an individual in a safe and accepting team may assume positive characteristics of all team members and only search for and utilize behavioral cues that confirm their expectations, thereby creating a form of halo that influences judgments of others (cf. Allen & O'Neill, 2015). Consistent with this explanation, research has found that higher team satisfaction can reduce self-serving biases among individual team members (Behfar, Friedman, & Oh, 2016).

We address the situation strength and social cognition explanations in the studies that follow. In study 1, we determine how much perceiver and target variance in individual performance ratings is attributable to team membership for people who are members of multiple teams. The results of that study provide estimates of the extent to which teams influence ratings via situation strength or perceiver cognitions within a cross-classified design that randomizes team membership. In the second study, we examine how the emergent state of psychological safety accounts for team-, perceiver-, and target-level variance in task performance and OCB ratings. Table 1 provides an explanation of the different variance components in studies 1 and 2.

Study 1

The purpose of study 1 was to provide a reliable estimate of the magnitude of team effects on perceiver and target variance in peer performance ratings. We did this using a stronger design than is typically employed, that is, by invoking cross-classified modeling (see O'Neill et al., 2012) of people who were members of *multiple* teams. This research design accounts for unobserved individual differences of targets and perceivers while estimating the proportion of target and perceiver variance that is attributable to team membership. Given our previous arguments, it is plausible that teams create situations that reinforce consistent behavior among targets and consistent performance judgments by perceivers; however, the current theory does not provide precise explanations about which mechanism has the strongest influence on ratings. We thus expect team membership to account for some perceiver and target variance in peer ratings, but we cannot predict if there will be differences in the magnitude of variance components. This first study was critical in order to understand the extent to which team membership affects individuals' ratings, and to identify which variance components (i.e., target versus perceiver) are affected most by the team context. More specifically, if teams create unique, strong situations that restrict within-team variance in behavior, then the target's team membership will account for more rating variance than the perceiver's team membership. Such a result would indicate that the team may be influencing target behavior more than perceiver cognitions and would support the situation strength argument. On the other hand, the social cognitive argument will receive support if the perceiver's team membership accounts for more rating variance than the target's team membership. This would suggest that perceivers' ratings are influenced by cognitions driven by membership in specific teams. The within-team variance of ratings given by perceivers should be restricted, and the between-team variance enhanced, if teams are influencing the perceiver cognitions. In other words, if this effect is large, perceivers' rate their targets very similarly within teams but very differently across teams. Unraveling these effects will inform us about whether peer ratings of performance are largely target-based (thereby contributing to validity) versus whether they may be biased by the perceivers' assessment of team characteristics (thereby detracting from validity, but offering an increasing understanding of where this source of variance originates).

Hypothesis 1: Team membership accounts for (a) target and (b) perceiver variance in peer ratings.

Research Question 1: Are there differences in the amount of perceiver and target variance accounted for by team membership?

Study 1 Methods

Sample

In the current study, we report on empirical analyses of a database of peer feedback ratings within ITPmetrics.com. ITPmetrics is an internet software platform that hosts team-based assessments that are currently used in many post-secondary institutions, primarily in Australia, Canada, Europe, and the USA (O'Neill et al., 2018). The website manages the collection of performance data; that is, team members log into the site to submit performance ratings of others and

Table 1 Variance components in study 1 and study 2 models

Study 1 variance components	Interpretation
Target	Variance in ratings received by targets from all perceivers, controlling for team membership. A large target variance component indicates that all ratings received by a target are highly consistent and that targets are differentiated by their perceivers (i.e., the between-target rating variance is relatively large and the within-target rating variance is relatively small).
Target team Membership	Variance in ratings received by targets specific to each team. A large target team membership variance component indicates that ratings received by a target are consistent within teams, but variable between teams.
Perceiver	Variance in ratings given by perceivers to all targets, controlling for team membership. A large perceiver variance component indicates that all ratings given by a perceiver are highly consistent and that rating patterns are differentiated between perceivers (i.e., the between-perceiver rating variance is relatively large and the within-perceiver rating variance is relatively small).
Perceiver team membership	Variance in ratings given by perceivers specific to each team. A large perceiver team membership variance component indicates that ratings given by a perceiver are consistent within teams, but variable between teams.
Study 2 variance components	Interpretation
Target	Similar definition as above, but team membership was not controlled as targets were only members of one team.
Perceiver	Similar definition as above, but team membership was not controlled as perceivers were only members of one team.
Dyad	Rating variance specific to unique target-perceiver dyads. A large dyad variance component indicates that dyad members give each other consistent ratings that are unique to the ratings given to other team members.
Team	Variance in ratings specific to teams. A large team variance component indicates that all team members give each other consistent ratings and that rating patterns are differentiated between teams (i.e., the between-team rating variance is relatively large and the within-team rating variance is relatively small).

receive their own ratings. Participants in the database were from many disciplines that employ teamwork in post-secondary education and all had given consent to share their results for research purposes. Typical team projects included research reports, presentations, proposals, business ventures, engineering designs, software development, experiential learning, and field projects. The ratings were from students working as part of a team for one or more semesters within the context of a course in higher education. Although precise statistics are not available, instructors in business as well as engineering faculties are likely the most common users of the platform. Typical courses in business that use teamwork involve organizational behavior and entrepreneurship. Typical courses that use teamwork in engineering are design-based and they usually reside in electrical or mechanical engineering fields. However, the system is available to any instructor and therefore the teams may be from a multitude of disciplines. The teams are embedded in courses and likely would have team-based deliverables associated with course grades.

