Challenges and Opportunities for Data-Centric Peer Evaluation Tools for Teamwork

WENXUAN WENDY SHI, University of Illinois, Urbana-Champaign, USA
AKSHAYA JAGANNADHARAO, University of Illinois, Urbana-Champaign, USA
JAEWOOK LEE, University of Illinois, Urbana-Champaign, USA
BRIAN P. BAILEY, University of Illinois, Urbana-Champaign, USA

Peer evaluations are critical for assessing teams, but are susceptible to bias and other factors that undermine their reliability. At the same time, collaborative tools that teams commonly use to perform their work are increasingly capable of logging activity that can signal useful information about individual contributions and teamwork. To investigate current and potential uses for activity traces in peer evaluation tools, we interviewed (N=11) and surveyed (N=242) students and interviewed (N=10) instructors at a single university. We found that nearly all of the students surveyed considered specific contributions to the team outcomes when evaluating their teammates, but also reported relying on memory and subjective experiences to make the assessment. Instructors desired objective sources of data to address challenges with administering and interpreting peer evaluations, and have already begun incorporating activity traces from collaborative tools into their evaluations of teams. However, both students and instructors expressed concern about using activity traces due to the diverse ecosystem of tools and platforms used by teams and the limited view into the context of the contributions. Based on our findings, we contribute recommendations and a speculative design for a data-centric peer evaluation tool.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Collaborative and social computing systems and tools*; **Computer supported cooperative work**; • **Applied computing** → *Collaborative learning*.

Additional Key Words and Phrases: teams, teamwork, team assessment, peer evaluation, education, activity traces

ACM Reference Format:
Wenxuan Wendy Shi, Akshaya Jagannadharao, Jaewook Lee, and Brian P. Bailey. 2021. Challenges and Opportunities for Data-Centric Peer Evaluation Tools for Teamwork. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 432 (October 2021), 20 pages. https://doi.org/10.1145/3479576

1 INTRODUCTION

Peer evaluations are critical for assessing the contributions of team members, discovering problems within teams, and helping team members improve their team skills [12]. Instructors are unlikely to be able to observe all team processes, especially when there are many teams in a course or the teamwork is performed online. Peer evaluations shift the responsibility of assessment to the students,

Authors' addresses: Wenxuan Wendy Shi, wshi16@illinois.edu, University of Illinois, Urbana-Champaign, 201 N Goodwin Ave, Urbana, Illinois, USA, 61801; Akshaya Jagannadharao, University of Illinois, Urbana-Champaign, 201 N Goodwin Ave, Urbana, Illinois, USA, 61801; Jaewook Lee, University of Illinois, Urbana-Champaign, 201 N Goodwin Ave, Urbana, Illinois, USA, 61801; Brian P. Bailey, bpbailey@illinois.edu, University of Illinois, Urbana-Champaign, 201 N Goodwin Ave, Urbana, Illinois, USA, 61801.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2573-0142/2021/10-ART432 $15.00
https://doi.org/10.1145/3479576

who have a greater awareness of their team dynamics, and reduce the burden on instructors for assessing teams [15]. Despite their necessity, peer evaluations may not always be reliable due to cognitive and social factors that students experience when completing the evaluations [4, 29, 39]. In addition, instructors may not always know how to best interpret the evaluation results.

In this paper, we envision a future in which the activity traces captured by the collaborative tools commonly used by student teams can be leveraged to compose fair, accurate, and meaningful peer evaluations of teamwork. We use the term "activity trace" to refer to an objective record of contribution automatically captured by an online tool. For example, Google Docs logs each edit, who made the edit, and the time of the edit. Other tools that log activity traces include version control systems (e.g., Git), group messaging applications (e.g., Slack), and video conferencing tools (e.g., Zoom). Activity traces from different tools can capture different aspects of the teamwork, including technical contributions, communication, and meeting participation.

Research has shown that activity traces affect impression formation between users in online peer production communities, which can affect how they evaluate each other's contributions [27, 28]. However, there is little knowledge about how students and instructors might leverage activity traces for peer evaluations of teamwork. What opportunities and challenges do students and instructors perceive for incorporating activity traces into the peer evaluation process? How can we extend existing peer evaluation tools or create new tools to incorporate the insights from these stakeholders? To answer these and other research questions, we interviewed (N=11) and surveyed (N=242) students who performed peer evaluations as part of their teamwork experience. The students were sampled from different courses and disciplines at our university. We also interviewed instructors (N=10) teaching courses that involve teamwork with different enrollment sizes and in different disciplines.

Our results indicate that there is a movement towards a data-centric peer evaluation process. We found that students value and are already beginning to use activity traces such as document edits shown in Google Doc's version history and code commits on Github in supporting peer evaluations. Nearly all (96%) survey participants considered at least one type of activity trace when evaluating their teammates. Despite this, student interviewees reported relying on their subjective experiences the most, which can be affected by students' orientation towards grades, personal relationships, and other social and cognitive factors. From the instructor interviews, we identified the emerging usage of activity traces as a corroborative source of data when instructors desired a more robust and objective picture of the teamwork. However, both students and instructors expressed concerns about fully capturing all aspects of teamwork given the diversity of tools and platforms across which contributions were made and having a limited view into the context of the contributions. These findings suggest that data-centric peer evaluation tools should provide support for a variety of activity trace sources, allow students to express the relative value of different types of contributions, and enable contextualization of the activity trace data.

The contributions of our work to the CSCW community are 1) insights into the opportunities and concerns perceived by instructors and students for incorporating activity traces into the peer evaluation process; 2) descriptions of emergent uses of activity traces for peer evaluation processes; and 3) implications for leveraging activity traces in peer evaluation processes and an exploration of these implications through a speculative design of a data-centric peer evaluation tool.

2 RELATED WORK

We situate our work in the context of prior literature on teamwork assessment, with a focus on peer evaluations and systems designed to support them. Teamwork assessment is a holistic concept that covers all types of feedback, evaluation, and assessment that occur within teams [12]. It can be applied in a formative or summative manner and at both an individual and team level. Our work

focuses on team work assessment at an individual level and on assessing the contributions that each member of a team made to the teamwork process and outcomes. For this paper, we define a "team" as a group of 2-8 people working toward a common goal.

2.1 Peer Evaluations of Teamwork

Peer evaluations are a well-known method for assessing teamwork processes. For example, teammates might rate each other's contributions to the teamwork deliverable, interpersonal interactions, knowledge and skills, and timeliness of work. The evaluation results allow instructors to gain insights into individual behaviors and group dynamics that they would not normally be able to observe and from students who should be most knowledgeable of the team's functioning. Peer evaluations can also help instructors to account for differences in individual contributions and assign grades fairly based on students' contributions [11, 41], alleviating student concerns about "free-riding" group members that contribute little to no work [1, 33]. Prior research has also argued that students can benefit from using peer evaluations because peer evaluations can promote reflection on group processes and increase feelings of accountability towards their teammates [5], even improving students' teamwork skills in the long-term [10]. Instructors often incorporate self-assessment into the peer evaluation process as well to give students greater input into how they are evaluated [20].

Our work contributes to the literature by exploring emergent and future uses of activity traces in the peer evaluation process. We investigate the types of activity traces that are being leveraged for peer evaluations, how these traces are interpreted, and the new opportunities and potential problems they bring to peer evaluation processes. Our investigation also includes the perspectives of both students and instructors.

2.2 Peer Evaluation Tools

In a peer evaluation, each member of the team must differentiate individual contributions to the teamwork process and outcomes over time. Most research on peer evaluation tools has focused on alleviating the administrative burden of conducting the evaluations [14, 23]. For example, tools have been developed to aid the evaluation process by providing standardized criteria and scales [3, 31], automating the aggregation and release of the results within the teams and to the instructor [18], and highlighting patterns in the results that deserve closer inspection [24].

