The Value of Activity Traces in Peer Evaluations: An Experimental Study

WENXUAN WENDY SHI, University of Illinois, Urbana-Champaign, USA
SNEHA R. KRISHNA KUMARAN, University of Illinois, Urbana-Champaign, USA
HARI SUNDARAM, University of Illinois, Urbana-Champaign, USA
BRIAN P. BAILEY, University of Illinois, Urbana-Champaign, USA

Peer evaluations are a well-established tool for evaluating individual and team performance in collaborative contexts, but are susceptible to social and cognitive biases. Current peer evaluation tools have also yet to address the unique opportunities that online collaborative technologies provide for addressing these biases. In this work, we explore the potential of one such opportunity for peer evaluations: data traces automatically generated by collaborative tools, which we refer to as "activity traces". We conduct a between-subjects experiment with 101 students and MTurk workers, investigating the effects of reviewing activity traces on peer evaluations of team members in an online collaborative task. Our findings show that the usage of activity traces led participants to make more and greater revisions to their evaluations compared to a control condition. These revisions also increased the consistency and participants' perceived accuracy of the evaluations that they received. Our findings demonstrate the value of activity traces as an approach for performing more reliable and objective peer evaluations of teamwork. Based on our findings as well as qualitative analysis of free-form responses in our study, we also identify and discuss key considerations and design recommendations for incorporating activity traces into real-world peer evaluation systems.

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

Additional Key Words and Phrases: team assessment; peer evaluation; activity traces; online collaboration

ACM Reference Format:
Wenxuan Wendy Shi, Sneha R. Krishna Kumaran, Hari Sundaram, and Brian P. Bailey. 2023. The Value of Activity Traces in Peer Evaluations: An Experimental Study. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 151 (April 2023), 39 pages. https://doi.org/10.1145/3579627

1 INTRODUCTION

Peer evaluations are a well-established framework for evaluating individual and team performance in collaborative environments. They often form the basis for critical decisions made in academia and industry, including assigning individual grades, giving promotions and rewards, and identifying problems within teams [1, 12, 55]. However, peer evaluations are often susceptible to leniency errors, memory limitations, and other social and cognitive factors that compromise the reliability and validity of the evaluations [4, 21, 30, 49]. Peer evaluations are also inherently subjective. Raters tend to evaluate each other based on their own set of heuristics and standards, rather than using

Authors' addresses: Wenxuan Wendy Shi, wshi16@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Sneha R. Krishna Kumaran, srkrish2@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Hari Sundaram, hs1@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Brian P. Bailey, bpbailey@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA.

2573-0142/2023/4-ART151 $15.00
https://doi.org/10.1145/3579627

objective, independent criteria [40]. The shift towards technology-mediated collaboration in teams has further complicated these problems. Many teams in academia and industry now work remotely with the majority of their work being performed online. Distributed teams have lower visibility into each members' individual contributions which may lead to greater reliance on potentially biased assumptions and perceptions [38].

Researchers have explored many tools and approaches for improving the quality of peer evaluations, including different rating scale formats and criteria [3, 19], maintaining confidentiality of ratings [37], and rater training exercises [34]. Advances in technology have also led to the development of online peer evaluation systems that automate the process of administering evaluations and incorporate such features [14, 24, 34]. However, issues of social and cognitive bias still persist in these systems [13, 33]. Furthermore, current peer evaluation tools have yet to address the challenges that advancements in online collaborative technologies have introduced. At the same time, these advancements provide unique opportunities for scaffolding the peer evaluation process.

In this paper, we explore the value of one such opportunity for peer evaluations: data traces automatically generated by collaborative tools. We refer to these traces as "activity traces". For example, Google Docs logs each edit, who made the edit, and the time of the edit in a collaborative document. These traces can be aggregated into a visual or written summary such as Google Doc's document version history. Activity traces show more than just the final outcome of a team; they build a representation of the individual and team processes that occurred over time to reach the outcome. Activity traces from collaborative tools, such as Slack, Github, and Zoom, can also provide evidence for different aspects of teamwork, such as technical contributions, communication, and meeting participation.

The CSCW community has made extensive contributions to the development of systems and tools that draw on activity traces to support awareness and coordination in collaborative settings [11, 46, 54]. Research has also shown that activity traces influence social inferences that individuals make about each other in online groups and communities [27, 42]. While prior work has explored tools for collaborative awareness and observed how activity traces can influence informal impressions of others, little work has examined the effects of using activity traces for peer evaluations. As objective records of contribution, activity traces may reduce team members' reliance on memory and subjective standards to evaluate each others' contributions. In a classroom setting, instructors consider activity traces a robust source of data and use them to corroborate peer evaluations in student teams [43] However, in this context, activity traces are only incorporated as evidence after the evaluations actually occur. There is a need for greater empirical work exploring the impact of using activity traces during the peer evaluation itself. For example, to what extent does reviewing activity traces affect team members' initial evaluations from memory? Do traces improve the consistency of evaluations received by each team member? How does reviewing activity traces affect team members' attitudes towards the peer evaluation process?