The initial data comprised 116,973 individual ratings of 17,727 targets by 17,729 perceivers. A total of 37.4% of the sample were female and the average age was 22.27 years (*SD* = 5.34). Of the rating targets in the sample, 28.27% were members of only one team, 41.42% were members of two distinct teams, 17.46% were members of three teams, 10.04% were members of four teams, and the remainder were members of up to seven different teams. Although not identical, the team membership characteristics for perceivers in the sample were very similar to that of targets (i.e., there were slight differences because not all of targets and perceivers were members of the same number of teams). We chose to retain data from participants who had been members of four different teams to ensure there were enough teams to estimate team membership variance in ratings, while maximizing the total sample size of perceivers and target for the analyses. This reduced the final sample to 7160 performance rating observations conducted by 567 perceivers for 568 targets who were members of 559 separate teams.[1]

Measures

Round-robin peer ratings of team member behavior were provided on five dimensions developed by Ohland and colleagues (e.g., Loughry, Ohland, & Moore, 2007; Ohland et al., 2012). The ratings were provided on a 5-point scale (*1 = strongly agree* to *5 = strongly disagree*) with multiple behaviors describing the high-scoring end of the continuum (see O'Neill et al., 2019). The dimensions were identified through a review of the teamwork literature, a review of existing peer evaluation forms used by instructors, thematic sorting of behaviors, subject-matter expert review, and multiple factor-analytic studies (Loughry et al., 2007). The following dimensions were adapted for use in the ITPmetrics platform (adapted terms in parentheses): contributing to the team's work (commitment); strong foundation of knowledge, skills, and

[1] A portion of the data in study 1 appears in a published paper (O'Neill et al. 2019). That paper is intended for educators and is focused on how to apply peer ratings in the classroom. As such, it only reports the means, standard deviations, inter-rater reliability, and correlations for each item and the aggregated scale score of the performance rating measure.

abilities (KSAs); interacting with teammates (communication); keeping the team on track (focus); and expecting quality (emphasizing high standards). Inter-rater reliability involving the dimensions assessed on ITPmetrics was found in earlier research to be satisfactory (O'Neill et al., 2019). The internal consistency reliability was .92 in both the target and perceiver samples.

Analysis

The round-robin rating design of this study meant that each target was rated multiple times, and that each perceiver rated multiple targets. Thus, the dependent variable, performance rating, was cross-clustered by both perceivers and targets. Perceivers and targets were also members of four distinct teams. We analyzed the data using cross-classified modeling with Mplus 7.31. The cross-classified model we created partitions the variance of the peer ratings of performance into the following sources: target, target's team membership, perceiver, perceiver's team membership, and residual variance.

In specifying the model, performance rating observations were entered at the within-level and perceivers and targets were the cross-classified level 2 clusters, entered at the between-level. Four dummy variables that represented team membership were created for each of the perceivers and targets (i.e., eight dummy variables in total) and entered as predictors of the performance ratings at the within-level. Team membership was treated as a within-person variable because each individual was a member of multiple teams and the team membership dummies were predicting the ratings given by perceivers or received by targets within each team. Similar to the procedure for multilevel social relations modeling (see Kenny, 2016; Snijders & Kenny, 1999), the dummy variable slopes were allowed to randomly vary at level 2 in both the perceiver and target clusters. The regression slopes of the four perceiver team membership dummy variables were also constrained to be equal, as were the regression slopes for the four target team membership dummy variables. By treating the dummy variable slopes as random effects and imposing equality constraints, all four dummy variables were retained in the model and the effects are interpreted as the extent to which team membership accounted for perceiver and target variance in the performance ratings, while controlling for unobserved individual differences of targets and perceivers (see Table 1 for a description of the variance components and Fig. 1 for a depiction of the model).[2] One limitation of the cross-classified modeling approach we applied is that specific perceiver-target dyad effects were not directly estimated and this source of variance is embedded within the residual term.

Study 1 Results

Table 2 contains the results of the cross-classified analysis described above. After controlling for team membership, perceiver and target characteristics account for similar proportions of model variance at 21.07% and 19.47%, respectively. On the other hand, team membership of the perceiver comprised 24.27% of the total model variance, whereas team membership of the target was 6.13% of the model variance. Hypothesis 1 was supported as the 95% confidence interval excluded zero for both variance estimates (.08 to .10 for perceiver team membership and .02 to .03 for target team membership). The confidence intervals also did not overlap, indicating that team membership has substantially stronger effects on the ratings provided by perceivers than on the behavior of targets; thus, providing an initial answer to the first Research Question.

Study 1 Discussion

This study investigated the extent to which team membership accounts for variance in perceiver ratings and target behaviors, while controlling for all unobserved individual differences of perceivers and targets via the study design feature of having individuals as members of multiple teams. This study provides important methodological and empirical contributions because it allowed us to determine the variance due to team membership that was specific to targets and perceivers, which, to our knowledge, has not been done previously when modeling multiple team memberships. The results indicate that team membership has stronger effects on the cognitions of perceivers than on the behavior of targets, suggesting that the situation strength explanation accounts for a relatively smaller portion of the between-team variance in individual performance ratings, and that a social-cognitive perspective might be a more accurate representation of the rating processes that are contributing to team effects in individual performance ratings.

To elaborate on the above, if the teams in this research had a powerful impact on individual performance behavior, and hence, performance ratings, then the target's team membership should have accounted for a substantial portion of the total model variance. Or, if all teams restricted target behavior in similar ways, the overall proportion of model variance attributable to targets would be reduced as compared to variance attributable to team membership. Instead, the target and perceiver components accounted for very similar proportions of model variance and the proportion of variance accounted for by the perceiver's team membership was four times larger than that of the

[2] This differs from the usual application of dummy variables, where the model would contain one fewer ($k - 1$) dummy variables than there are teams. In those types of models, a dummy variable slope is interpreted as the difference in the y-variable between the team represented by the dummy and the reference category, or the team that was excluded from the model.