Peer evaluation tools are distinct from the class of tools for assessing peer outcomes. A peer outcome assessment tool helps students evaluate a specific product or outcome created by peers (e.g., design prototypes) [26]. To improve peer outcome assessments, prior work has investigated intelligent assignment of peers and the work to be evaluated [22], the use of comparative judgements [6], and guidance for writing the assessments [7].

While there has been abundant prior work on developing new techniques and supporting tools for assessing peer outcomes [2, 21, 26, 37], there has been comparatively little research on developing new techniques and tools for assessing peer contributions to teamwork. Through the work presented in this paper, we hope to contribute to a new thread of research in the CSCW community on improving peer evaluations of teamwork processes.

The goal of our work is to determine whether and how the class of peer evaluation tools might incorporate access to the activity data captured by the online tools used for the team work. This kind of data could potentially help students develop more accurate ratings by having access to summaries and specific examples of contributions made by each team member. Such data might also help students overcome factors (e.g., friendship, peer pressure, and bias) that can affect their ratings and provide stronger methods for instructors to interrogate the evaluation outcomes.

2.3 Online Impression Formation

In online peer production communities, users often form impressions of each other based on activity traces found in the systems and tools they use. These impressions can then influence how they evaluate other users and their contributions [27, 34]. For example, developers make social inferences about other users based on activity traces such as recency and volume of code commits in open-source communities [9, 38]. Different visualizations of these traces can also affect impressions of the work [28]. Such prior work in the area of online impression formation has primarily examined how individual strangers form impressions of each other [28, 35, 36]. Our work examines the use of activity traces for peer evaluations in an academic team setting where both instructors and students need to form impressions of individual team members. Students have already spent time working with each other by the time they complete the peer evaluation. They may also continue to interact with each other after the course or the evaluation is complete. Thus, we also examine the interpersonal factors that may affect their use and interpretation of activity traces.

3 RESEARCH QUESTIONS

As shown in prior work on online impression formation, activity traces can influence how people evaluate each other's contributions. To effectively leverage activity traces in peer evaluations where students must evaluate the contributions of each of their teammates, we need to understand what challenges and concerns instructors and students have with using existing peer evaluation processes. We also investigate what role activity traces currently play in the peer evaluation process. Therefore, our work asks three research questions:

RQ 1: What social and cognitive factors affect students' attitudes and approaches toward peer evaluations?

RQ 2: What are instructors' goals for using peer evaluations and what challenges do they face in operationalizing these goals?

RQ 3: How might instructors and students leverage online activity traces when evaluating students' contributions in teams?

4 METHODS

To answer the above research questions, we conducted a mixed-methods study involving university instructors and students. We first discuss the semi-structured interviews with students in Section 4.1, followed by the student survey in Section 4.2, and conclude with the instructor interviews in Section 4.3. Participant demographics for the interviews and survey are shown in Table 1. This study was approved by the Institutional Review Board at our university.

4.1 Student Interviews

We conducted semi-structured interviews with eleven students who had participated in a recent teamwork experience in which they evaluated the other members of their team. In the interviews, we asked about what factors students considered when evaluating their teammates, the role of data and tools during the evaluation process and overall teamwork experience, and their attitudes towards peer evaluations, including perceived accuracy of the evaluations they completed and social pressures they experienced when evaluating teammates. While participants focused on a specific teamwork experience for the majority of the interview, we also asked about their experiences with peer evaluations in past teamwork experiences when discussing their attitudes towards peer evaluations. We augmented the discussion of data and tools by asking participants to walk through

Table 1. Participant demographics

	Student Interview (N=11)	Student Survey (N=242)	Instructor Interview (N=10)
Gender			
Female	7 (63.6%)	116 (48%)	7 (70%)
Male	4 (36.4%)	121 (50%)	3 (30%)
No answer	-	5 (2%)	-
Course Subject			
Social Sciences	7 (63.6%)	123 (50.8%)	1 (10%)
Formal and Applied Sciences	4 (36.4%)	62 (25.6%)	7 (70%)
Natural Sciences	-	35 (14.5%)	1 (10%)
Humanities	-	22 (9.1%)	1 (10%)
Year			
Freshman	1 (9%)	5 (2%)	-
Sophomore	2 (18%)	12 (5%)	-
Junior	2 (18%)	39 (16%)	-
Senior	5 (46%)	41 (17%)	-
Graduate student	1 (9%)	143 (59%)	-

various artifacts they created during the teamwork experience, such as project documents and chat histories. If the tool logged activity traces, the students were also asked to review the log and describe their impressions of the information shown.

Students were recruited through online posts in a Reddit community targeted towards students at our university. Interviews took place from mid- to late Spring of 2020. All but one interview were conducted remotely using a video conferencing tool. Interviewees received a $10 gift certificate as compensation for completing the interview. Five of the participants were seniors, two were juniors, two were sophomores, one was a freshman, and one was a graduate student. Participant majors belonged to various colleges at our institution, including Engineering (N=4), Liberal Arts and Sciences (N=4), Media (N=2), and Business (N=1). Seven participants identified as female and the rest identified as male.

4.2 Student Survey

We designed an online survey to further investigate our research questions and the themes that arose from the interviews. For example, we identified 10 types of activity traces from the interviews and then asked survey respondents to rate the extent to which they considered each type of activity trace during their peer evaluations. Survey participants first identified a specific teamwork experience and then responded to all questions on the survey with regard to that experience. In the survey, we asked participants about their consideration of activity traces when evaluating their teammates and their perceptions (perceived accuracy, fairness, social pressures) of the peer evaluation/s they completed during the teamwork experience. These survey questions are listed in the Appendix. Most of the survey questions were structured as Likert items, but we also included four optional open-ended questions to allow participants to elaborate on the characteristics of their evaluation, other factors and data that they considered during the evaluation, and any other comments they had about the topics in the survey.

We distributed the survey in Fall 2020 to 5385 students using a student sampling service provided by our university. For compensation, participants could opt-in to a lottery with a one in 10 chance to win a $10 gift certificate. We received survey responses from 265 students, indicating a 5% response rate. We filtered out responses that failed attention checks within the survey or did not qualify for participating in the survey. We were left with a total of 242 survey responses. Of those responses, 48% of participants identified as female and 50% identified as male. 59% were graduate students, 17% were seniors, 16% were juniors, 5% were sophomores, and 2% were freshmen. We did not collect specific course information in the survey but 52% of respondents' majors were in Social Sciences, 28% were in Formal or Applied Sciences, 14% were in Natural Sciences, and 6% were in Humanities.

4.3 Instructor Interviews

We conducted semi-structured interviews with ten instructors who taught courses in which they assigned team-based work. During the interviews, we asked instructors what methods they used to assess students in teams, their motivation for using these methods, how they interpreted and acted on information collected, the perceived strengths and weaknesses of these methods, challenges they faced, and opportunities for improvement. We did not constrain our interviews with instructors to only peer evaluations so as not to preclude other methods that instructors may use to assess teams. Because we prioritized obtaining a diversity of perspectives, we also did not limit our recruitment to instructors that taught courses mentioned in the student interviews. The courses taught by our instructor interviewees were different from the courses that participants identified in the student interviews, though it is possible that student interviewees may have taken courses with the instructor interviewees in the past. Both instructor and student interviewees primarily discussed long-term projects in which student teams had to produce multiple deliverables of varying artifact types, allowing us to find commonalities in their perspectives on activity traces.