To answer these questions, we conducted a between-subjects experiment (N=101), investigating the effects of using activity traces when evaluating team members on a simulated collaborative task. The context of the study captured several of the social and cognitive factors that exist in real-world peer evaluation cases. Participants performed a creative, open-ended task within teams and evaluated each other's contributions to the task after a one-day delay. They were informed that the evaluations would impact the bonuses that each member receives, adding real stakes to the evaluations for participants. In addition, each ratee's evaluation ratings would be aggregated, anonymized, and shared with them at the end of the study. Participants were recruited from two different populations: university students and Amazon Mechanical Turk (MTurk) workers.

In our experiment, participants first completed an initial evaluation of each team member's contributions based on memory. Participants were then assigned to complete one of two reflection

activities. In one condition, participants were asked to review activity traces from the collaborative task while writing a reflection on their team experience. The activity traces consisted of the version history and chat log that were generated by the Google document each team worked on (shown in Figure 2). These traces revealed information about the editing behavior and communication of each team member throughout the task. In the other condition, participants wrote the reflection without being aware of or given access to the activity traces. These conditions represented our treatment and control respectively. After writing the reflection, participants in both conditions were then able to revise their evaluations of their teammates. By comparing pre-post differences between the initial and revised evaluations, we isolate the effect of the activity traces on participants' evaluations. Follow-up questions after the evaluation asked participants about their attitudes towards the evaluation process. Participants in the treatment condition also provided potential benefits and concerns that they perceived with using activity traces to evaluate peers.

Our findings show that participants who reviewed their activity traces during the reflection made larger revisions to their initial evaluations and revised more team members' evaluations. The size of revisions made by participants in the treatment condition was on average more than twice the size of revisions made in the control condition. Furthermore, the consistency of evaluations among team members significantly increased after participants revised their evaluations in the treatment condition, but not in the control condition. In other words, the final evaluations that each participant received from their team members became more aligned with each other compared to the initial evaluations. Participants in the treatment condition also perceived the evaluation scores they received as being more accurate. We observed that the perceived difficulty of the evaluation did not significantly differ between conditions, suggesting that reviewing activity traces can improve peer evaluations without increasing perceived cognitive load. Finally, the open-ended responses in our study show that participants perceived activity logs as being helpful in scaffolding and sometimes debiasing their memory of their team members' contributions. At the same time, participants shared concerns about the information presented in the logs and how that information might be interpreted or actuated within real-world teams.

This paper makes two major contributions to the CSCW community:

Empirical evidence for the value of activity traces. We provide empirical evidence that using activity traces can address social and cognitive limitations of peer evaluations and increase the consistency among evaluators in the context of a creative collaborative task. While prior work has primarily examined how activity traces can improve awareness and coordination in teams, we examine the effects of using activity traces during the peer evaluation process itself. Based on a controlled experiment, our findings show that activity traces led participants to make more and larger revisions to their initial evaluations from memory and those revisions increased the consistency of evaluations. These findings demonstrate the value of activity traces as an approach for performing more reliable and objective peer evaluations of contributions to teamwork to better inform critical decisions within academia and industry.

Recommendations for design and use. We provide implications based on our findings for incorporating activity traces in real-world evaluation systems in a manner that is considerate of users' needs and concerns, the purpose of the evaluation, and the context of the teamwork. For example, we recommend that data-driven peer evaluation systems provide greater control to users over how their contributions are represented and incorporate strategies to identify quality contributions. Our qualitative analysis of participants' perceived benefits and concerns for using activity traces and other open-ended responses in our study suggest opportunities for new tools and techniques that can more effectively leverage activity traces for peer evaluations.

2 RELATED WORK

We situate our work in the context of prior literature on peer evaluations and collaborative awareness. Our goal is to examine the value of activity traces as a new approach to performing peer evaluations. To motivate this goal, we first provide an overview of how peer evaluations are used today, the challenges that arise, and existing tools and approaches to address some of these challenges. We then describe how peer evaluations intersect the literature on collaborative awareness through the usage of activity traces.

2.1 Peer Evaluations

Team assessment comprises many types of evaluation and feedback that occur within teams [13]. This work focuses on peer evaluations as a prominent and critical method for assessing teams at an individual and team level.