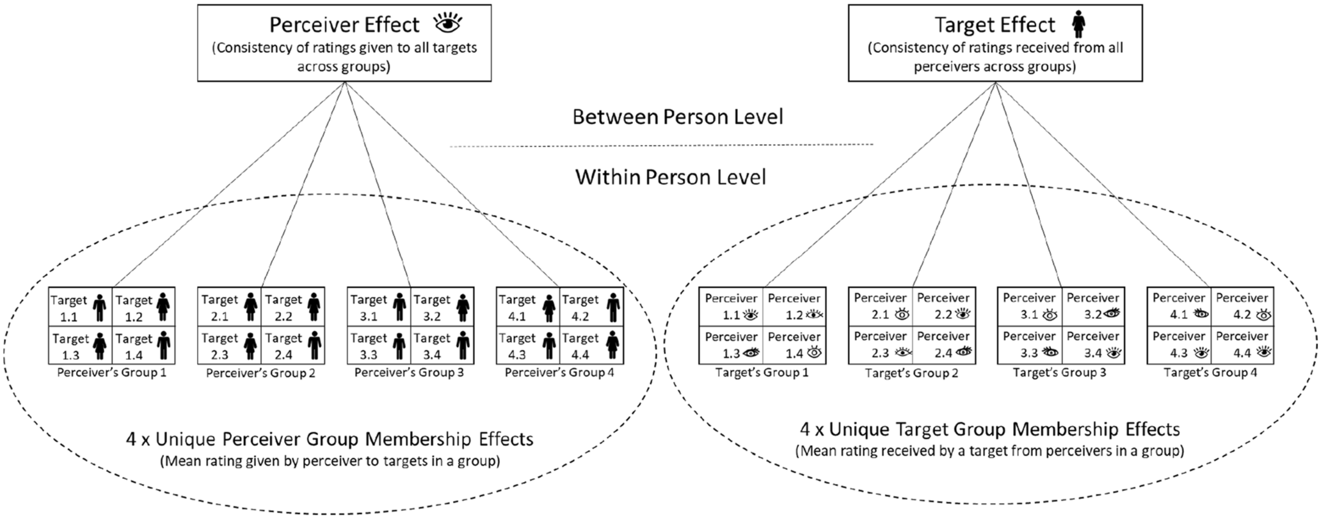

Fig. 1 Depiction of the study 1 research design. The sample consisted of 567 perceivers and 568 targets who were members of four unique teams. Every rating associated with each perceiver and target were included in the analysis regardless of whether the rating was given (or received) from an individual who was included in the main perceiver or target sample. For example, we retained every rating that perceiver A provided to targets 1.1 through 1.4, even though target 1.4 was excluded from the main target sample because she did not participate in four distinct teams. The design allowed us to account for the cross-classified nature of the data (i.e., some participants were a perceiver and a target) and to provide the most reliable estimates of perceiver, target, and team membership effects by retaining all of the ratings associated with each perceiver and target in the main sample

target. Together, these results suggest that teams play a larger role in influencing how perceivers *judge* behavior in teams than they do in actually influencing the behavior of team members. In other words, these findings suggest that teams create "situations" that influence perceivers' social judgments, rather than individual differences in performance-related behavior (the latter of which would be the assumed meaning of these ratings if used in an HRM system). Although the results of study 1 provided more precise estimates about the degree to which team membership accounts for perceiver and target variance, the study design did not allow us to make inferences about how teams influence perceiver cognitions. Thus, we conducted a second study to examine how a theoretically relevant team-level construct may act to influence or distort perceiver judgments.

Study 2

The results of study 1 indicated that teams have a stronger effect on perceiver judgments than they do on target behaviors. As discussed previously, one explanation for this finding is that the team context serves as a heuristic that influences, and possibly distorts, perceivers' judgments about their peers' behavior. The team context may inhibit perceivers' tendencies to detect and utilize certain behavioral cues, particularly if the cues are inconsistent with the team climate. This line of reasoning is consistent with a large corpus of research showing that performance appraisal is often inaccurate and influenced by factors other than the behaviors being evaluated. Specifically, ratings are often distorted due to error-prone memory and information processing (e.g., Hilbert, 2012; Ilgen, Barnes-Farrell, & McKellin, 1993). And as noted by

Table 2 Cross-classified modeling results

			95% confidence interval		
Variance component	σ^2 estimate	Posterior SD	Lower	Upper	% of model variance
Target	.073	.007	.060	.087	19.47
Target team membership	.023	.003	.018	.029	6.13
Perceiver	.079	.009	.064	.097	21.07
Perceiver team membership	.091	.005	.081	.101	24.27
Residual	.109	.002	.104	.114	29.07

The model contained 7160 performance rating observations by 567 perceivers for 568 targets on 559 separate teams

Erez, Schilpzand, Leavitt, Woolum, and Judge (2015), "individuals generally face some degree of uncertainty when rating and rely upon person impressions to 'fill in the gaps' (Wherry & Bartlett, 1982), and may also reweight performance criteria to justify decisions reflecting their own social preferences or biases" (p. 1766). The results of study 1 extend the logic of appraisal heuristics to the team level and suggest that teams may influence social cognition by acting on impressions of individual team members, providing some of the "gap-filling" information used to derive performance judgments. That is, perceivers use the team context as a heuristic to derive general impressions of a target, which affects how subsequent target behavior is recalled and interpreted. Thus, we believe that the specific characteristics of the team context may be responsible for driving idiosyncratic perceiver variance, and if this is true, it would deepen our understanding of perceiver cognitions, but also call attention to developing appropriate interpretations of peer ratings in HRM applications. As we describe below, we identified psychological safety as particularly relevant to perceiver cognitions because it has a strong affective component and may cause perceivers to overlook individual behavior.

The Effects of Psychological Safety on Perceiver Judgments

Psychological safety is an emergent state that is defined as "a shared belief that the team is safe for interpersonal risk taking" (Edmondson, 1999, p. 354). It is derived mainly from perceptions of peer and organizational support and has a powerful influence on team member attitudes (Frazier, Fainshmidt, Klinger, Pezeshkan, & Vracheva, 2017). Members of teams with higher levels of psychological safety likely perceive that their standing within the team is secure and that they can freely express themselves without fear of reprimand.

Given the tendency for individuals to view other team members as a generalized category (Savitsky et al. 2005), higher levels of psychological safety could increase the likelihood that perceivers adopt a cognitive heuristic implying that *any* given team member is an effective and supportive contributor, irrespective of that team member's actual behavior. As stated earlier, psychological safety may contribute to social categorization processes (Srull & Wyer, 1989), where people use the team context, in addition to small samples of others' behaviors, to derive general impressions of individual team members. Thus, the team context may "fill the gaps" in performance and OCB ratings for unobserved behavior. If this argument is accurate, perceptions of the team context may produce social cognitions that are only weakly related to the target's actual behaviors and lead to inaccurate evaluations of those behaviors. A positive association between team-level psychological safety and the team intercepts of individual performance and OCB ratings would be consistent with this perspective. That is, the more psychologically safe the team context is, the more favorable the peer ratings of individual behavior in that team. However, such a pattern of results would also be consistent with the notion that psychological safety improves team functioning, that is, it truly increases performance and OCB among all team members. To the extent that this occurs, an association of psychological safety with peer ratings would reflect accurate ratings of performance.