Instructors were recruited via a faculty and staff mailing list at our university, email, and word-of-mouth. All interviews were conducted using an online conferencing tool; they ranged from 40 to 90 minutes. The instructors were affiliated with a range of disciplines: Agricultural and Biological Engineering, Software Engineering, Bioengineering, Industrial and Systems Engineering, Human-Computer Interaction, Molecular and Cell Biology, Business Administration, Information Science, Physics, and Population Health Nursing. The typical size of classes that they taught ranged from 10 to 200 students. Seven participants identified as female and the rest identified as male.

4.4 Data Analysis

We employed thematic analysis [17] to analyze the student and instructor interview data. A member of the research team first performed open coding on the data and then refined these codes in an iterative and reflexive process. The same person then used axial coding to group these codes into larger themes, centered on our research questions. For the student interviews, we extracted themes related to students' attitudes and approaches towards peer evaluations to address RQ1. For the instructor interviews, we extracted themes related to instructors' goals for using peer evaluations in teams and the types of challenges they face when doing so to address RQ2. For both student and instructor interview data, we also extracted themes related to usage of activity traces when evaluating team contributions and the challenges associated with it to address RQ3. Subsequent passes and discussion by the research team further refined the themes. Themes that did not directly relate to our research questions are excluded from discussion. To analyze the open-ended responses from the student survey, a member of the research team used thematic analysis to categorize the responses.

	Format		
Frequency	Purely qualitative	Purely quantitative	Mixed-format
Only at the end	I = 4 (36.4%) S = 27 (11.2%)	I = 2 (18.2%) S = 68 (28.1%)	I = 7 (63.4%) S = 63 (26%)
Multiple times	I = 3 (27.3%) S = 10 (4.1%)	I = - S = 18 (7%)	I = 7 (63.4%) S = 20 (8.3%)

Table 2. Format and frequency of peer evaluations completed by students in the interviews and survey. 'I=' represents the percentage of interview participants who had ever completed that type of peer evaluation and 'S=' represents the percentage of survey participants who completed a peer evaluation of that type for the specific group work experience they focused on during the survey.

5 RESULTS

We present the results of our interviews and survey. We begin by discussing the challenges and concerns that students had with using traditional peer evaluations. We then discuss instructors' goals for assessment and challenges they faced in trying to achieve these goals. Finally, we discuss the emergent use of activity traces during the evaluation process for both students and instructors.

5.1 Student Perspectives (RQ1)

5.1.1 What types of peer evaluations did students complete? Although students in our interviews came from different courses and used different types of peer evaluations, we were able to categorize the peer evaluations that they described completing based on format and frequency. We then asked our survey participants to describe their peer evaluations along these two axes through a series of multiple-choice questions. We identified three categories of format: peer evaluations that were completely quantitative, peer evaluations that were completely qualitative, and peer evaluations that had both quantitative and qualitative components (mixed-format). We then applied a binary classification to frequency of the evaluation: evaluations that only occurred once at the end of the course or teamwork experience and evaluations that occurred multiple times during the course or teamwork experience. Table 2 summarizes the reported use of each type of peer evaluation in the student interviews and survey.

5.1.2 Social and cognitive factors that affect student attitudes and approaches. In this section, we discuss the social and cognitive factors that influenced students' attitudes and approaches towards peer evaluations. When applicable, we support each section with findings from our survey on students' perceptions towards peer evaluations. Perception questions on the survey were measured on a 5-point agreement scale (1=Strongly disagree to 5=Strongly agree) except for satisfaction which was rated from 1=Extremely dissatisfied to 5=Extremely satisfied. In the survey, we only asked participants about the peer evaluation/s they completed for a single teamwork experience.

Bias and Collusion. Seven students in our interviews reported being lenient with teammates in previous peer evaluations. They were afraid of being harsh on their teammates, especially if that could result in penalizing their grade. 34% of students in our survey agreed to some extent that they felt pressured to give a teammate a more favorable evaluation during their teamwork experience (μ=2.61, s=1.45). A larger percent (61%) of students in the survey reported not wanting to hurt their teammates' grades as a result of the evaluation (μ=3.67, s=1.25).

Personal relationships also introduced bias into students' evaluations: *"I feel like a lot of times people are in group projects with their friends and they're not going to score their friends a lower score even if they deserve it. Or you'll definitely overscore people you'll have class with again."* (S4) Another student described becoming closer friends with a teammate due to cultural compatibility, thus

biasing their evaluation. In addition, three students described cases where their teams had a *"game mentality"* (S5) and would negotiate to give each other good scores. Other biases described in the student interviews included a tendency to remember only the most extreme or recent events, and to be lenient with team members as long as everything worked out in the end.

Personal Standards and Dispositions. Personal standards and dispositions also led to concern about or instances of subjectivity for ten students. S2 (Grad student, Computer Science) compared evaluators who were more easygoing with evaluators who were more high-strung and tried to *"convey their angst through their evaluation"*. Another student described having *"overpowering personalities"* in their team and worried that these members might overreact in the peer evaluations (S8, Senior, Advertising). Students also demonstrated different standards when evaluating themselves; two students were more critical when evaluating themselves because they felt they had put in less effort than they were capable of. Some students also calibrated their evaluations of other people based on their own self-evaluations, which resulted in overrating team members.

Work Style. Seven students stated that members of their team were assigned or took on different roles or responsibilities. It was harder for students to meet together when teamwork was remote and students liked being able to fit each other's strengths or skill sets. Three of these students described being able to evaluate their team members more easily when each person had their own roles because they knew which tasks a team member was supposed to do. Differing roles also sometimes meant that some members were more active during certain stages of a project than others. This complicated the evaluation process when there were multiple evaluations because some team members may not have had as much to do prior to an evaluation. When there was only one evaluation at the end, it meant that students had to recall and make sense of their team members' contributions at each stage of the project. Two students also found it easier to evaluate their team members because they often worked or met synchronously with their teams.

Limitations of Peer Evaluations. Eight students felt that questions or prompts on peer evaluations were sometimes not specific enough to the different components of their teamwork or to the context of their project. In addition, three participants thought that peer evaluations could not effectively capture extenuating circumstances that contribute to variance in teammates' contributions over time. As S1 stated, *"It's hard to give concrete singular answers to questions which ask about someone's performance over the course of the entire project because everyone has their own lives, everyone has their own things going on every week, every day, that changes."* One student also mentioned that quantitative evaluations where students assigned their teammates a numeric rating made it easy to resort to *"defaulting"*, without reflecting deeply on the team's relative contributions (S10). As a result, students valued open-ended responses where they could elaborate on team members' contributions, whether to highlight specific strengths and accomplishments or to discuss issues they were dealing with.

Communication of Expectations. Differing or unclear expectations increased anxiety for three students about how they would be evaluated by their teammates: *"I usually get really worried about peer evaluations. It's just because I never really know what the other person is thinking if we don't communicate... But if we don't communicate, then I have no idea what they expect from me."* (S6) S9 even worried that there was not much proof of the effort they were making. Communication and transparency ensured that teammates were on the same page, allowing them to have a better understanding of how they would be evaluated by their teammates. For example, S3 felt more confident about how they were evaluated because they communicated with their team regularly and had received positive feedback from their team members outside of the evaluations.

5.2 Instructor Perspectives (RQ2)

We now describe the instructor's goals and challenges in administering peer evaluations.

5.2.1 Instructor Goals. We identified three primary goals that instructors wanted to accomplish that framed their approaches to assessing teams. These goals are neither mutually exclusive nor exhaustive of all potential objectives for assessing teamwork. Assessment approaches ranged from simple measures like asking teams to state the contributions of each member when submitting their deliverables to multi-method approaches utilizing peer evaluations, team meetings, reflections, etc.

