ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/feduc.2020.00049

Team Health and Project Quality Are Improved When Peer Evaluation Scores Affect Grades on Team Projects

Thomas A. ONeill[1], Melissa Boyce[1] and Matthew J. W. McLarnon[2]*

[1] University of Calgary, Calgary, AB, Canada, [2] Mount Royal University, Calgary, AB, Canada

OPEN ACCESS

Edited by:
Meryem Yilmaz Soylu,
University of Nebraska-Lincoln,
United States

Reviewed by:
Michaéla C. Schippers,
Erasmus University, Netherlands
Miguel Ángel Conde,
Universidad de León, Spain

***Correspondence:**
Thomas A. ONeill
toneill@ucalgary.ca

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

Received: 07 November 2019
Accepted: 14 April 2020
Published: 15 May 2020

Citation:
ONeill TA, Boyce M and
McLarnon MJW (2020) Team Health
and Project Quality Are Improved
When Peer Evaluation Scores Affect
Grades on Team Projects.
Front. Educ. 5:49.
doi: 10.3389/feduc.2020.00049

The use of team projects is common in higher education. Teamwork offers an avenue to help students learn to collaborate and develop the interpersonal skills needed for career success. However, student teams are not always effective, which may undermine learning, growth, and development. In the current research, we integrate accountability, valence, motivation, and social loafing theories to advance an understanding of the role of peer evaluations conducted at the end of a team project. We use a state-of-the-art peer feedback system that allows students to assess and evaluate each other on five competencies critical to teamwork. We also used the system for the assessment of overall team functioning. Finally, grades on team projects were collected as a measure of team performance. Over three cohorts and using a total sample size of 162 teams and 873 students, we found that the use of peer evaluations for grading purposes, compared to a control group, promoted effective team member behavior, overall team health, and higher grades on team projects (i.e., team performance). Future research is needed to further investigate the optimal use of peer evaluations in a variety of contexts using a variety of methods.

Keywords: team, peer-evaluation, assessment & education, feedback – FB, motivation, motivating and demotivating factors

INTRODUCTION

The use of teamwork in the classroom has been steadily increasing due to both the potential educational benefits and the professional skill development required to meet 21st century work requirements (e.g., Crossman and Kite, 2012). The educational benefits of cooperative learning, for example, includes higher achievement and learning through mutually shared goals, interaction and information exchange, and higher levels of effort (Johnson and Johnson, 2009; Hammond et al., 2010). However, decades of research clearly indicates that effective teamwork is difficult to achieve (Hackman, 1990, 1998, 2002). For example, Wageman et al. (2008) found that only one in five teams met standards for excellent performance and that nearly half were functioning poorly. Moreover, teamwork can be difficult to teach during post-secondary because of large classes, limited student contact, and lack of evidence-based teaching supports that are readily available and usable. Therefore, there is a need for assessment and evaluation interventions that are easy to implement, scale well, and have few transaction costs from the perspective of students, instructors, and institutions (Riebe et al., 2016). Ultimately, practical and evidence-based solutions are needed to equip students with the collaborative problem-solving skills needed to successfully navigate the complexity of modern and future work activities (Prichard et al., 2011; Borton and Anderson, 2018; Fiore et al., 2018).

One of the key challenges in the use of teamwork in the context of higher education is to create a healthy team environment that is conducive for learning and that leads to high performance (Hansen, 2006). A healthy team is one in which team members communicate well, adapt to each other and the context, manage relationships, and capitalize on educational and learning opportunities by working together (O'Neill et al., 2017). A high performing team delivers high-quality output to stakeholders (Hackman, 2002). To achieve these, students must be motivated to work on their team projects with their team members, and not fall prey to social loafing. Indeed, social loafing can be a cause of team dysfunction in higher education (Colbeck et al., 2000; North et al., 2000). Unfortunately, being part of a dysfunctional team undermines students' potential to grow, learn, and benefit from the teamwork course components and course learning objectives (O'Neill et al., 2018). Moreover, working in a dysfunctional team can, at times, be worse than independent study (Oakley et al., 2004).

One avenue for addressing social loafing and promoting healthy, high-performing teamwork is the introduction of a structural contingency that increases accountability and perceptions of the value of effective team member behavior (i.e., valence). Round-robin peer evaluations of team member effectiveness may be useful in this respect. Specifically, each team member provides a rating of each other team member on multiple dimensions (Loughry et al., 2007). If team members know that their contributions are being evaluated by their peers, they may be likely to feel more accountable, perceive a higher valence for engaging in teamwork behaviors, and perform better.

Importantly, peer evaluations have been used both for completion (i.e., general feedback purposes) and for grading absolute levels of contributions (Patchan et al., 2018). Grading the absolute levels of contributions involves assigning students a portion of their course grade based on the magnitude of the ratings provided by peers (e.g., the% of the total possible peer-rated score). It is noteworthy that peer evaluations used for general feedback purposes versus for grades based on peer-rated contribution levels are often confounded in previous research (e.g., Brutus and Donia, 2010; Donia et al., 2018). General feedback may be unlikely to influence student or team outcomes as it is informational only, whereas there are reasons to believe that grading based on peer-rated contribution levels could have stronger implications for individual behaviors and team functioning. We believe that these two assessment purposes could affect student mindsets quite differently by creating different perceptions of accountability, but evidence-informed, empirically-based recommendations regarding the implications of these two approaches to conducting peer evaluations appear to be absent. In our experience, many instructors end up guessing about the impact of one purpose versus the other on student teamwork effectiveness and these guesses could be inaccurate. Accordingly, in the current research we considered general feedback versus grading based on peer-rated contribution levels on peer ratings of team members' behavior, overall team health, and team project grades.

The remainder of this article is organized as follows. First we introduce our theory of accountability in more depth, and then offer our research hypotheses. We then report on the research methods, results, and implications for future research and practice.

THEORETICAL BACKDROP

We integrate two theoretical perspectives to generate our hypotheses: accountability theory and the construct of valence from expectancy theory. Both of these ultimately support Gielen et al.'s (2011) proposition to use peer ratings as "an external motivator to work harder and perform better" (p. 722).