Fortunately, the distinction above can be disentangled through social relations modeling. Specifically, the degree to which individual perceptions of psychological safety account for perceiver- and target-level variance in individual ratings reveals more about social cognition than does the effects of team-level psychological safety on average team ratings. It is arguably the *perceptions* of psychological safety held by a perceiver (notwithstanding the perceptions of psychological safety held by other members of the team) that will directly affect how targets are evaluated by that perceiver. Results showing that individual psychological safety perceptions account for more perceiver than target rating variance would imply that perceivers are using psychological safety, rather than actual target behavior, to derive their judgments. Although psychological safety is normally operationalized as a team construct, Frazier et al. (2017) demonstrated that it is identically homologous, meaning that "relationships within a construct's nomological network will be identical in magnitude and direction across levels of analysis" (pp. 144–145). We therefore utilized perceivers' assessments of psychological safety at the individual level for the purposes of evaluating its effect on both perceiver and target variance components of task performance and OCB ratings, while also accounting for the effects of team-level psychological safety.

Hypothesis 2: Perceiver assessments of psychological safety account for perceiver variance in (a) task performance and (b) OCB ratings.

Study 2 Methods

Sample

The sample consisted of 217 students (46.3% male) from three undergraduate classes and three MBA classes of a Western Canadian university. The undergraduate classes occurred over a 3-month semester and the MBA classes were structured as an intensive 3-week course. Both types of classes had 39 h of in-class time (i.e., contact hours). Participation was voluntary and students who chose to participate were given course credit. The response rate was 91.5% as 236 students were registered in the six classes. Students were randomly assigned to teams of four to six members and the team members worked closely with each other throughout the course. The teams were required to deliver a presentation at the midpoint of the course, complete an online leadership and team simulation, and submit a report about their simulation experience at the end of the course. Team members were also required to interact in nearly every class as the teams participated in competitive trivia sessions where the top scoring teams won additional bonus marks at the end of the course.

The students chose to participate in the research within the first week of the course. Participants completed an online personality questionnaire at the beginning of the course, completed three questionnaires about team states and processes throughout the course, and rated the individual task performance and citizenship behaviors of each team member at the end of the course. For this research, we used the psychological safety measure gathered during the third wave of data collection, which occurred when two-thirds of the course was complete (i.e., 2 months into the semester for undergraduate students and 2 weeks into the course for MBA students). This timing allowed students to become acquainted with their team members and develop a stable impression of their team's emergent states and processes. Because the peer task performance and OCB ratings were collected immediately after the last scheduled class, the timing also allowed for temporal separation between ratings of the team construct and outcome measures. All of the survey measures described below were rated on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*).

Measures

Psychological Safety Edmondson's (1999) seven-item psychological safety measure was used in this research. Two example items include, "If you make a mistake in this group, it is often held against you," and "Members of this group are able to bring up problems and tough issues." Internal consistency reliability was adequate at both levels of analysis as the individual-level ω statistic was .71 and the team-level statistic was .98 (see Geldhof, Preacher, & Zyphur, 2014). The ICC(1) value of .35 indicates that a substantial proportion of the variance was between teams.

Individual Task Performance and OCB Participants rated the task performance of each team member using a three-item scale developed by Griffen, Neal, and Parker (2007). The items included, "completed individual tasks well," "completed individual tasks using appropriate methods/procedures," and "ensured individual tasks were completed properly." The Cronbach alpha was .91. OCB was measured with Lee and Allen's (2002) eight-item OCB-Individual scale. Two sample items include "helped others who have been absent," and "gave up time to help others who had work or non-work problems." The Cronbach alpha for this scale was .92.

Team-Level Control Variable To determine if psychological safety accounted for variance in individual ratings over and above team performance, we created a team performance composite score that consisted of two measures: team presentation grade and team simulation score. All teams had delivered a presentation to the class by the midpoint of the course and received feedback about their performance before completing the team construct surveys. The teams also participated in the Harvard Business Publishing online Everest Simulation, which required teams to overcome unexpected challenges while negotiating conflicting individual and team goals to complete the simulation successfully. Teams were scored based on the number of goals they completed and team members participated in an extensive debrief session that provided them with an opportunity to conduct a "post-mortem" about their interactions and compare their performance to other teams in the class. Each team score was standardized within the respective class and the two standardized scores were averaged to create an overall performance composite.

Analysis

This was a round-robin design with missing data, so we applied the multilevel approach to social relations modeling (Kenny, 2016; Snijders & Kenny, 1999) using Mplus 7.31. To conduct the analyses, we specified two separate multilevel models: one for task performance and the other for OCB. Each task performance or OCB observation was entered as the level 1 dependent variable. The performance or OCB observations were clustered within unique dyads at level 2 and the dyads were clustered within teams at level 3. We created dummy variables for each unique perceiver and target within the teams and regressed the task performance or OCB ratings on these dummy variables at level 1. The slopes of the dummy variables for the perceiver were constrained to be equal as were the slopes for the target dummy variables. The perceiver and target slopes were freely estimated at the team level (i.e., level 3). As in study 1, all of the dummy variables are retained in the model because the slopes were constrained to be equal and allowed to randomly vary at a higher level. The perceiver and target slopes were also allowed to covary at level 3. This modeling approach allows for the estimation of perceiver, target, dyad, and team variance components as well as the perceiver-target covariance (Fig. 2 shows the design of study 2).

To investigate the research questions, we conducted a three-step model building process and examined changes in variance components at each step. First, we estimated models without any covariates to partition the rating variance into each component. Second, we entered the team performance control variable at level 3. Third, we aggregated the psychological safety variable to level 3 by calculating the team means on this variable. Then, we regressed the team-level intercept of