Assessing Individual Contributions. Instructors sought to understand how each team member contributed to the team in terms of both the member's part in the final outcome as well as their contributions to the team dynamic throughout the teamwork. It was important for instructors to ensure that every team member was being held accountable and was contributing to the team. This knowledge was especially important for seven instructors who were concerned about assigning grades fairly based on each member's contributions to the team. Assigning grades based on individual contributions helped *"moderate the effect of being on a lucky team"* (I10, Industrial and Systems Engineering) and avoid situations where students were *"being carried by a group grade"* (I3, Population Health Nursing). Some instructors also wanted the opportunity to reward students who went above and beyond in contributing to the team. Thus, several instructors valued having a quantitative component because it was more objective and easier to interpret and turn into a grade. Two instructors also talked about how having multiple peer evaluations over time created a stronger signal of students' contributions.

Identifying Potential Dysfunctions or Conflicts. Seven instructors described needing to identify potential problems or dysfunctions within teams, whether that was a member being completely absent from the team or a deeper issue with the team culture. Even when students directly presented issues to instructors first, instructors wanted to make sure that they were not unintentionally misinterpreting the situation or missing any information. Four instructors also emphasized identifying dysfunctional behaviors or conflicts early so that they could determine appropriate strategies for helping students to resolve them in time. This further motivated instructors to conduct peer evaluations multiple times during the semester, rather than wait until the end of the semester. Instructors also looked at the open-ended comments in peer evaluations and students' self-reflections to identify issues or validate outliers in the quantitative data.

Helping Students Develop Teamwork Skills. Nine instructors discussed the importance of helping students develop and get feedback on their teamwork skills. Because the goal of assigning teamwork in courses was often to help prepare students for careers which involved teamwork or collaboration, instructors wanted students to be able to learn how to work more effectively in teams. Two instructors also described the motivating effect that feedback can have on teams, helping to build trust between team members. Four instructors ran peer evaluations multiple times so that students would have opportunities to improve based on the feedback they got. One instructor even assigned a grade to students based on their improvement in peer evaluation scores from the middle to the end of the semester.

5.2.2 Challenges of Operationalizing Goals. Though instructors used other methods to assess teams, eight of the instructors relied primarily on peer evaluations, with four using the automated peer evaluation system, CATME [24]. Peer evaluations allowed instructors to obtain in-depth information about the contributions of each member of the team, identify potential conflicts and dysfunctions through outliers in the data, and help students develop their teamwork skills by giving and receiving feedback. However, they also introduced new challenges with administering the evaluations, interpreting the responses, and facilitating feedback.

Administering Evaluations. Instructors had to balance their goals for using peer evaluations with the limited time and resources available to them in a classroom setting. Thus, a significant challenge experienced by nine instructors was a trade-off between the quality and the efficiency of

the assessment process. This constrained the length and format of peer evaluations, the number of times instructors were able to administer peer evaluations, and their ability to provide feedback. For example, even though qualitative responses could reveal more insights about the team, quantitative ratings were simply easier and more time-efficient to interpret.

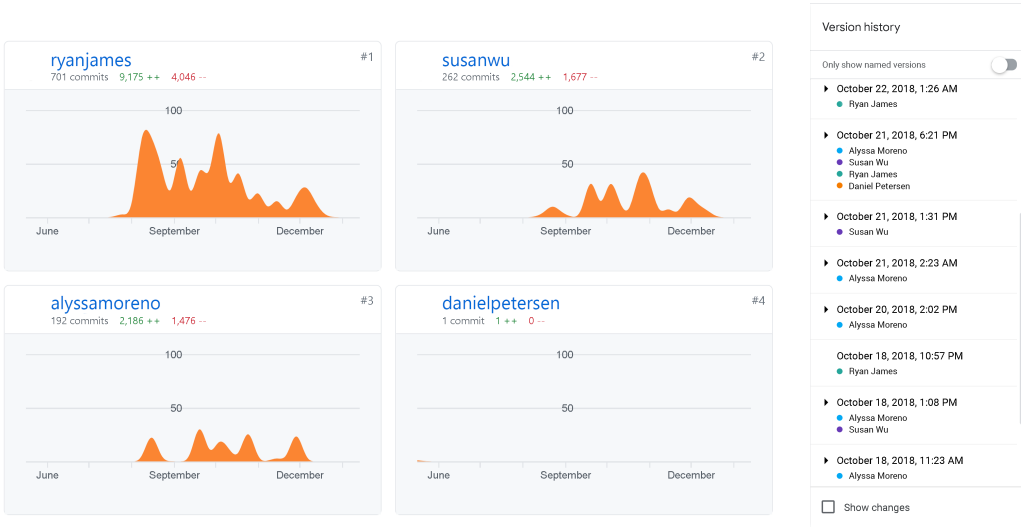
Even the four instructors who used an automated peer evaluation system reported that it was time-consuming to deploy the assessments, especially in large classes of more than a hundred students. Thus, although instructors preferred to administer peer evaluations multiple times throughout the course, it was not always feasible to do so: *"I sometimes feel a 16-week semester is quite short when it comes to achieving that in a university environment."* (I7, Agricultural and Biological Engineering) In such cases, there would only be an evaluation at the end of the course when it would be too late for instructors to identify and resolve dysfunctional behaviors and for students to learn from the evaluation feedback. For an instructor who lacked funding to license an automated peer evaluation system or extra labor provided by teaching assistants, administering evaluations was even more labor-intensive.

Facilitating Feedback. Constraints on time and resources sometimes limited instructors' ability to train students to constructively evaluate each other as well. Three instructors expressed concern about the negative impact that peer evaluations could have on team dynamics, especially when the students are able to see how they were evaluated. When students are not taught how to provide feedback to each other and instead only know how to judge each other's performance, this can result in *"a chilling effect on people's willingness to take risks and make themselves feel vulnerable"* (I2, Business Administration). In turn, the lack of psychological safety can lead to students assigning blame instead of working together to resolve conflicts. According to I5, an Information Sciences instructor, one of the biggest challenges of team assessment is giving students feedback without creating a *"surveillance environment"* where students feel judged for every mistake: *"That's why it's really important to create some sort of culture where it's safe for students to give each other feedback. I would value that sort of safety in the actual learning more than the accuracy of the assessment."*

Interpreting Evaluations. Instructors also experienced challenges with interpreting the peer evaluation responses. Eight instructors expressed concern about whether students were being honest in their peer evaluations and whether they could actually trust the data. They cited reasons for potential dishonesty such as bias from interpersonal relationships, fear of hurting a member's grade, and lack of true anonymity when team sizes are small. I8 (Bioengineering) brought up an instance where they suspected students were not being honest after the first peer evaluation: *"I think that the fact that their feedback was released to their teammates bothered some of them and caused them to be less honest the second time. I noticed that all the ratings went up and I don't think that it was necessarily because people were super satisfied within the team."* The inflation of evaluations made it more difficult for instructors to identify potential issues within the team that needed to be resolved. Instead, instructors were only able to see conflicts emerge from evaluations at the end of the project or course, which I10 attributed to frustration from team members.

Two instructors also brought up the possibility of students themselves not knowing the full context of their teammates' contributions. Examples mentioned were members within the team splitting off and not knowing how much work each faction did, or students not realizing how much effort may have been put into the work. Not being on the same page can lead to conflicts in which students tell different stories about the same situation, creating additional work for instructors to process, as experienced by four instructors in our interviews. Instructors also discussed how differing motivations or expectations of team members and external or contextual circumstances can contribute to this problem.