Accountability Theory

Accountability is the "perceived expectation that one's decisions or actions will be evaluated by a salient audience, and that rewards or sanctions are believed to be contingent on this expected evaluation" (Hall and Ferris, 2011, p. 134). Accountability has two key components (Hall et al., 2017). First, there is an expectation of an evaluation or at least the possibility of an evaluation (Erdogan et al., 2004). Second, there are consequences associated with the evaluation in terms of rewards and punishments, such that individuals believe they will be answerable for their decisions and actions (Schlenker et al., 1994). Accountability has been identified as a fundamental requirement for social order, because it enables social systems to operate with stability and predictability (Hall et al., 2004). Given that a team is a social system with structures that facilitate or inhibit performance, high accountability would seem to be critical. For example, high accountability may make individual contributions more clear, and social loafing more difficult. Moreover, accountability likely creates role clarity by clearly defining expectations, and having role clarity ensures that students know where to direct effort toward the team task (Lacerenza et al., 2018). Thus, accountability adds a layer of structure to student teams by making individual contributions visible and team expectations clear, and ultimately this should lead to stronger contributions to the team's task.

Valence

Notwithstanding the above, the rewards or sanctions determined by the evaluation must be valued by the individual in order to have motivating potential. The concept of valence, from Vroom's (1964) expectancy theory, captures the "degree to which the outcome is viewed as desirable" (Karau and Williams, 1993, p. 685). In addition, valence involves a consideration of the costs associated with achieving the outcome, with the maximum valence occurring for low cost, high reward outcomes (Sheppard and Taylor, 1999). Thus, all else equal, an individual will exert effort to achieve desirable outcomes with high utility. In the current context, we make the assumption that most students value their grades, and therefore the valence of the peer evaluations should be high in a grade-based assessment. Moreover, students often over-estimate their teamwork skills (Organization for Economic Cooperation and Development [Oecd], 2015), suggesting that students would associate a low cost to engaging in effective teamwork behaviors. Thus, implementing a grade-based peer evaluation should evoke a high-valence mindset, greater effort dedicated toward being an effective team member, and ultimately healthier and higher performance teamwork.

HYPOTHESIS DEVELOPMENT

Implications of Grade-Based Peer Evaluations for Team Member Effectiveness

We propose that grade-based peer evaluations should lead to higher and more consistent engagement in effective team member behaviors than general peer feedback that carries no formal consequences. Consider the psychological mindset engendered by teamwork when only general feedback is expected at the conclusion of the project. In this case, team members are both aware that peer ratings will not be used to determine grades, and that team members will not be held accountable for acting on the feedback because the team will have disbanded at the time of the evaluation. From an accountability perspective, there would be no expectation of an evaluation with consequences or that peers could hold each other answerable to the feedback. Further, without a grade contingency or a strong likelihood that team members will work together again in the future, the expected valence placed on the general feedback outcome is relatively minimal.

The grade-based peer evaluation assessment used in the current research required team members to rate each other's effectiveness at the end of the team project, as was the case for the general evaluation. The magnitude of the mean peer rating provided to each member, however, comprised a grade associated with the quality of the individual's contribution to the team project. In this case, the mindset created by knowledge of the grade-based assessment should consist of high accountability, along with high valence, for engaging in effective team member behaviors. Consider theories of social loafing, which suggest that motivation in groups is decreased when members perceive low potential for evaluation, low attention to individual contributions, and low impact of individual efforts (see Karau and Williams, 1993). By definition, the grade-based evaluation in the current study involved high evaluation potential (i.e., the peer evaluation was used for grades). Given knowledge of the grade-based evaluation, self-attention and self-regulation should be enhanced because team members are aware that their individual behavior is being observed and evaluated by their peers (McLarnon et al., 2019). This should also increase the perceived impact of individual efforts, because those contributions have a stronger potential to be recognized by peers. Taken together, grade-based peer evaluations, relative to peer evaluations that do not affect performance outcomes for students, should create high accountability around a high-valence outcome. This should enhance motivation to exert effort toward being an effective team member.

Hypothesis 1: *Peer evaluations of team member effectiveness will be higher when they are used for grades compared to general feedback purposes.*

Implications of Grade-Based Evaluations for Team Health

A healthy team can be operationalized as one that scores high on communication, adaptability, relationships, and education/learning (i.e., CARE). This framework was developed by O'Neill et al. (2017), which was based on an extensive review of the teamwork literature. According to our reasoning above, grade-based peer evaluations should create stronger perceptions of individual team member accountability as well as introduce a meaningful outcome associated with the peer evaluations (i.e., high valence). This should discourage social loafing because the presence of evaluations will draw attention to personal behaviors and the higher perceived impact of these behaviors on the team. When team members perform well by displaying behaviors that indicate they are committed, communicative, focused, high-standards oriented, and utilizing relevant knowledge, skills, and abilities (i.e., all features of individual effectiveness in teams; Ohland et al., 2012), the team as a whole will be healthier. With high individual member involvement, the team will establish better strategies, plans, and roles (communicate), monitor and back each other up (adapt), resolve conflicts and trust each other (relationships), and learn from each other by sharing existing and new knowledge needed for the team task (education).

Hypothesis 2: *Team health scores will be higher when peer evaluations are used for grades compared to general feedback purposes.*

Implications of Grade-Based Evaluations for Team Performance

High performance occurs when team members deliver high-quality output to their stakeholders (Hackman, 2002). In the current context, teams were responsible for a team project and they were graded on the quality of that work. A strong test, therefore, of whether grade-based peer evaluations produce better teamwork is to consider the effects on team project grades. Above we argued that grade-based peer evaluations should produce more effective team member behavior as well as healthier teamwork in general (i.e., the CARE model). This is because grade-based peer evaluations may create a climate of accountability, minimize social loafing, and create stronger individual contributions to the team effort. The literature is clear that, all else equal, stronger team motivation, effort, and persistence leads to higher team performance (Kanfer and Chen, 2016). However, grades offer a challenging test of the effect of grade-based peer evaluations because they are affected by a multitude of factors, and therefore the effect of grade-based peer evaluations on teamwork must be strong to produce a measurable effect on grades.