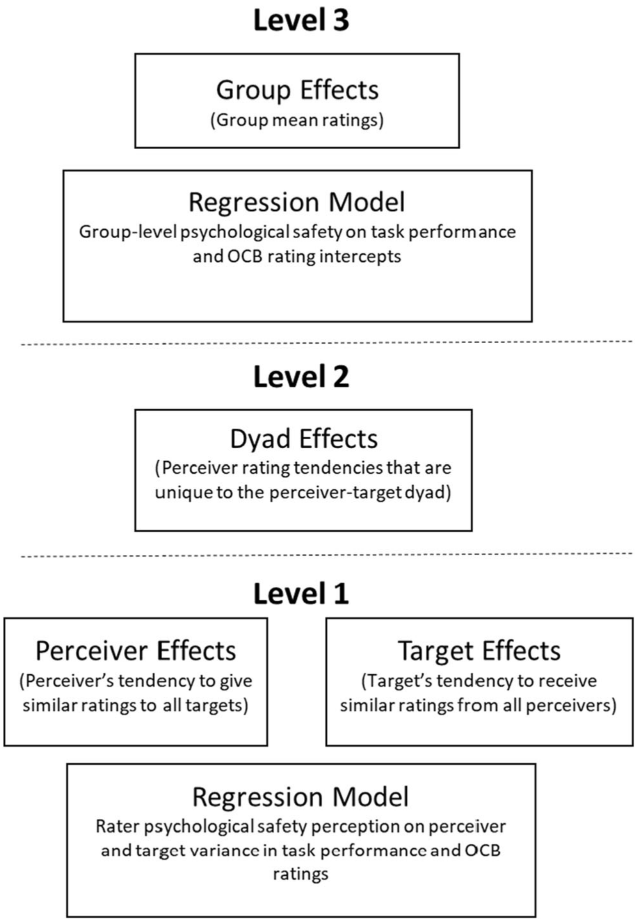
Level 3
Group Effects
(Group mean ratings)
Regression Model
Group-level psychological safety on task performance and OCB rating intercepts
Level 2
Dyad Effects
(Perceiver rating tendencies that are unique to the perceiver-target dyad)
Level 1
Perceiver Effects
(Perceiver's tendency to give similar ratings to all targets)
Target Effects
(Target's tendency to receive similar ratings from all perceivers)
Regression Model
Rater psychological safety perception on perceiver and target variance in task performance and OCB ratings

Fig. 2 Depiction of study 2 research design

task performance (OCB) ratings on team-level psychological safety at level 3 and also entered the perceivers' own ratings of psychological safety as a fixed-effect predictor variable at level 1. Because the hypothesis focused on each perceiver's idiosyncratic perceptions of the team constructs, we group-mean centered all the level 1 covariates in the final step. This process allowed us to estimate the degree to which the added covariates accounted for variance in each component of the model.

Study 2 Results

The correlations among the study variables are reported in Table 3, while Table 4 reports the results of the social relations modeling analyses for the task performance and OCB ratings. In Table 4, model 1 for task performance ratings indicates that target and perceiver effects accounted for the majority of model variance at 30.88% and 22.71%, respectively. The unique characteristics of rating dyads accounted for 5.58% and the team context accounted for 6.97% of the variance. In this sample, the team variance in peer ratings was somewhat smaller than that reported in the first study and past research. As we discuss subsequently, this may be due to the nature of performance behaviors that were rated in this study or design differences as participants in the current research were members of only one team. The second model of Table 4 contains the results when the team performance covariate was added to the model. The association between team performance and the task performance intercept was marginally significant ($\gamma = .11$, $p = .059$) and this effect accounted for 17.17% of the team variance.

Model 1 for OCB ratings shows that target effects accounted for 20.09% of the model variance, perceiver characteristics accounted for 38.46% of the variance, the relationship between members of the rating dyad accounted for 5.77%, and the team accounted for 13.89%. Model 2 shows that the team performance composite was not significantly associated with the level 3 intercept and it accounted for a relatively small amount of team variance (3.01%).

Hypothesis 2 states that perceiver judgments about psychological safety account for perceiver variance in peer ratings of

Table 3 Descriptive statistics and correlations among study 2 variables

	M	SD	1	2	3	4	5	6
Level 1 variables								
1. Perceiver psychological safety	4.07	.51	(.71)					
2. Task performance rating	4.09	.70	.28**	(.91)				
3. Organizational citizenship rating	3.71	.68	.29**	.64**	(.92)			
Level 3 variables								
4. Team performance composite	− .13	.76	–	–	–	–		
5. Psychological safety	4.03	.35	–	–	–	.12	(.98)	
6. Task performance rating	4.03	.34	–	–	–	.33*	.48**	
7. Organizational citizenship rating	3.67	.36	–	–	–	.15	.54**	.62**

$N = 679$ for level 1 correlations, $N = 46$ for level 3 correlations. Reliability statistics are reported in the diagonals where appropriate. Multilevel ω values are reported for psychological safety, Cronbach's α are reported for level 1 task performance and organizational citizenship behavior. Correlations were calculated with MPlus and account for the nested structure of the data.

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

Table 4 Effects of psychological safety on task performance and OCB ratings

	Task performance						OCB					
	Model 1		Model 2		Model 3		Model 1		Model 2		Model 3	
Variance components	σ^2 estimate	% of model variance	σ^2 estimate	R^2 for each component	σ^2 estimate	R^2 for each component	σ^2 estimate	% of model variance	σ^2 estimate	R^2 for each component	σ^2 estimate	R^2 for each component
Target	.155	30.88	.155	.000	.151	.026	.094	20.09	.094	.000	.095	− .011
Perceiver	.114	22.71	.114	.000	.087	.237	.180	38.46	.180	.000	.158	.122
Dyad	.028	5.58	.029	− .036	.028	.000	.027	5.77	.027	.000	.025	.074
Team	.035	6.97	.029	.171	.002	.943	.065	13.89	.063	.031	.007	.892
Residual (within-person)	.170	33.86	.170		.170		.102	21.79	.102		.104	
Perceiver-target covariance	− .006		− .007		− .013		− .026		− .026		− .025	
Model Variables			Coefficient		Coefficient				Coefficient		Coefficient	
γ_{00}: intercept			4.06 (.05)**		4.05 (.04)**				3.69 (.06)**		3.69 (.04)**	
γ_{10}: individual psychological safety					.36 (.07)**						.37 (.08)**	
γ_{01}: team performance			.11 (.06)†		.07 (.05)				.07 (.07)		.03 (.04)	
γ_{02}: team psychological safety					.61 (.09)**						.71 (.12)**	

$N = 679$ performance observations nested in 441 dyads within 46 teams. Negative R^2 values are the result of very small changes in variance estimates between models. Standard errors are reported in parentheses for the variable coefficient estimates.

$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$

task performance and OCB. Hypothesis 2a was supported because model 3 for task performance shows that perceiver psychological safety judgments were positively associated with the task performance ratings given to peers ($\gamma = .36$, $p < .001$), accounting for 23.68% of perceiver variance. Team-level psychological safety was strongly associated with the level-3 intercept for task performance ($\gamma = .61$, $p < .001$) and accounted for substantial team-level variance (77.14%) over and above the team performance composite variable. As shown in model 3 for OCB, Hypothesis 2b also received support because perceiver psychological safety judgments were positively associated with the OCB ratings given to peers ($\gamma = .37$, $p < .001$), accounting for 12.22% of perceiver variance. Team-level psychological safety was significantly associated with the OCB intercepts at level 3 ($\gamma = .71$, $p < .001$) and accounted for 86.15% of team-level variance over and above the team performance composite.