These instances created irregularities in the data that instructors needed to be aware of. CATME's automatic flagging of exceptional or unusual rating patterns was praised by two instructors as



(a) Github Visual Summary

(b) Google Doc Version History

Fig. 1. Fictitious examples of activity traces that students mentioned during the interviews.

helpful in identifying these cases. However, CATME does not elucidate the underlying reasons for the irregularities or detect patterns between multiple evaluations. Missing or incomplete data as a result of varying response rates also weakened the reliability of the peer evaluations, forcing instructors to either try to fill in the gaps or discard the evaluations.

Addressing Challenges with Interpreting Evaluations. Given the problems discussed so far with interpreting peer evaluations, instructors that wanted to assign grades fairly during team work generally needed overwhelming evidence that a student was an excellent or poor team member before considering making a grading decision. Instructors usually adjusted or assigned individual grades only at the end of the project because more data was available to them. For example, three instructors calculated individual grades for students by aggregating all of the peer evaluation scores that students received. Other instructors would only adjust grades for exceptional cases that appeared as noticeable outliers in the data obtained from across different assessment methods used. For example, a student that consistently received negative evaluations from all of their teammates, rarely contributed during instructor check-ins, and submitted low-quality work was more likely to receive a lower individual grade.

Instructors followed a similar strategy of identifying outliers or discrepancies in one source of data and corroborating with further sources when deciding other courses of action for responding to team conflicts. In the absence of a single high-quality signal and in order to obtain as much information as possible, the primary strategy that eight instructors used was to corroborate data collected using multiple methods. For example, I1, a Computer Science instructor, used a three-prong approach in which they checked contributions to Github and Wiki, conducted peer evaluations, and consulted the teaching staff that regularly met with teams. This allowed them to *"get a reliable overall signal from these rather unreliable sub-signals.*

5.3 Use of Activity Traces (RQ3)

5.3.1 Instructor Use. Five instructors corroborated other types of assessment data with activity traces from tools that students used during the teamwork experience. These traces ranged from Github logs to log-in activity reports on a learning management system (LMS). Activity traces provided more objective metrics of individual contributions and team processes, where *"there's not an incentive to misstate"* (I5) as in peer evaluations. For example, I8 used Google Docs for assignments because they could look at the version history and *"visually see efforts that have been put in by students and where they are contributing."*

Activity traces could also be used to identify or validate dysfunctional behaviors in teams: *"[Google Docs' version history] would particularly be useful to validate if I had a suspicion that someone did little or no work. It would be possible to say "well, I can see that in the data" and justify it"* (I5). I7 asked students to share their group's communication history if they complained to the instructor about an unresponsive team member. This allowed the instructor to confirm the problem, check whether the team had tried to resolve the problem on their own first, and then decide the most appropriate strategy for helping the team. Because activity traces are automatically logged in real time, this also allowed instructors to construct a temporal view of each team and each member's contributions without the administrative overhead of having multiple meetings or peer evaluations throughout the semester.

5.3.2 Student Use. During the student interviews and survey, we focused on what information students considered when assessing each other's contributions. In the interview, we further dug into when students might actually seek further information.

What information do students consider? When students in our interviews were asked about additional information that they thought could support their peer evaluations or help them to evaluate their team members, almost all students (N=10) cited online activity traces, such as version histories, communication logs, deliverables, and time logs. With these traces, S4 stated, *"there's evidence right clear on the screen that everyone has access to that shows who did what"*. To preserve the privacy of the participants, we do not include the actual activity traces they used in their teams. Instead, in Figure 1, we showcase realistic examples of two types of traces students mentioned: code contributions as visualized in contribution graphs on Github and document contributions as displayed in the version history feature on Google Docs.

To investigate how students weigh different types of activity traces when evaluating their teammates, we categorized the tools mentioned in the student interviews and instructor interviews into 10 different types of activity traces (listed in Appendix A.1). We then asked participants in our survey to rate the extent to which they considered each type when evaluating their teammates during their peer evaluations (1=Not at all to 5=A great deal).

96% of survey respondents considered at least one type of activity trace when evaluating their teammates. Document contributions on tools such as Google Docs and Microsoft Word was the most popular type of activity trace (μ=3.88, s=1.25), likely because documents are the most common deliverable used in courses. This was followed by audio/video i.e. synchronous communication on tools such as Zoom, Google Hangouts, and Discord (μ=3.51, s=1.4) then asynchronous communication on messaging platforms like Slack, GroupMe, and WhatsApp. We found that students taking courses in Formal or Applied Sciences also weighed code contributions highly (μ=3.53, s=1.55), as to be expected. Meanwhile, tasks created on task management tools (μ=2.09, s=1.31) and meeting logs (μ=2.34, s=1.33) were the least considered activity traces. Respondents also had the option to list additional data that they considered. Responses that did not correspond to one of the categories that we provided included research materials collected, file shares, and forum posts.

When do students seek this information? Students generally did not consult activity traces while actively completing their evaluation, with the majority of students feeling that they could rely on their memory and experiences instead. This confidence likely stems from working closely with the artifacts that generated these traces throughout the project, as evidenced by six students describing active monitoring of logs throughout the teamwork experience. Students primarily checked these logs to make sure that their teammates were on-task, especially near deadlines, and to review the contributions that their teammates had made. However, if there was a problem in their team or they were going to give a team member a more critical evaluation, three students stated they would consult the activity traces for *"peace of mind"* (S2).

When students actually examined their activity traces more closely, students felt that the traces mostly confirmed their initial evaluations of their team members. However, two students described experiencing shifts in their initial evaluation of their team members. For example, after inspecting the version history on one of their project documents, S6 realized that a teammate had done an even greater share of the work than they remembered.

5.3.3 Challenges for using activity traces. Four of the five instructors that used activity traces in their assessment process only collected one type of trace. However, both instructors and students acknowledged that students use a wide array of tools to conduct team activities, each of which may capture different aspects of students' contributions. Seven students felt that it was most useful to combine the activity traces from different tools in order to understand what teams were doing. This challenged the feasibility for instructors to collect or parse through information from all of these tools. Three instructors pointed out that for coding projects that use version control systems, it is easy to keep track of commits or lines of code, with systems like Github even providing visual summaries of the data. However, not every tool has a built-in version control system and even fewer has user-friendly visualizations to help interpret who or how much someone has contributed. Therefore, the ability to interpret activity on each tool was extremely platform-dependent.

Three students also worried that specific tools may not accurately capture the quality of their contributions or the brainstorming, planning, and overall effort they had put in. For example, S1 stated that although they had the least amount of code and commits on Github, their code was actually one of the most complex portions of the project and seeing the low numbers on Github was *"disheartening"*. Differing roles on the team also meant that some students contributed mostly through one tool such as writing a report while another student may be responsible for creating design files: *"If we're working on a report, some people will type a lot more, some people will type a lot less, some people will spend a lot more time doing an image off the report and then import it into the report or do the actual design in Autocad 3D."* (S5) Students also sometimes worked together, but only one team member was responsible for documenting the contributions. This made it important to understand the context of the contributions, such as whether that contribution took place during a team meeting or while the member was working by themselves. Another concern that two instructors expressed was the potential for students to game the metrics: *"If you just tell someone you must contribute some number of lines, it's trivial to add 10,000 lines that don't do anything useful."* (I1)

6 DISCUSSION

In this study, we identified the emerging use of activity traces in evaluating contributions within teams. We also identified challenges that instructors and students face when using traditional peer evaluations. We now provide design implications for how activity traces can be leveraged to help address these challenges. For several of the design implications, we provide examples demonstrating how the implications might be operationalized in a new genre of data-centric peer evaluation tools.

Visualizing activity traces across tools and channels. The multitude of tools that students use to collaborate makes it difficult for instructors to collect and interpret activity traces efficiently. Automatic collection, aggregation, and visualization of traces will help address this challenge. From our interviews and survey, we identified the most popular types of activity traces, which can serve as an initial testbed for collecting and visualizing activity traces. However, different disciplines may value different types of traces. Therefore, data-centric peer evaluations should consider the contexts in which they may be used in order to decide which tools or platforms they should incorporate and allow the users to weight each data source based on its perceived value.