> **Hypothesis 3:** *Team performance will be higher when peer evaluations are used for grades compared to general feedback purposes.*

METHOD

Study Overview

In the current research we examined team member behaviors, overall team health, and team project grades under conditions of general feedback versus grading based on peer-rated contribution levels. Specifically, in our grade-based peer evaluation condition, peer evaluations had a direct impact on each student's grade. In the general peer feedback assessment condition, however, peer evaluations did not affect grades and were used at each student's discretion for personal development. Theoretically, knowledge that the grade-based evaluations would occur at the end of the term could lead students to engage in more effective team member behaviors, healthier teamwork overall, and achieve higher project grades (i.e., team performance). Note that, across all three cohorts involved in the current research, students were aware that the assessment of team health had no bearing whatsoever on their peer ratings or any other course grading component (i.e., had no grading implications), and students were invited to be as honest as possible in their responses.

In order to test our hypotheses we report on a cohort-based "ABA" intervention study. In Cohort 1 we introduced a teamwork project into an introductory psychology course along with a general peer assessment of team member effectiveness that was graded only on completion (described further below; i.e., "A_1," baseline condition). In Cohort 2, the course was delivered in an identical way except that the peer assessment was grade-based, with peer-rated contribution levels having implications for grades (i.e., "B," intervention). Because this is a pseudo-experimental design, even with the large sample sizes it is not possible to rule out changes in the dependent variables from Cohort 1 to Cohort 2 that may be extraneous, such as chance differences in students' interest in psychology. Accordingly, in Cohort 3 we removed the grading component of the peer evaluation intervention in order to investigate whether the dependent variables would return to baseline levels (i.e., "A_2," return to baseline). Hence the ABA design – if the dependent variable scores in Cohort 2 are superior to Cohorts 1 and 3, we can be more confident that implementing a framework for high accountability through the use of peer evaluations intervention was successful at improving team health and quality of team's project.

Participants

Participants comprised a total of 162 teams and 873 students taking an introductory psychology course for non-majors offered across three consecutive Fall terms. The average team size was 5.39. In Cohort 1 there were 284 students and 51 teams (A_1 condition). In Cohort 2 there were 297 students and 57 teams (B condition). In Cohort 3 there were 291 students and 54 teams (A_2 condition). Across the three cohorts, a small number of students declined to participate in this study in Cohorts 1 (n = 7), 2 (n = 3), and 3 (n = 9). Aside from Cohort 3, the majority of students in each cohort were in their first year of university (70.1, 60.7, and 48.3%, respectively) and came from a wide range of Faculties across the university, including Arts, Science, Business, Nursing, Kinesiology, and Engineering.

Procedure

Students self-selected into teams to complete a team project over the course of the semester that required each team to use the empirical literature to develop an action plan to achieve a personal goal of their choice, and to then record and reflect on their progress. After project completion, each student was emailed a link to complete peer evaluation and team health measures (see below) on the ITPmetrics.com platform. Students received 2% toward their final grade in the course for completing the peer ratings and team health measures regardless of whether they consented to participate in this study. At the end of the course, students were given access to personalized reports through ITPmetrics based on the average peer evaluation scores awarded to them by their team members. With respect to self-selection of the teams, given that the majority of students were non-psychology majors representing a cross-section of programs who were largely in their first term of their first year at university, very few students likely knew each other. This minimizes concerns due to self-selection affecting the study's findings.

In Cohorts 1 and 3 (the Baseline conditions, A_1 and A_2, respectively), peer assessment was used for general feedback only; that is, all members within a team received the same grade (worth 2% of their final grades) on the team project regardless of the quality of their contributions as viewed by their team members provided they simply completed the peer evaluation. In Cohort 2 (Intervention, B), students were informed at the beginning of the term that although all team members would receive the same grade on the written portion of their projects (worth 6% of their final grades), an additional 4% of their course grade would be awarded individually based on the average peer evaluation scores assigned to each student by that student's team members. That is, the peer evaluation scores would be used as a graded (summative) assessment in the course. Students were also provided with the peer evaluation survey questions at the beginning of the course to ensure they were knowledgeable of the components of the peer assessment.

To ensure that valence was sufficiently high to expect students to be invested in the team projects, we viewed the allocation of grades to teamwork to be sufficient to gain the attention of the students. First, the 4% for the peer evaluations approximates a grade level (e.g., B+ versus A–) and a .3 difference in GPA, which is a meaningful change in student achievement that can mean the difference between meeting or failing a program admission requirement. Second, the peer evaluations (2% to complete it; 4% for the feedback received) constitutes *half* of the project grade (the overall teamwork component of the course was worth 10%), which is proportionally salient and difficult to ignore. Therefore, we anticipated that the grades assigned to the peer evaluations were meaningful enough to create sufficiently high levels of valence.

For Cohort 1 (i.e., A_1), the course was taught twice a week at 5 pm, whereas for Cohorts 2 (i.e., B) and 3 (i.e., A_2), the course was taught on the same days at 11 am. Thus, comparisons of B to A_2 eliminate time of day as an explanatory variable for any differences in the dependent variables between baseline and intervention conditions. As with many course projects, the instructor made some refinements to the project after Cohort 1 to clarify expectations for students in Cohorts 2 and 3. Specific refinements included the provision of an example paper, feedback for teams on their goal statements prior to their implementation of their action plans, and revisions to the instructions provided by the instructor when she introduced the project in class to prevent the reoccurrence of areas of confusion experienced by A_1. As a result of this, we note that the comparison of A_1 to B is weaker test of the hypotheses than is A_2 to B. Specifically, A_1 versus B compares the addition of grade-based peer evaluations *and* these project clarifications, whereas A_2 versus B compares only peer evaluations for grades versus only for general feedback purposes (i.e., holding constant the project clarifications in both cohorts).

Measures

Peer Evaluations

Peer evaluations were conducted by inviting students to the ITPmetrics.com assessment platform. This website uses an adapted version of Ohland et al.'s (2012) five dimensions of team member effectiveness. Specifically, peers provided round-robin ratings on commitment, communication, focus, emphasizing high standards, and having relevant knowledge, skills, and abilities. The average of each of these five dimensions was used in our analyses. Supporting reliability evidence was provided in previous research using ratings from 30,486 raters, which mainly consisted of student teams in higher education settings, but also some industry teams (see O'Neill et al., 2019). The reliability of the peer ratings *of each individual team member* in the current study, as estimated by the intra-class correlation coefficient [i.e., ICC(2)], was 0.81. The reliability *of the overall team means* (aggregated to the team level) was 0.77, according to the ICC(2). Peer ratings were provided based on behaviors describing each of the five dimensions using a frequency based scale. The Likert-type five-point response scale ranged from "to no extent" (1) to "a great extent" (5) with the possibility to select "not familiar with team member's behavior" instead of making a rating [1].