While the changes in variance components described above suggest that psychological safety influenced perceivers more than targets, we conducted additional tests of this proposition. First, we calculated perceiver and target effects for round-robin designs as described by Kenny, Kashy, and Cook (2006) for the set of participants who both gave *and* received performance ratings ($n = 165$) and correlated the effects with the individuals' psychological safety perceptions (see Greguras et al., 2001 for a similar example of this approach). Providing further support to Hypothesis 2, the correlations between psychological safety and the perceiver effects for task performance and OCB were both significant ($r = .27$, $p = .001$ and $r = .23$, $p = .003$ for the task performance and OCB perceiver effects, respectively). On the other hand, the correlations between psychological safety and the target effects were not significant ($r = .02$, $p = .800$ and $r = .05$, $p = .503$ for the task performance and OCB target effects, respectively). This evidence supports the notion that psychological safety perceptions produced more consistency in the ratings given by perceivers than received by targets.

We also followed Keeler et al.'s (2019) suggestion to test situation strength by examining variance restriction in perceiver and target effects for teams that were above versus teams that were below the mean level of psychological safety. Levene's test for equality of variances showed that the team level of psychological safety did not affect the variance in perceiver effects for either task performance or OCB ($F(1,163) = .08$, $p = .772$ and $F(1,163) = .02$, $p = .898$ for task performance and OCB, respectively). Similarly, the variance differences in target effects did not reach the level of statistical significance ($F(1,163) = 1.97$, $p = .162$ and $F(1,163) = 2.26$, $p = .134$ for task performance and OCB, respectively). These results indicate that psychological safety was likely not creating a strong situation that restricted variance in either perceiver or target effects.

Study 2 Discussion

The results of study 2 indicated that perceiver psychological safety judgments accounted for more perceiver than target variance and team-level psychological safety accounted for much of the team-level variance over and above the team performance composite variable. Further tests showed that psychological safety perceptions were correlated with the perceiver effects, but not the target effects, and that psychological safety did not appear to restrict variance in either the perceiver or target effects. Thus, when these findings are considered together with the study 1 results, it is reasonable to conclude that team-level variance in peer ratings is primarily due to the influence of the team on the rating heuristics of the perceivers rather than compelling team members to behave consistently.

General Discussion

This research was among the first to examine the extent to which multiple team contexts, and specific perceptions of those team contexts, influence perceiver versus target variance in peer ratings of individual performance and OCB. We conducted two studies to test different theoretical perspectives—situation strength and social cognition—that may explain how the team context influences individual ratings. The first study employed a large cross-classified sample to provide reliable estimates of team variance specific to perceivers and targets who were each members of four separate teams. The results showed that team membership accounted for nearly four times as much perceiver variance as it did target variance, revealing an important contribution to the theory and practice of peer performance ratings. From a practical perspective, the findings suggest that absolute levels of peer performance ratings for members of different teams may not be comparable at face value, which has implications for organizations using peer ratings for administrative or developmental purposes (e.g., 360-degree assessments). Theoretically, the findings deepen our understanding of perceiver effects in team contexts by suggesting that teams may have a stronger influence on how perceivers attend to and evaluate performance information than on the targets' actual performance behaviors, indicating that social cognitive arguments may provide a better explanation of how teams influence peer performance ratings than situation strength.

The second study informed how perceivers' idiosyncratic understanding of psychological safety influenced their ratings of their peers. The results suggested two possible conclusions. First, team-level psychological safety accounted for substantially more between-team variance in peer performance and

OCB ratings than actual team performance, suggesting that the peer ratings were influenced more by the social context of the team than by the behavior of the team members. When considered together with the fact that assignment into teams was exogenous, it is unlikely that a team's ability to attract, select, and retain higher-performing members (i.e., attraction-selection-attrition mechanisms) might explain between-team variance in performance ratings in this sample. Second, perceptions of psychological safety accounted for more perceiver than target variance and were correlated with perceiver effects, but not target effects, suggesting that characteristics of the team, as understood by the perceiver, have a strong effect on peer ratings. This is further evidence that situation strength does not account for much variance in peer ratings of performance and OCB. Moreover, it could indicate that perceivers are using psychological safety as a heuristic or to "fill in gaps" for unobserved behavior and adds to an increasing body of knowledge which suggests that ratings may not reflect a substantial amount of "true" performance variance. Researchers need to continue to investigate this issue given the criticality of behavioral ratings for research (as criteria) and for practice. Below, we identify the theoretical and practical implications of these findings and we suggest future research directions to advance our understanding of the role of peer ratings in HRM contexts.

Theoretical Implications and Future Research

Given that social cognitive processes, rather than situation strength, appear to be a source of team influences on individual ratings, theories of social judgment and cognition require further development to incorporate the team context. We describe how our results may apply to two specific theories of social judgment, while recognizing that there is an expansive corpus of research in the fields of social perception, judgment, and decision-making and the applications are certainly not limited to our suggestions.

The first suggestion is to further incorporate the effects of team context within theories of social categorization. As discussed previously, Srull and Wyer's (1989) theory suggests that perceivers form rapid judgments about targets based on behaviors observed during initial interactions and then assign targets to pre-existing social categories or prototypes (e.g., ambitious, outgoing, pessimistic). The target's future behaviors are subsequently anchored to the category, thereby profoundly affecting the perceiver's interpretations. It is not unlikely, based on the current study's findings, that the team context influences how perceivers actually assign targets to categories. For example, people in cohesive and productive teams may generally assume that all team members must be prosocial and hard-working, given the positive team context. Moreover, perceivers may make assumptions about unknown or unobserved information based on behaviors that are prototypical for that category. Bias occurs when judges make incorrect assumptions about the target's behavior based primarily on category membership. Future research could focus on explicitly defining a taxonomy of social categories and specifically testing how the team context influences perceivers' social categorization processes and their judgments of others.