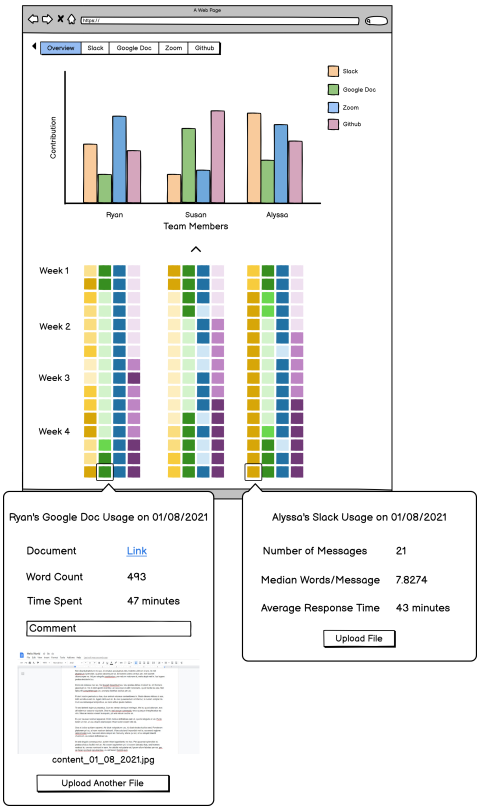
In Figure 2a, we present an example of a visualization dashboard that incorporates tools often used in programming courses. The dashboard visualizes document contributions on Google Doc, communication on Slack and Zoom, and code contributions on Github. Each group of four columns represents a member of the team while each hue and column within a group represents a different tool. The top graph in the overview aggregates contributions over time, allowing users to quickly identify how much each team member contributed during the entire teamwork experience. The bottom graphs represent heat maps where users can see changes in contribution level (represented by color saturation) over each week of the teamwork experience, providing a more nuanced view of each member's contributions.

We can also view specific metrics associated with each tool in Figure 2b to better understand the different dimensions of students' contributions. To satisfy instructors' desire for evidence, a link to this dashboard could be automatically shared in a peer evaluation system to provide objective metrics to support the student's evaluations, as shown in Figure 2c. From the student's perspective, being able to attach impartial evidence for their evaluation addresses students' anxiety over being critical of their team members and helps them interpret the evaluations received.

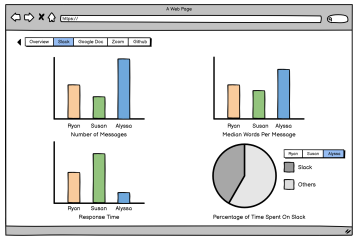
Calibrating evaluation standards. The peer evaluations described by instructors and students were often susceptible to inaccurate or inconsistent data. In particular, students typically erred on the side of leniency when submitting evaluations. Students and instructors attributed these cases to bias and collusion between students, fear of penalizing others' grades, lack of communication within teams, and differing personal standards. We propose that data-centric peer evaluation tools can help deter student misjudgements by calibrating evaluation standards with the activity traces from each team. The system can then "fact-check" students who submit evaluations that are significantly outside the range of what the data shows (e.g. outliers). Because the data itself might not be a perfect signal, students can still be allowed to proceed with their evaluation after seeing such a notification. We demonstrate a simple example of this feature in Figure 2d. The system can also automatically flag discrepancies between the evaluation responses and teams' activity traces to instructors. This will add an additional signal to help instructors identify potential dysfunctions where they should take a closer look at the peer evaluations or at the activity traces.

Allowing student-configured criteria. Another approach to configuring a data-centric peer evaluation system is to allow teams to set their own weights for different types of contributions. For example, if a team values the effort of each member, the peer evaluation system can process total edits in a document. If the team values each member's final contribution to the project, the system can filter for net edits (the content that made it into the final submission). Teams may even want to configure the system differently for different members of the team. Providing students more configuration options within the tool may help to address different work styles, such as when each member has a different role within the team. Allowing students to explicitly set criteria within the peer evaluation system can also help teams to establish clear expectations for contributions and communication at the beginning of a project.

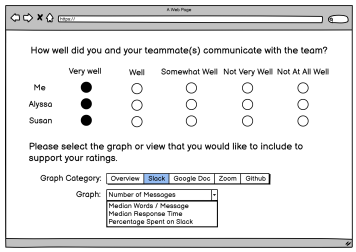
Contextualizing contributions. In our study, we found that instructors value objective, easy-to-interpret quantitative measures, while students value being able to provide specific, open-ended



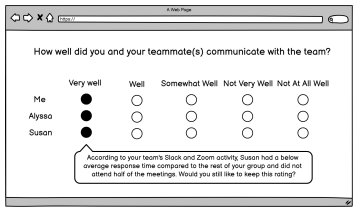
(a) Activity Trace Visualization - Overview



(b) Activity Trace Visualization - Slack



(c) Attaching Data to Support Evaluation



(d) Calibrated Fact-Checker

Fig. 2. Examples of proposed data-centric peer evaluation tool and features. (a) Dashboard that visualizes activity traces of each team member's contributions across different tools and over time. (b) Detailed view of the visualization dashboard showing metrics specific to Slack. (c) Attachment of relevant activity traces to support evaluations. (d) Automatic fact-checker that notifies the student if their evaluation is outside the range of what the data suggests.

comments. A data-centric peer evaluation tool could integrate both perspectives by allowing students to annotate their activity traces. While the traces themselves may indicate quantitative metrics such as number of edits or response time, students can annotate the data with information such as when and where the contribution took place, the difficulty or effort put into the contribution, or extenuating circumstances that affected their ability to contribute. This suggestion also addresses instructor and student concerns about the ability of logs to capture the context of contributions. Figure 2a displays a simple comment box where users can input these annotations.

Tagging highlights and areas of improvement. One of the primary goals for instructors in assessing teams is to help students develop their teamwork skills through feedback. Prior work has shown that students provide higher-quality feedback after repeated uses of peer evaluation systems [10]. However, instructors sometimes had limited opportunities to administer the feedback process or teach students how to give constructive feedback to each other. Only four students in the interviews even mentioned the value of getting feedback during the peer evaluation process,

indicating untapped potential. A data-centric peer evaluation tool could help facilitate feedback by requiring students to highlight both meaningful contributions that each of their team members made and areas for improvement. Therefore, students can see specific examples of their strengths and weaknesses within the tool.

Other approaches to facilitating feedback in data-centric peer evaluations might be drawn from prior work on facilitating high-quality feedback in peer outcome assessments [8, 13, 19]. For example, auto-generated feedback statements could be suggested to an evaluator based on how they rated their team members. Alternatively, the system can ask a student what they would like to improve on and suggest tailored feedback to their team members or highlight the activity traces that are most relevant to the area of improvement. Given enough data over time, the system could even generate feedback statements directly from the patterns surfaced in the activity traces.

7 LIMITATIONS AND FUTURE WORK

Participants were recruited from a single, large public research university. Therefore, our findings may not be representative of students and instructors at other universities. In addition, we tried to recruit a diverse range of disciplines for our interviews, but we did not interview students that focused on courses in the Humanities or Natural Sciences.

We explored implications for using activity traces in peer evaluations in university courses. Future work can further explore the use of activity traces in peer evaluations in other contexts, such as in the workplace and in online work platforms. Prior work on activity reporting in the workplace has showcased difficulties with collecting, composing, and delivering progress reports for knowledge workers [25]. Employees may benefit from being able to use systems that can automatically compose reports for them based on their activity traces or help them to evaluate coworkers impartially.