Team Health

Team health ratings were also collected using ITPmetrics.com. The website calculates scores based on the mean of several variables for each of the four CARE dimensions. Evidence for reliability and validity was provided by O'Neill et al. (2017). Each dimension is measured using multiple well-validated scales (see O'Neill et al., 2017, 2020 for further details and relevant references). Specifically, communicate involves cooperative conflict management, role clarity, and strategy formulation and planning. Adapt involves team monitoring and back up, goal progression, and coordination. Relate involves contribution equity, healthy fact-driven conflict, lack of personal conflict, and trust. Education involves constructive controversy, exploitative learning, and exploratory learning. Items used a five-point Likert-type scale with response options ranging from "strongly disagree" (1) to "strongly agree" (5). Prior to aggregating scores to the team-level, using the within-team mean, ICC(2) inter-rater reliabilities were calculated: communicate (0.69), adapt (0.73), relate (0.73), and educate (0.60). These dimensions of team health were averaged to create an overall team health score for each team.

Team Performance Scores

The team projects for each year were graded by a unique pair of teaching assistants who were blind to the study hypotheses. To facilitate objective grading standards, all teaching assistants were provided with identical grading instructions and rubrics, in addition to graded exemplars that were assessed using the rubric and had received scores of 70, 80, and 90%. The completed rubrics were provided with each of these exemplars.

Analytical Approach

Recalling that this study used an ABA design [A_1 = baseline (Cohort 1), B = intervention (Cohort 2), and A_2 = return to baseline (Cohort 3)], to model the effect of the intervention, two dummy-coded variables were created (Dummy 1: A_1-condition = 0, B-condition = 1, A_2-condition = 1; Dummy 2: A_1-condition = 1, B-condition = 1, A_2-condition = 0). These dummy codes were chosen to represent the hypothesized positive mean differences in the peer ratings, team health, and team performance variables in the B-condition versus either of the A-conditions. Typical p-values were used to estimate statistical significance of our hypotheses. However, Hypothesis 1, which reflects a student-level effect, invoked an individual-level model in conjunction with a complex sampling adjustment (Muthén and Satorra, 1995) for the standard errors to adjust for the nesting of individuals in teams. Hypotheses 2 and 3, which reflect team-level effects, used team-level data only given that the ICC(2) estimates presented above provide sufficient support for aggregation to the team level (Allen and O'Neill, 2015). All focal analyses used M*plus* 8.4 (Muthén and Muthén, 2019).

RESULTS

Table 1 presents the team-level correlations and descriptive statistics across the three conditions of this study. The strong

[1] We only usereport ICC(2) values because they reflect the reliability of, in the first estimate provided, (a) the totality of an individual's ratings of others. T and then the second estimate we provide is (b) the reliability of the aggregate of the team member's' ratings of each other. The latter estimate is the coefficient associated with the team-level peer ratings as operationalized in our hypotheses testing, and is the appropriate reliability coefficient. An ICC(1)'s in this context would not be meaningful because it would reflect the variance explained in individual ratings by either (a) the individual or (b) the team, respectively. ICC(1) is commonly used in teamwork reseach to aggregate shared-unit constructs but this is an additive construct based on round-robin ratings and as a result ICC(1) is not a meaningful coefficient to report or interpret in this context.

TABLE 1 | Correlations and descriptive statistics (team level of analysis).

	Mean	SD	1.	2.	3.	4.	5.
1. Dummy 1	0.69	0.46	–				
2. Dummy 2	0.67	0.47	−0.48**	–			
3. Peer Evaluation	4.40	0.49	0.45**	0.03	–		
4. Team Health	4.23	0.41	0.39**	0.07	0.87**	–	
5. Team Performance	81.40	10.05	0.28**	0.07	0.17*	0.11	–

*n = 162 teams (based on n = 873 individuals). Dummy 1 coded as: A_1-condition = 0, B-condition = 1, A_2-condition = 1; Dummy 2 coded as: A_1-condition = 1, B-condition = 1, A_2-condition = 0. *p < 0.05; **p < 0.01.*

correlations involving peer evaluations, team health, and team performance supports the expected positive relations and suggests that teams with members receiving high peer evaluations tend to be the healthiest and perform the strongest[2].

Supplementary results can be obtained from the first author. Hypothesis 1 proposed that peer ratings would be higher when graded evaluations were used (i.e., the B condition). The regression coefficients accompanying both of the dummy-coded variables were significant and positive, $b_{Dummy\ 1} = 0.69$, $p < 0.01$, $b_{Dummy\ 2} = 0.35$, $p < 0.01$. This indicates that peer evaluations were significantly higher in the graded evaluation condition ($M = 4.70$, SD = 0.48) compared to the general feedback conditions at A_1 ($M = 4.01$, SD = 0.80) and at A_2 ($M = 4.36$, SD = 0.64). Thus, Hypothesis 1 was supported (see **Figure 1**).

Hypothesis 2 posited that team health, operationalized as the average level across the four team CARE variable scores, would be highest for graded evaluations. The results indicated positive, significant relations between the intervention conditions and the overall CARE variable: $b_{Dummy\ 1} = 0.48$, $p < 0.01$, $b_{Dummy\ 2} = 0.28$, $p < 0.01$. Accordingly, team health was significantly stronger in the graded evaluation condition ($M = 4.47$, SD = 0.28) compared to the general feedback conditions at A_1 ($M = 3.96$, SD = 0.44) and at A_2 ($M = 4.18$, SD = 0.32). These results lend support to Hypothesis 2 (see **Figure 2**).