A second theoretical implication is that team factors could be incorporated into the fast-and-frugal tree (FFT) approach to judgment and decision-making (e.g., Luan & Reb, 2017). This theory suggests that judgments are based on series or "trees" of non-compensatory steps, where people make a number of successive binary yes/no decisions about each criterion within a decision tree and will cease to consider an alternative if the criterion does not meet the minimum standard at one of the decision steps. As applied to the context of peer evaluations, it is possible that team constructs, such as psychological safety, are part of perceiver judgment trees and people may exit the tree to provide a positive evaluation if they determine the team construct is at a sufficient level.

Comparison of Variance Components

The findings also revealed that the magnitude of variance components differed between the task performance and OCB rating criteria. Targets accounted for more variance than perceivers in task performance, whereas the pattern was reversed for OCB. Teams also accounted for more variance in OCB than task performance. These findings may reflect the observation that task performance is more clearly defined and observable than OCB; thus, peer evaluators have less need to rely on rating heuristics or the team context to make inferences about targets' performance behaviors. These explanations are consistent with Dierdorff and Morgeson's (2009) findings that ratings of task descriptors had higher inter-rater reliability than ratings of other behavioral descriptors, which they attributed, in part, to higher observability of task behaviors. Moreover, observers may make different types of attributions about the motives for OCB (e.g., Halbesleben et al., 2010) and the distinctiveness of different types of OCB behaviors are conditional on team performance (Oh, Chen, & Sun, 2015), which may explain both the larger perceiver and team variance components for OCB as compared to task performance. Future research should continue to explore why there are discrepancies in team- and perceiver-level variance for different performance dimensions and examine if frame-of-reference training for OCB can reduce team variance as it has for task performance ratings (Loignon et al., 2017).

Methodological and Practical Implications

The findings of this research suggest that team context influences perceptions of task performance and OCB enough to produce substantial between-team variance in peer ratings of these constructs. The ratings in the current research were not seen by targets, which may have mitigated the team-level effects on perceiver judgments. When ratings are seen by targets and impact important work outcomes (i.e., ratings are used for administrative decisions), perceivers are likely to engage in conscious distortion that may be partially based on team climate and norms (e.g., Levy & Williams, 2004; Murphy & Cleveland, 1995). Therefore, it is possible that team- and perceiver-level variance components will be even larger when peer ratings are used for developmental or administrative purposes that will be seen by targets (e.g., ratings from 360-degree assessments). This leads to somewhat pessimistic conclusions about the accuracy of peer ratings in applied contexts and may suggest that ratings cannot be compared across people working in different teams. However, this conclusion may be somewhat premature until more research is conducted and, following Greguras et al.'s (2001) suggestions to create person-level SRM indices, we offer two ideas about how researchers and practitioners may mitigate the effects of bias or irrelevant variance on peer performance and OCB ratings.

The first suggestion involves a relatively straightforward approach for parsing irrelevant variance out of individual ratings. In the context of social relations modeling, observed ratings are calculated as follows:

$$\begin{aligned}\text{observed rating} = {} & \text{team effect} + \text{perceiver effect}\\ & + \text{target effect} + \text{relationship effect},\end{aligned}$$

where the team effect is the mean rating for the team, the perceiver effect is the tendency for the perceiver to give similar ratings to all targets, the target effect is the tendency for the target receive similar ratings from all perceivers, and the relationship effect is the perceiver's unique rating of the target (see Christensen & Kashy, 2012). Bonito and Kenny (2010) explain how to calculate each of these effects for individual targets and perceivers as well as estimate the reliability of each component in the equation. A practitioner interested in comparing performance ratings of people in different teams could calculate the target effect for each person, which is essentially a performance rating with the team, perceiver, and relationship effects parsed from the score. The practitioner could also estimate the standard error of measurement of targets by calculating the target effect reliability and then applying this estimate and the standard deviation of all target effects to create confidence interval around each score (see Furr & Bacharach, 2014). If the confidence intervals around the two target effects overlap, then the ratings would be deemed to be statistically equivalent and should not be used as justification to make different developmental or administrative decisions about the two individuals.

Another possible approach to mitigate the effects of team constructs on peer ratings of individual performance and another important method to explore in future research is the relative percentile approach (e.g., Goffin et al., 2009). This technique involves instructing perceivers to explicitly compare an individual's behavior (e.g., performance, citizenship) to the behavior of other individuals *in the same population* (e.g., in relation to all members of student work teams at the university). Thus, to the extent that perceivers are able to accurately judge the percentile location of their fellow team members in the population, the relative percentile ratings could reduce team and perceiver variance components. If team context represents spurious rather than performance-relevant source of variance in individuals' peer ratings, a relative percentile method may be advantageous over traditional Likert-type or graphic rating scales.

Limitations

Although there are some positive aspects of the designs, including a large sample of individuals who were members of multiple teams in study 1 and members who were randomly assigned to teams working on standardized tasks in study 2, there are some limitations that require consideration. One limitation is that both studies were conducted with samples of student work teams and there may be important differences between the student and employee teams. First, students did not receive any feedback about the ratings given to them by peers, nor did the ratings impact important outcomes for the students (e.g., grades). In a work context, organizations will likely use peer ratings for developmental feedback or to make administrative decisions, which may impact rating leniency (per Greguras et al., 2003) and how peers behave toward each other. Second, the student teams received explicit rewards for team performance (i.e., simulation scores, grades) and the performance management structures in some organizations may not track or reward team performance to the same extent. Finally, unlike many employee teams, the student teams had a relatively short lifecycle, which could further affect how much effort teams put into resolving conflict and addressing performance deficiencies. Research should continue to examine these important boundary conditions.

A second limitation of this research is that we did not have more objective performance metrics to compare to the peer ratings. As Kenny and Albright (1987) argued, a substantial amount of target variance is a necessary, but not sufficient condition to determine that ratings are accurate without an objective comparator. It is also possible that variance in perceiver ratings was due to some perceivers providing more accurate ratings than others. Future research should seek to address these issues by comparing peer ratings with more objective behavioral data to further determine how ratings reflect actual behavior irrespective of the team context.