We used interviews and surveys with students and instructors across different disciplines in order to gain insights into the use of activity traces in peer evaluations. Future work could employ more specific methodologies such as think-alouds and user studies to supplement and evaluate these insights. For example, students could be asked to construct peer evaluations out of their activity traces in order to investigate what information or signals they find most valuable for evaluating their team members. Case studies on a single course or student population may also provide more specific insights into how activity traces are utilized within that setting.

Future research is needed in order to implement and evaluate the design implications we proposed in this paper. These implications may change or evolve when tested in an actual course setting. We acknowledge that some of the features we suggested may require a significant amount of time and resources to implement. However, research on group awareness tools has already made strides in collaborative activity tracking and analytics [16, 30, 40]. For example, Wang et al. were able to identify collaborative writing patterns within groups of students by visualizing and analyzing document revisions [32]. Utilizing methods from this body of work can facilitate the implementation of the design implications.

Finally, we note that researchers should consider issues of privacy and transparency when collecting activity traces in order to avoid creating the *"surveillance environment"* mentioned in Section 5.2.2. However, we have shown that instructors are already using activity traces when evaluating students' contributions. Designing systems that can produce visualizations to abstract some of the details away while still conveying what is useful may help to preserve students' privacy compared to directly collecting the activity traces.

8 CONCLUSION

Peer evaluations are a staple for assessing individual contributions to teamwork, yet are susceptible to bias, social pressure, and collusion. In this paper, we reported the results from a mixed-methods study with university students and instructors aimed at understanding the types of peer evaluations they use, factors that affect how students complete the evaluations, challenges instructors faced with interpreting and using the evaluations in courses, and how activity traces from online collaborative tools are being incorporated throughout the peer evaluation process. We demonstrated how the findings from our study could be implemented in a speculative data-centric peer evaluation tool. By grounding peer evaluations in the activity traces that teams produce and allowing students to contextualize these activity traces, we believe that more fair, accurate, and meaningful evaluations of teams can be achieved.

9 ACKNOWLEDGEMENTS

This work was supported in part by NSF award IIS-2016908. We would like to thank all the students and instructors who voluntarily participated in our study. Additional thanks to the members of our research lab, Silas Hsu, Tiffany Wenting Li, and the anonymous reviewers for their invaluable feedback.

REFERENCES

[1] Praveen Aggarwal and Connie L. O'Brien. 2008. Social Loafing on Group Projects: Structural Antecedents and Effect on Student Satisfaction. *Journal of Marketing Education* 30, 3 (2008), 255–264. https://doi.org/10.1177/0273475308322283 arXiv:https://doi.org/10.1177/0273475308322283

[2] Dmytro Babik, Edward F. Gehringer, Jennifer Kidd, Ferry Pramudianto, and David Tinapple. 2016. Probing the Landscape: Toward a Systematic Taxonomy of Online Peer Assessment Systems in Education. *CEUR Workshop Proceedings* 1633 (2016).

[3] Diane F. Baker. 2008. Peer Assessment in Small Groups: A Comparison of Methods. *Journal of Management Education* 32, 2 (2008), 183–209. https://doi.org/10.1177/1052562907310489 arXiv:https://doi.org/10.1177/1052562907310489

[4] John H. Bernardin, Donna Cooke, and Peter Villanova. 2000. Conscientiousness and Agreeableness as Predictors of Rating Leniency. *The Journal of Applied Psychology* 85 (2000), 232–236. https://doi.org/10.1037/0021-9010.85.2.232

[5] Stéphane Brutus and Magda Donia. 2010. Improving the effectiveness of students in groups with a centralized peer evaluation system. *Academy of Management Learning and Education* 9 (2010), 652–662. Issue 4. https://doi.org/10.5465/AMLE.2010.56659882

[6] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173868

[7] Amy Cook, Steven Dow, and Jessica Hammer. 2020. Designing Interactive Scaffolds to Encourage Reflection on Peer Feedback. In *Proceedings of the 2020 ACM Conference on Designing Interactive Systems*. 1143–1153. https://doi.org/10.1145/3357236.3395480

[8] Amy Cook, Jessica Hammer, Salma Elsayed-Ali, and Steven Dow. 2019. How Guiding Questions Facilitate Feedback Exchange in Project-Based Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300368

[9] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1277–1286. https://doi.org/10.1145/2145204.2145396

[10] Magda Donia, Thomas A. O'Neill, and Stephane Brutus. 2015. "Peer Feedback Increases Team Member Performance, Confidence and Work Outcomes: A Longitudinal Study". *Academy of Management Proceedings* 2015, 1 (2015), 12560. https://doi.org/10.5465/ambpp.2015.21 arXiv:https://doi.org/10.5465/ambpp.2015.21

[11] Martin R. Fellenz. 2006. Toward Fairness in Assessing Student Groupwork: A Protocol for Peer Evaluation of Individual Contributions. *Journal of Management Education* 30, 4 (2006), 570–591. https://doi.org/10.1177/1052562906286713 arXiv:https://doi.org/10.1177/1052562906286713

[12] Johan Forsell, Karin Forslund Frykedal, and Eva Hammar Chiriac. 2020. Group Work Assessment: Assessing Social Skills at Group Level. *Small Group Research* 51, 1 (2020), 87–124. https://doi.org/10.1177/1046496419878269 arXiv:https://doi.org/10.1177/1046496419878269

[13] C. Ailie Fraser, Tricia J. Ngoon, Ariel S. Weingarten, Mira Dontcheva, and Scott Klemmer. 2017. CritiqueKit: A Mixed-Initiative, Real-Time Interface For Improving Feedback. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3131785.3131791

[14] Mark Freeman and Jo McKenzie. 2002. SPARK, a confidential web–based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects. *British Journal of Educational Technology* 33, 5 (2002), 551–569. https://doi.org/10.1111/1467-8535.00291 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8535.00291

[15] Karin Forslund Frykedal and Eva Hammar Chiriac. 2011. Assessment of students' learning when working in groups. *Educational Research* 53, 3 (2011), 331–345. https://doi.org/10.1080/00131881.2011.598661 arXiv:https://doi.org/10.1080/00131881.2011.598661

[16] Siwei Fu, Jian Zhao, Hao Fei Cheng, Haiyi Zhu, and Jennifer Marlow. 2018. T-Cal: Understanding Team Conversational Data with Calendar-Based Visualization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174074

[17] Graham R Gibbs. 2007. Thematic coding and categorizing. *Analyzing qualitative data. London: Sage* (2007), 38–56.