Hypothesis 3 posited that team performance, operationalized as team-level grades achieved on the final course project, would be highest for teams in the grade-based peer evaluation condition. The results indicated positive, significant relations between the intervention conditions and teams' grades: $b_{Dummy\ 1} = 8.87$, $p < 0.01$, $b_{Dummy\ 2} = 5.53$, $p < 0.01$. Thus, team performance was significantly stronger in the graded evaluation condition ($M = 86.01$, SD = 7.04) compared to the general feedback conditions at A_1 ($M = 77.14$, SD = 11.90) and at A_2 ($M = 80.48$, SD = 8.19). These findings support Hypothesis 3 (see **Figure 3**).

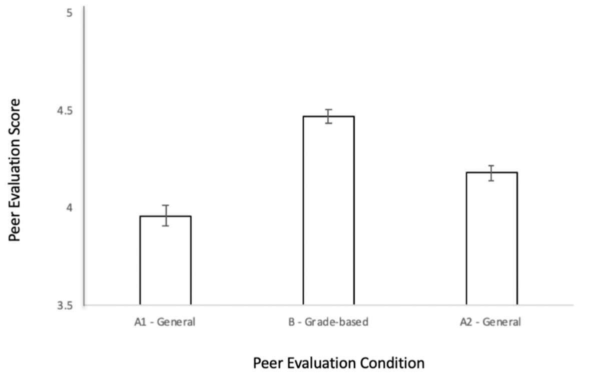

FIGURE 1 | Peer evaluation scores across conditions. Error bars represent 95% confidence intervals.

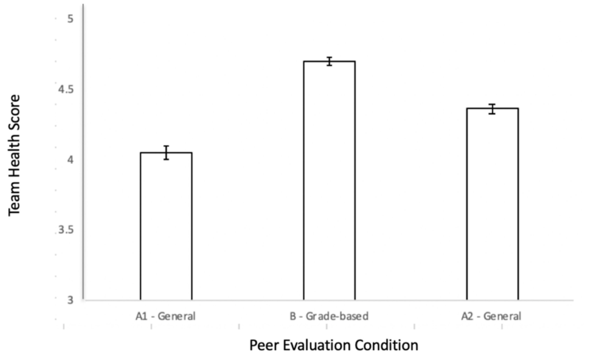

FIGURE 2 | Team health scores across conditions. Error bars represent 95% confidence intervals.

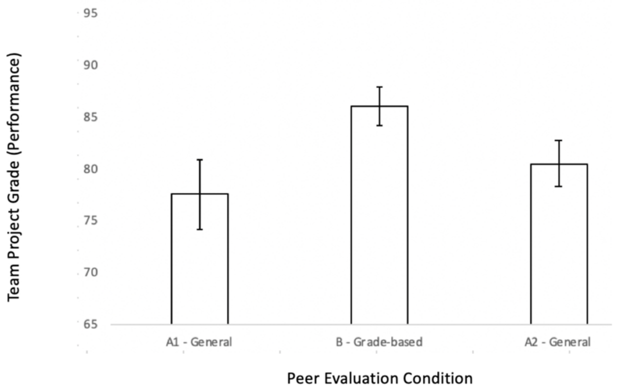

FIGURE 3 | Team project grades across conditions. Error bars represent 95% confidence intervals.

DISCUSSION

It is clear that teamwork is an important skill set that is valued by employers, and that developing teamwork in higher education is an ongoing priority (Bravo et al., 2019). Accordingly, researchers

[2] For completeness, we conducted three one-way Analysis of Variance (ANOVA) tests for peer ratings, team health, and team performance (team project grades). Results indicated that there were significant *F* tests of all three omnibus effects, and all pairwise comparisons of cell differences were significant except for team performance for A_1 versus A_2. **Figures 1–3** indicate the direction of the effects and the 95% confidence intervals. As these were not the focal analyses needed for testing the hypotheses, detailed.

need to develop evidence-based interventions that scale well for instructors of many disciplines, are easy to implement without extensive teamwork-specific theoretical knowledge, and that students can intuitively engage and interact with (Riebe et al., 2016). Indeed, in previous research, peer evaluations were viewed as accurate and useful by students using the ITPmetrics.com feedback platform used in the current research (O'Neill et al., 2019). Not examined in that research was whether the use of peer evaluations would improve teamwork behavior and functioning, leaving open an important empirical question about the role of peer feedback in student teams in higher education.

Theoretical Implications

The present research offers new theoretical propositions through which grade-based versus generic peer evaluations may influence teamwork behavior, team health, and team project performance as measured through grades. Much of the existing research considers the role of peer feedback as a learning mechanism (Brutus and Donia, 2010; Brutus et al., 2013). Specifically, much of the past research suggests that by providing and receiving peer feedback, team members learn effective behaviors and receive correction to previous actions. Indeed, peer ratings of team member behavior show consistent gains across repeated feedback episodes, although the gain from any given individual feedback episode is small (Donia et al., 2018). We adopted a different approach, arguing that grade-based peer evaluation motivates effective team member behavior (cf. Mory, 2003). This is based on previous research indicating that social loafing is more likely to occur under certain conditions: (a) when there is no evaluation of individuals, (b) when there is an emphasis on the team rather than on the individuals, and (c) when individuals perceive that their personal efforts are dispensable (cf. Karau and Williams, 1993). Thus, by integrating these previously disparate lines of theoretical rationale, we contribute an understanding of how grade-based peer evaluations address issues associated with social loafing by directly evaluating individual contributions and rewarding students for individual efforts that support the team's objectives.

We proposed that evaluations without high accountability and valence, however, would not function well as motivational incentives. Accordingly, we invoked two key components of accountability theory (Schlenker et al., 1994; Hall and Ferris, 2011), namely, that there must be an expectation of evaluation and that the evaluation results in consequences (i.e., rewards or punishments). Moreover, that the evaluation and rewards were reflected in the peer ratings received links the outcome (grades) to high valence perceptions. Valence is a component of Vroom's (1964) motivation model and central to many behavior modification theories (Michaels, 1977). If individuals do not see the contingency as appealing, they will not be motivated to work for it. By creating a direct, unambiguous correspondence between peer evaluations and course grades, students who value grades will understand that it is necessary to be an effective team member and that they could be held accountable if they are perceived as engaging in social loafing.