Conclusions

As organizations continue to leverage teamwork to enhance firm performance, peer evaluations will likely continue to underpin important HR processes related to employee development and personnel decision-making. Similarly, as post-secondary institutions continue to use teamwork in courses, educators may also use peer evaluations to assess the performance of individuals within student work teams (e.g., Ohland et al., 2012). The current research highlights the substantive effects of team context on peer evaluations and, in particular, that teams have a much stronger effect on perceiver judgments than target behavior. Consequently, practitioners and educators need to proceed cautiously when comparing evaluations between members of different teams. Moreover, researchers should seek to account for team context in theories of social cognition, perception, and judgment. Peer ratings will likely continue to play an important role in performance assessment and feedback, and therefore research needs to uncover how to make the ratings as useful as possible.

Data Availability The MPlus analysis scripts and outputs are available via the following URL: https://osf.io/9turm/

References

Aguinis, H. (2013). *Performance management* (3rd ed.). Essex: Pearson.

Allen, N. J., & O'Neill, T. A. (2015). The trajectory of emergence of shared group-level constructs. *Small Group Research, 46*, 352–390.

Bamberger, P. A. (2007). Competitive appraising: A social dilemma perspective on the conditions in which multi-round peer evaluation may result in counter-productive team dynamics. *Human Resource Management Review, 17*, 1–18.

Behfar, K. J., Friedman, R., & Oh, S. H. (2016). Impact of team (dis) satisfaction and psychological safety on performance evaluation biases. *Small Group Research, 47*, 77–107.

Bonito, J. A., & Kenny, D. A. (2010). The measurement of reliability of social relations components from round-robin designs. *Personal Relationships, 17*, 235–251.

Christensen, P. N., & Kashy, D. A. (2012). Using the social relations model to understand interpersonal perception and behavior. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 3, pp. 425–437). Washington, DC: American Psychological Association.

Dierdorff, E. C., & Morgeson, F. P. (2009). Effects of descriptor specificity and observability on incumbent work analysis ratings. *Personnel Psychology, 62*, 601–628.

Dierdorff, E. C., & Surface, E. A. (2007). Placing peer ratings in context: Systematic influences beyond ratee performance. *Personnel Psychology, 60*, 93–126.

Dominick, P. G., Reilly, R. R., & McGourty, J. W. (1997). The effects of peer feedback on team member behavior. *Group & Organization Management, 22*, 508–520.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*, 350–383.

Ellington, J. K., & Wilson, M. A. (2017). The performance appraisal milieu: A multilevel analysis of context effects in performance ratings. *Journal of Business & Psychology, 32*, 87–100.

Erez, A., Schilpzand, P., Leavitt, K., Woolum, A. H., & Judge, T. A. (2015). Inherently relational: Interactions between peers' and individuals' personalities impact reward giving and appraisal of individual performance. *Academy of Management Journal, 58*, 1761–1784.

Favero, J. L., & Ilgen, D. R. (1989). The effects of ratee prototypicality on rater observation and accuracy. *Journal of Applied Social Psychology, 19*, 932–946.

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology, 70*, 113–165.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91.

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.

Greguras, G. J., Robie, C., & Born, M. P. (2001). Applying the social relations model to self and peer evaluations. *Journal of Management Development, 20*, 508–525.

Greguras, G. J., Robie, C., Schleicher, D. J., & Maynard, G. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1–21.

Griffen, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*, 327–347.

Halbesleben, J. R. B., Bowler, W. M., Bolino, M. C., & Turnley, W. H. (2010). Organizational concern, prosocial values, or impression management? How supervisors attribute motives to organizational citizenship behavior. *Journal of Applied Social Psychology, 40*, 1450–1489.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*, 211–237.

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.

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.

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

Keeler, K. R., Kong, W., Dalal, R. S., & Cortina, J. M. (2019). Situational strength interactions: Are variance patterns consistent with the theory? *Journal of Applied Psychology, 104*, 1487–1513.

Kenny, D. A. (2016). *Estimation of the SRM using specialized software.* Unpublished manuscript. Retrieved from: http://davidakenny.net/srm/srm.htm.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102*, 390–402.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*, 1060–1072.

Lee, K., & Allen, N. J. (2002). Organizational citizenship behavior and workplace deviance: The role of affect and cognitions. *Journal of Applied Psychology, 87*, 131–142.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork process: Towards a better understanding of the dimensional structure and relationships with team effectiveness criteria. *Personnel Psychology, 61*, 273–307.

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.

Loignon, A. C., Woehr, D. J., Thomas, J. S., Loughry, M. L., Ohland, M. W., & Ferguson, D. M. (2017). Facilitating peer evaluation in team contexts: The impact of frame-of-reference rater training. *Academy of Management Learning & Education, 16*, 562–578.

Loughry, M. L., Ohland, M. W., & Moore, D. D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement, 67*, 505–524.

Luan, S., & Reb, J. (2017). Fast-and-frugal trees as noncompensatory models of performance-based personnel decisions. *Organizational Behavior and Human Decision Processes, 141*, 29–42.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121–140.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252–283.

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

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*, 48–65.

O'Neill, T. A., Deacon, A., Gibbard, K., Larson, N., Hoffart, G., Smith, J., & Donia, M. (2018). Team dynamics feedback for post-secondary student learning teams. *Assessment and Evaluation in Higher Education, 43*, 571–585.

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings. *Organizational Research Methods, 15*, 436–462.

O'Neill, T. A., Larson, N., Smith, J., Deng, C., Donia, M., Rosehart, W., & Brennan, R. (2019). Introducing a scalable peer feedback system for learning teams. *Assessment and Evaluation in Higher Education, 44*, 848–862.

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

O'Neill, T. A., & Salas, E. (2018). Creating high performance teamwork in organizations. *Human Resource Management Review., 28*, 325–331. https://doi.org/10.1016/j.hrmr.2017.09.001.

Oh, S. H. D., Chen, Y., & Sun, F. (2015). When is a good citizen valued more? Organizational citizenship behavior and performance evaluation. *Social Behavior and Personality: An International Journal, 43*, 1009–1020.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self-and peer evaluation. *Academy of Management Learning & Education, 11*, 609–630.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology, 98*, 114–133.

Savitsky, K., Van Boven, L., Epley, N., & Wight, W. M. (2005). The unpacking effect in allocations of responsibility for group tasks. *Journal of Experimental Social Psychology, 41*, 447–457.

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

Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships, 6*, 471–486.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96*, 58–83.

Waldman, D. A., Yammarino, F. J., & Avolio, B. J. (1990). A multiple level investigation of personnel ratings. *Personnel Psychology, 43*, 811–835.

Wherry Sr., R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. Personnel Psychology, 35(3), 521-551.