[18] Lisa E. Gueldenzoph and Gary L. May. 2002. Collaborative Peer Evaluation: Best Practices for Group Member Assessments. *Business Communication Quarterly* 65, 1 (2002), 9–20. https://doi.org/10.1177/108056990206500102 arXiv:https://doi.org/10.1177/108056990206500102

[19] Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features That Support High Quality Peer Assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 458–469. https://doi.org/10.1145/2858036.2858195

[20] Edward J. Inderrieden, Robert E. Allen, and Timothy J. Keaveny. 2004. Managerial Discretion in the Use of Self-ratings in an Appraisal System: The Antecedents and Consequences. *Journal of Managerial Issues* 16, 4 (2004), 460–482. http://www.jstor.org/stable/40604464

[21] Sneha R. Krishna Kumaran, Deana C. McDonagh, and Brian P. Bailey. 2017. Increasing Quality and Involvement in Online Peer Feedback Exchange. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 63 (Dec. 2017), 18 pages. https://doi.org/10.1145/3134698

[22] Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (Vancouver, BC, Canada) *(L@S '15)*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/2724660.2724670

[23] Steve Loddington, Keith Pond, Nicola Wilkinson, and Peter Willmot. 2009. A case study of the development of WebPA: An online peer-moderated marking tool. *British Journal of Educational Technology* 40, 2 (2009), 329–341. https://doi.org/10.1111/j.1467-8535.2008.00922.x arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8535.2008.00922.x

[24] Misty L. Loughry, Matthew W. Ohland, and David J. Woehr. 2014. Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools. *Journal of Marketing Education* 36, 1 (2014), 5–19. https://doi.org/10.1177/0273475313499023 arXiv:https://doi.org/10.1177/0273475313499023

[25] Di Lu, Jennifer Marlow, Rafal Kocielnik, and Daniel Avrahami. 2018. Challenges and Opportunities for Technology-Supported Activity Reporting in the Workplace. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173744

[26] Andrew Luxton-Reilly. 2009. A systematic review of tools that support peer assessment. *Computer Science Education* 19 (2009), 209–232. Issue 4. https://doi.org/10.1080/08993400903384844

[27] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. *Impression Formation in Online Peer Production: Activity Traces and Personal Profiles in Github*. Association for Computing Machinery, New York, NY, USA, 117–128. https://doi-org.proxy2.library.illinois.edu/10.1145/2441776.2441792

[28] Jennifer Marlow and Laura A. Dabbish. 2015. The Effects of Visualizing Activity History on Attitudes and Behaviors in a Peer Production Context. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 757–764. https://doi.org/10.1145/2675133.2675250

[29] Neal P. Mero, Rebecca M. Guidice, and Amy L. Brownlee. 2007. Accountability in a Performance Appraisal Context: The Effect of Audience and Form of Accounting on Rater Response and Behavior. *Journal of Management* 33, 2 (2007), 223–252. https://doi.org/10.1177/0149206306297633 arXiv:https://doi.org/10.1177/0149206306297633

[30] Sean A. Munson, Karina Kervin, and Lionel P. Robert. 2014. Monitoring Email to Indicate Project Team Performance and Mutual Attraction. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 542–549. https://doi.org/10.1145/2531602.2531628

[31] Matthew W. Ohland, Misty L. Loughry, David J. Woehr, Lisa G. Bullard, Richard M. Felder, Cynthia J. Finelli, Richard A. Layton, Hal R. Pomeranz, and Douglas G. Schmucker. 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, 4 (2012), 609–630. https://doi.org/10.5465/amle.2010.0177 arXiv:https://doi.org/10.5465/amle.2010.0177

[32] Judith S. Olson, Dakuo Wang, Gary M. Olson, and Jingwen Zhang. 2017. How People Write Together Now: Beginning the Investigation with Advanced Undergraduates in a Project Course. *ACM Trans. Comput.-Hum. Interact.* 24, 1, Article 4 (March 2017), 40 pages. https://doi.org/10.1145/3038919

[33] Vicente Peñarroja, Virginia Orengo, and Ana Zornoza. 2017. Reducing perceived social loafing in virtual teams: The effect of team feedback with guided reflexivity. *Journal of Applied Social Psychology* 47, 8 (2017), 424–435. https://doi.org/10.1111/jasp.12449 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jasp.12449

[34] David Rozas, Nigel Gilbert, Paul Hodkinson, and Samer Hassan. 2021. Talk Is Silver, Code Is Gold? Beyond Traditional Notions of Contribution in Peer Production: The Case of Drupal. *Frontiers in Human Dynamics* 3 (2021), 12. https://doi.org/10.3389/fhumd.2021.618207

[35] Anita Sarma, Xiaofan Chen, Sandeep Kuttal, Laura Dabbish, and Zhendong Wang. 2016. Hiring in the Global Stage: Profiles of Online Contributions. In *2016 IEEE 11th International Conference on Global Software Engineering (ICGSE)*. 1–10. https://doi.org/10.1109/ICGSE.2016.35

[36] N. Sadat Shami, Kate Ehrlich, Geri Gay, and Jeffrey T. Hancock. 2009. *Making Sense of Strangers' Expertise from Signals in Digital Artifacts*. Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/1518701.1518713

[37] Amy Shannon, Jessica Hammer, Hassler Thurston, Natalie Diehl, and Steven Dow. 2016. PeerPresents: A Web-Based System for In-Class Peer Feedback during Student Presentations. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 447–458. https://doi.org/10.1145/2901790.2901816

[38] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. 2013. Mutual Assessment in the Social Programmer Ecosystem: An Empirical Investigation of Developer Profile Aggregators. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 103–116. https://doi.org/10.1145/2441776.2441791

[39] Simon Taggar and Travor C. Brown. 2006. Interpersonal affect and peer rating bias in teams. *Small Group Research* 37 (2006), 86–111. Issue 1. https://doi.org/10.1177/1046496405284382

[40] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1865–1874. https://doi.org/10.1145/2702123.2702517

[41] Bo Zhang and Matthew W. Ohland. 2009. How to Assign Individualized Scores on a Group Project: An Empirical Evaluation. *Applied Measurement in Education* 22, 3 (2009), 290–308. https://doi.org/10.1080/08957340902984075

A APPENDIX: STUDENT SURVEY

A.1 Role of Data in Peer Evaluation

We are interested in whether you considered any data when evaluating your group members. By data, we refer to any artifacts or information produced by the digital tools that you used throughout the group work experience. Examples include document edits, logs, meeting notes, messages, emails, code commits, etc.

The list of data we present below is not exhaustive but you will have the opportunity to share other types of data that you considered afterwards.

(1) To what extent did you consider the following types of data when evaluating your group members in this group work experience? *[Likert-7: Not at all, A little, A moderate amount, A lot, A great deal, Not applicable]*
   - Google Doc/Microsoft Word edits
   - Google Slides/Powerpoint edits
   - Code/commits
   - Digital drafts/prototypes/models
   - Tasks on a task management tool (e.g. Trello, Asana, etc.)
   - Wiki edits
   - Meeting logs
   - Emails
   - Messages (e.g. Slack, GroupMe, WhatsApp, text, etc.)
   - Video/audio calls (e.g. Zoom, Google Hangouts, Discord, etc.)
   - Please check "A great deal" for this row.

(2) What other data, if any, did you consider in evaluating the other members of your group? Please list them and separate each type of data by a new line. *[Optional open-ended response]*

A.2 Perceptions of Peer Evaluation

We would now like to understand your thoughts and perceptions about the peer evaluation/s that you completed in this group work experience. Please answer as honestly as you can.

(1) How much do you agree with the following statements about **your evaluation of your group members** for the peer evaluation/s that you completed in this group work experience? *[Likert-5: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree]*
   - I was able to accurately evaluate the other members' individual contributions to the group.
   - I expressed any conflicts or issues within my group in the peer evaluation.
   - I felt pressured to give one or more group members a more favorable evaluation than they deserved.
   - I felt pressured to give one or more group members a less favorable evaluation than they deserved.
   - I did not want to hurt my group members' grades as a result of my evaluation.
   - Please check "Somewhat agree" for this row.

(2) How much do you agree with the following statements about **how you and other members were evaluated by the rest of your group** for the peer evaluation/s that you completed in this group work experience? *[Likert-5: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree]*
   - I am confident that I received a fair evaluation from my other group members.
   - I am confident that my other group members received fair evaluations from the rest of the group.

(3) Overall, how satisfied were you with the peer evaluation/s that you completed for this group work experience? *[Likert-5: Extremely satisfied, Somewhat satisfied, Neither satisfied nor dissatisfied, Somewhat dissatisfied, Extremely dissatisfied]*

(4) Do you have any other comments on the topics mentioned in this survey? *[Optional open-ended response]*

Received January 2021; revised April 2021; accepted July 2021