Our findings offer support for the proposition that grade-based peer evaluations result in more effective team member behaviors and healthier teams. We used a quasi-experimental design, which is rare in peer feedback research but should be considered a strength (Gielen et al., 2010; though see McLarnon et al., 2019 for an exception). The advantage of our quasi-experimental design was the ability to establish a baseline for the dependent variables involving teamwork, add the grade-based evaluation as an intervention to investigate changes in outcomes, and then remove the grade-based evaluation to determine whether teamwork effectiveness would again decrease without the intervention. Although randomized experiments are ideal for identifying cause and effect, on ethical grounds it is difficult to design a defensible study for the classroom. The current quasi-experimental ABA design strongly suggests that the grade-based application of peer evaluations at the end of the semester results in better teamwork behavior and greater team health.

It should be noted that the current research does not indicate that grade-based evaluations are always superior to all other forms of feedback. Formative evaluations (i.e., peer feedback used for learning and development reasons only) offer peers the opportunity for honest and accurate feedback that will not affect their teammates' grades (Gielen et al., 2010). Moreover, formative evaluations provide an avenue to address aspects of learning that are separate from accountability and valence mechanisms. Specifically, formative feedback can provide opportunities to take corrective actions, learn from previous experiences, and enhance self-awareness (Mory, 2003). Even the act of simply giving peer feedback (and not receiving it) has been related to more effective team member behavior (Dominick et al., 1997). Effective delivery of formative feedback, especially at the midpoint of the semester when team members have time to use the feedback to adjust their behavior, may in fact be quite advantageous (but through separate mechanisms than grade-based feedback obtained after project completion). On the other hand, effect sizes for formative feedback on teamwork, leadership, and other soft-skills have generally been quite low (Smither et al., 2005; Donia et al., 2018), suggesting that formative feedback on its own may be insufficient. Thus, we believe that an optimized peer evaluation program should be designed to incorporate both formative and grade-based evaluations, as we describe below.

We also found in supplementary analyses (For completeness, we conducted three one-way Analysis of Variance (ANOVA) tests for peer ratings, team health, and team performance (team project grades). Results indicated that there were significant *F* tests of all three omnimbus effects, and all pairwise comparisons of cell differences were significant except for team performance for A_1 versus A_2. **Figures 1–3** indicate the direction of the effects and the 95% confidence intervals (As these were not the focal analyses needed for testing the hypotheses, detailed supplementary results can be obtained from the first author). As both of these conditions involved general peer feedback only rather than peer evaluatons for grades, this difference would appear to reflect the effect of adding project clarifications and guidelines that students requested after the first year of the course (i.e., A_1). This is not particularly surprising, but was not the focus of the study given that previous research and meta-analyses find support for the benefits of increasing instructional clarity

(Titsworth et al., 2015; Blaich et al., 2016). However, the finding does generalize the evidence to projects involving student teams.

Practical Implications

Our results suggest that instructors should consider utilizing grade-based peer evaluations of team member effectiveness at the end of the semester, after team projects are completed. However, we also recommend incorporating a formative evaluation given the advantages of obtaining rich feedback that may provide students suggestions for improvement and foster self-reflection (Dochy et al., 1999). We suggest that an ideal time to administer formative peer evaluations may be at approximately the mid-point of the team activities. Specifically, seminal research has shown that the mid-point is a time of transition when team members begin to pay more attention to the team's progress and the remaining time before the deadline (Gersick, 1989). This period is often recommended for teamwork interventions, such as engaging in reflection related to current progress and needed adjustments (Hackman et al., 2009). Administering peer feedback at this juncture would offer students an opportunity to improve their teamwork behavior prior to the grade-based peer evaluations (O'Neill et al., 2019). Taken together, we recommend a multi-pronged approach that leverages the advantages of both formative and grade-based peer evaluations.

We also wish to point out that peer evaluations represent one of the least costly teamwork interventions in terms of instructor and student training, time to implement, and administration (e.g., accessibility to automated software platforms; Ohland et al., 2012; O'Neill et al., 2019). As well, access to automated platforms that are intuitive eases the burden of adoption on instructors and students. With respect to ITPmetrics.com (O'Neill et al., 2020), an interesting side benefit involves the user dashboards. These dashboards allow students to centralize their feedback reports and later integrate them into a skills portfolio that might be useful during employment seeking upon graduation. For instructors, the dashboards provide access to the raw data from their classes, which facilitates research ranging from local experimentation with new instructional designs to data collection for peer-reviewed knowledge dissemination (provided applicable ethical guidelines are adhered to). For institutions, the dashboards allow for a repository of data that can be helpful in various activities, such as accreditation. Finally, peer evaluation is highly scalable, given that it can be implemented effectively even in large classroom environments (such as the current research) without requiring instructors to have extensive knowledge of teamwork or access to their own software and assessments. Overall, these benefits address potential "transaction costs," that can often be a barrier to implementing new assessments, techniques, and materials into the classroom (see Riebe et al., 2016). In sum, the current research adds to a growing evidential basis supporting the use of peer evaluations as a practical tool to enhance the teamwork experience in higher education (e.g., Erez et al., 2002).

Limitations and Future Research

A potential limitation of the current research is that the peer evaluations were not used for a large portion of the students' final grades. Given that in our theorizing we expect that students value grades and therefore deliverables associated with grades will be emphasized and attended to, we wonder if it might actually be possible to obtain even greater benefits of peer evaluations if they are worth a greater proportion of course grades. On the other hand, too much emphasis on peer evaluations would likely be inappropriate and place too much responsibility on students for grading others' work. Overall, we believe that the proportion of the teamwork grade assigned to peer evaluations created adequate valence in the current study to engender motivation, but this was not directly tested.

Our theorizing invoked elements of accountability theory, valence from expectancy theory, social loafing, and motivation. However, within the current research design, we were not able to directly measure constructs associated with this theorizing. We recommend that future research adopt longitudinal designs in which perceptions of accountability and valence are measured early in the team project, social loafing and motivation are measured later in the team project, and peer evaluations (general feedback versus grade-based), and team health are measured at the end of the project. With such a design, it would be possible to test whether links between accountability, valence, social loafing, and motivation differ across formative versus grade-based peer evaluation conditions. It would also be possible to estimate the relation between team member behaviors and team health from the earlier social loafing and motivation scores, which themselves could be predicted by accountability and valence perceptions. Furthermore, including individual difference variables (e.g., achievement-striving) would allow for an investigation of whether some individuals are more strongly influenced by grade-based evaluations versus general feedback (cf. Schippers, 2014). It is possible that grade-based evaluations are more motivational for some students, whereas this could be a turn-off for other students who prefer general feedback without a grading component. It is also possible that some students do/do not feel accountable regardless of whether the grade-based evaluation is used (see Mero et al., 2006). Clearly, much future research could be done on peer evaluations and feedback.

The current study suggests that the peer evaluations were highest in the grade-based evaluation condition, a finding which follows the theoretical reasons advanced in this research. Specifically, grade-based evaluations may enhance perceptions of accountability, thereby engendering deeper commitment and motivation to contribute to teamwork. Incorporating the role of formative feedback will be particularly valuable in future research to produce an integrated, comprehensive peer feedback system in student learning teams.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Calgary Conjoint Faculties Research Ethics Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TO conceived the study and wrote the manuscript. MB collected the data, provided input on the study conception, and edited the manuscript. MM analyzed the data and edited the manuscript.

FUNDING

This research was supported by a grant to TO from the Social Sciences and Humanities Council of Canada and to TO and MB from The Taylor Institute for Teaching and Learning, University of Calgary.

REFERENCES

Allen, N. J., and O'Neill, T. A. (2015). The trajectory of emergence of shared group-level constructs. *Small Group Res.* 46, 352–390. doi: 10.1177/1046496415584973

Blaich, C., Wise, K., Pascarella, E. T., and Roksa, J. (2016). Instructional clarity and organization: It's not new or fancy, but it matters. *Change Mag. Higher Learn.* 48, 6–13.

Borton, K., and Anderson, O. S. (2018). Metacognition gains in public health graduate students following in-class peer evaluation. *Assess. Eval. Higher Educ.* 43, 1286–1293. doi: 10.1080/02602938.2018.1458211

Bravo, R., Catalán, S., and Pina, J. M. (2019). Analysing teamwork in higher education: an empirical study on the antecedents and consequences of team cohesiveness. *Stud. Higher Educ.* 44, 1153–1165. doi: 10.1080/03075079.2017.1420049

Brutus, S., and Donia, M. B. (2010). Improving the effectiveness of students in groups with a centralized peer evaluation system. *Acad. Manag. Learn. Educ.* 9, 652–662. doi: 10.5465/amle.9.4.zqr652

Brutus, S., Donia, M. B., and Ronen, S. (2013). Can business students learn to evaluate better? Evidence from repeated exposure to a peer-evaluation system. *Acad. Manag. Learn. Educ* 12, 18–31. doi: 10.5465/amle.2010.0204

Colbeck, C. L., Campbell, S. E., and Bjorklund, S. A. (2000). Grouping in the dark: What college students learn from group projects. *J. Higher Educ.* 71, 60–83. doi: 10.2307/2649282

Crossman, J. M., and Kite, S. L. (2012). Facilitating improved writing among students through directed peer review. *Act. Learn. Higher Educ.* 13, 219–229. doi: 10.1177/1469787412452980

Dochy, F., Segers, M., and Sluijsmans, D. (1999). The use of self-, peer and co-assessment inhigher education: A review. *Stud. Higher Educ.* 24, 331–350.

Dominick, P. G., Reilly, R. R., and McGourty, J. W. (1997). The effects of peer feedback on team member behavior. *Group Organ. Manag.* 22, 508–520. doi: 10.1177/1059601197224006

Donia, M. B., O'Neill, T. A., and Brutus, S. (2018). The longitudinal effects of peer feedback in the development and transfer of student teamwork skills. *Learn. Individ. Differ.* 61, 87–98. doi: 10.1016/j.lindif.2017.11.012

Erdogan, B. E., Robert, T. S., Liden, C., and Kenneth, J. D. (2004). Implications of organizational exchanges for accountability theory. *Hum. Resour. Manag. Rev.* 14, 19–45. doi: 10.1016/j.hrmr.2004.02.002

Erez, A., Lepine, J. A., and Elms, H. (2002). Effects of rotated leadership and peer evaluation on the functioning and effectiveness of self-managed teams: A quasi-experiment. *Pers. Psychol.* 55, 929–948. doi: 10.1111/j.1744-6570.2002.tb00135.x

Fiore, S. M., Graesser, A., and Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. *Nat. Hum. Behav.* 2, 367. doi: 10.1038/s41562-018-0363-y

Gersick, C. J. G. (1989). Marking time: predictable transitions in task groups. *Acad. Manag. J.* 32, 274–309. doi: 10.2307/256363

Gielen, S., Dochy, F., Onghena, P., Struyven, K., and Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Stud. Higher Educ.* 36, 719–735. doi: 10.1080/03075071003759037

Gielen, S., Peeters, E., Dochy, F., Onghena, P., and Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learn. Instr.* 20, 304–315. doi: 10.1016/j.learninstruc.2009.08.007

Hackman, J. R. (1990). *Groups that work (and those that don't)*. San Francisco, CA: Jossey-Bass.

Hackman, J. R. (1998). "Why teams don't work," in *Theory and research on small groups*, ed. R. S. Tindale (New York, NY: Plenum), 245–267.

Hackman, J. R. (2002). *Leading teams: Setting the stage for great performances*. Boston, MA: Harvard Business School Press.

Hackman, J. R., Wageman, R., and Fisher, C. M. (2009). Leading teams when the time is right: Finding the best moments to act. *Organ. Dyn.* 38, 192–203. doi: 10.1016/j.orgdyn.2009.04.004

Hall, A. T., Blass, F. R., Ferris, G. R., and Massengale, R. (2004). Leader reputation and accountability in organizations: Implications for dysfunctional leader behavior. *Leadersh. Q.* 15, 515–536. doi: 10.1016/j.leaqua.2004.05.005

Hall, A. T., and Ferris, G. R. (2011). Accountability and extra-role behavior. *Employee Responsibil. Rights J.* 23, 131–144. doi: 10.1007/s10672-010-9148-9

Hall, A. T., Frink, D. D., and Buckley, M. R. (2017). An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *J. Organ. Behav.* 38, 204–224. doi: 10.1002/job.2052

Hammond, J. A., Bithell, C. P., Jones, L., and Bidgood, P. (2010). A first year experience of student-directed peer-assisted learning. *Acti. Learn. Higher Educ.* 11, 201–212. doi: 10.1177/1469787410379683

Hansen, R. S. (2006). Benefits and problems with student teams: Suggestions for improving team projects. *J. Educ. Bus.* 82, 11–19. doi: 10.3200/JOEB.82.1.11-19

Johnson, D. W., and Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educ. Res.* 38, 365–379. doi: 10.3102/0013189X09339057

Kanfer, R., and Chen, G. (2016). Motivation in organizational behavior: History, advances and prospects. *Organ. Behav. Hum. Dec. Process* 136, 6–19. doi: 10.1016/j.obhdp.2016.06.002

Karau, S. J., and Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *J. Personal. Soc. Psychol.* 65, 681–706. doi: 10.1037/0022-3514.65.4.681

Lacerenza, C. N., Marlow, S. L., Tannenbaum, S. I., and Salas, E. (2018). Team development interventions: Evidence-based approaches for improving teamwork. *Am. Psychol.* 73, 517–531. doi: 10.1037/amp0000295

Loughry, M. L., Ohland, M. W., and Moore, D. D. (2007). Development of a theory-based assessment of team member effectiveness. *Educ. Psychol. Meas.* 67, 505–524. doi: 10.1177/0013164406292085

McLarnon, M. J. W., O'Neill, T. A., Taras, V., Law, D., Donia, M. B. L., and Steel, P. (2019). Global virtual team communication, coordination, and performance across three peer feedback interventions. *Can. J. Behav. Sci. Rev. Can. Sci. Comport.* 51, 207–218. doi: 10.1037/cbs0000135

Mero, N. P., Guidice, R. M., and Anna, A. L. (2006). The interacting effects of accountability and individual differences on rater response to a performance-rating task. *J. Appl. Soc. Psychol.* 36, 795–819. doi: 10.1111/j.0021-9029.2006.00044.x

Michaels, J. W. (1977). Classroom reward structures and academic performance. *Rev. f Educ. Res* 47, 87–98. doi: 10.3102/00346543047001087

Mory, E. H. (2003). "Feedback research revisited," in *Handbook of Research on Educational Communications and Technology*, eds D. Jonassen and M. Driscoll (New York, NY: Routledge), 745–783.

Muthén, B. O., and Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociol. Methodol.* 25, 267–316. doi: 10.2307/271070

Muthén, L. K., and Muthén, B. O. (2019). *Mplus 8.4*. Los Angeles, CA: Author.

North, A. C., Linley, P. A., and Hargreaves, D. J. (2000). Social loafing in a co-operative classroom task. *Educ. Psychol.* 20, 389–392. doi: 10.1080/01443410020016635

Oakley, B., Felder, R. M., Brent, R., and Elhajj, I. M. (2004). Turning student groups into effective teams. *J. Stud. Cent. Learn.* 2, 9–34.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., et al. (2012). The comprehensive assessment of team member effectiveness: development of a behaviorally anchored rating scale for self- and peer evaluation. *Acad. Manag. Learn. Educ.* 11, 609–630. doi: 10.5465/amle.2010.0177

O'Neill, T. A., Deacon, A., Gibbard, K., Larson, N. L., Hoffart, G. C., Smith, J., et al. (2017). Team dynamics feedback for post-secondary student learning teams. *Assess. Eval. Higher Educ.* 43, 571–585. doi: 10.1080/02602938.2017.1380161

O'Neill, T. A., Larson, N., Smith, J., Deng, C., Donia, M., Rosehart, W., et al. (2019). Introducing a scalable peer feedback system for learning teams. *Asses. Eval. Higher Educ.* 44, 848–862. doi: 10.1080/02602938.2018.1526256

O'Neill, T. A., McLarnon, M. J. W., Hoffart, G. C., Woodley, H. J., and Allen, N. J. (2018). The structure and function of team conflict profiles. *J. Manag.* 44, 811–836. doi: 10.1177/0149206315581662

O'Neill, T. A., Pezer, L., Solis, L., Larson, N., Maynard, N., Dolphin, G., et al. (2020). Team dynamics feedback for post-secondary student learning teams: Introducing the "Bare CARE" assessment and report. *Asses. Eval. Higher Educ.* 43, 1–15. doi: 10.1080/02602938.2020.1727412

Organization for Economic Cooperation and Development [Oecd] *Performance in collaborative problem solving, PISA 2015 Results (Volume V): Collaborative Problem Solving*. Paris: OECD Publishing.

Patchan, M. M., Schunn, C. D., and Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Stud. Higher Educ.* 43, 2263–2278. doi: 10.1080/03075079.2017.1320374

Prichard, J. S., Bizo, L. A., and Stratford, R. J. (2011). Evaluating the effects of team-skills training on subjective workload. *Learn. Instr.* 21, 429–440. doi: 10.1016/j.learninstruc.2010.06.003

Riebe, L., Girardi, A., and Whitsed, C. (2016). A systematic literature review of teamwork pedagogy in higher education. *Small Group Res.* 47, 619–664. doi: 10.1177/1046496416665221

Schippers, M. C. (2014). Social loafing tendencies and team performance: The compensating effect of agreeableness and conscientiousness. *Acad. Manag. Learn. Educ.* 13, 62–81.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., and Doherty, K. (1994). The triangle model of responsibility. *Psychol. Rev.* 101, 632–652. doi: 10.1037/0033-295x.101.4.632

Sheppard, J. A., and Taylor, K. M. (1999). Social loafing and expectancy-value theory. *Personal. Soc. Psychol. Bull.* 25, 1147–1158. doi: 10.1177/01461672992512008

Smither, J. W., London, M., and Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Person. Psychol.* 58, 33–66. doi: 10.1111/j.1744-6570.2005.514_1.x

Titsworth, S., Mazer, J. P., Goodboy, A. K., Bolkan, S., and Myers, S. A. (2015). Two meta-analyses exploring the relationship between teacher clarity and student learning. *Commun. Educ.* 64, 385–418.

Vroom, V. H. (1964). *Work and motivation*. Oxford: Wiley.

Wageman, R., Nunes, D. A., Burruss, J. A., and Hackman, J. R. (2008). *Senior leadership teams: What it takes to make them great*. Boston, MA: Harvard Business Review Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.