2.1.1 Usage and Benefits of Peer Evaluations. In a peer evaluation, each member of the team generally must differentiate individual contributions to the teamwork process and outcomes over time. As active participants, team members are able to observe individual behaviors and group dynamics that other stakeholders such as instructors or managers cannot [31]. Self-assessment is often incorporated in the peer evaluation process as well for similar reasons [17]. Within both educational and work settings, peer evaluations are used to make critical decisions. For example, instructors often use peer evaluations to account for differences in individual contributions and assign grades fairly [12, 55]. In work organizations, self- and peer evaluations contribute to administrative decisions such as promotions and raises [1]. Team members can also diagnose their own strengths and weaknesses as a team member and learn teamwork skills through consistent usage of peer evaluations [10]. Given these stakes, it is crucial that the peer evaluation process is reliable and provides high-quality information for all stakeholders.

2.1.2 Challenges for Peer Evaluations. In practice, self- and peer evaluations are susceptible to a number of social and cognitive biases. Raters form general impressions of their team members that strongly influence how they organize, interpret, and recall performance-related information [8]. These impressions can overwhelm and even distort their memory of the actual behaviors [32, 49]. Thus, peer evaluations are vulnerable to selective or biased recall. The timing of peer evaluations is also often delayed from when the team actually performs the project or task being evaluated, further straining team members' memory of events. In addition, self- and peer evaluations are prone to self-enhancement bias as well as anchoring and adjustment heuristics that compromise the reliability of the evaluations [40]. Poor performers may be especially ill-equipped to assess their team members' contributions accurately as they are unable to appropriately calibrate their standards for evaluation [40].

In online contexts, these issues may be exacerbated by the distributed nature of collaboration and reduced visibility into team members' contributions [53]. Without in-person cues, it can also be easier to misjudge online contributions [38]. Activity traces can potentially provide reliable and objective cues for contribution in online teamwork that can mitigate social biases and limitations of memory. Our study explores this potential in-depth by examining the effects of activity traces on the consistency and perceived accuracy of peer evaluations received in an online collaborative task.

2.1.3 Tools and Methods for Peer Evaluations. Peer evaluations may take a number of forms (e.g. peer rankings, peer nominations, peer ratings) and incorporate different dimensions of teamwork [19]. Tools have also been developed to aid the evaluation process by providing standardized criteria and scales [3, 19], incorporating training exercises for raters to practice evaluating team members [34], and highlighting patterns in the evaluation results that deserve closer inspection

[25]. Several automated peer evaluation systems combine multiple of the features described and increase the ease and efficiency of the evaluation process [14, 24, 34]. These systems and tools have been incorporated in organizations and educational institutions across the globe. Despite the benefits of these systems, issues of social and cognitive bias still persist with peer evaluations [13, 33].

Furthermore, over the past few decades, the development of collaborative technologies has rapidly outpaced the development of peer evaluation tools, with CSCW researchers spearheading many of the advancements. This gap motivates a greater push for the CSCW community to investigate how the same technologies developed to help people work together can also impact how people evaluate each other. In this work, we take a step towards addressing this gap by investigating the impact of using activity traces automatically generated from collaborative tools on the peer evaluation process.

2.2 Activity Traces for Collaborative Awareness

There is a long history of research on systems for supporting activity awareness in collaborative settings within the CSCW community. Much of this work has focused on facilitating monitoring and coordination of activities between collaborators, often in real-time [11, 18, 46]. These systems support social inferences about others' activities, characteristics of the individuals performing the activities, and the structure of a group of individuals [42, 47, 52]. 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 [6, 44]. These social inferences can, in turn, influence how people evaluate contributions in a group or community. For example, prior work has shown that a more detailed visual format of previous work history can induce positive impressions of another worker that persist into evaluations of that worker, regardless of the actual quality of their current outcome [28].

Aggregated activity traces have also been incorporated into awareness features in many commercial collaborative platforms. For example, Google Docs automatically logs each edit to the document, the user who made the edit, and the timestamp of the edit and displays this information in the version history of the document. Several platforms such as Github and Slack provide analytics dashboards that visualize users' activity traces over time. The abundance of activity traces available in these tools and their ability to reveal detailed information about the relative contributions of team members highlight the significant potential for using activity traces to support peer evaluations. Prior work has already identified the emergent usage of activity traces from collaborative tools to corroborate peer evaluations in student teams [43].

Overall, however, there have been few studies explicitly examining the impact of activity traces on peer evaluations of collaborative work. In this paper, we address this gap by conducting an experiment to investigate the effects of reviewing activity traces on peer evaluations for a simulated collaborative task. Our study extends prior work by isolating and quantifying the direct effects of activity traces on team members' initial evaluations from memory, assessing the quality of evaluations through consistency, and exploring participants' perceptions of the peer evaluation process.

3 RESEARCH QUESTIONS

The goal of this study is to understand whether and how using activity traces can improve the peer evaluation process. Evaluating the quality of a peer evaluation is not straightforward, however. Accuracy is difficult to measure directly as it typically requires rigorous constraints over the task being performed as well as multiple experts [48]. In many real-world teams, collaborative work is open-ended, creative, and not well-structured. Consistency among feedback sources (i.e. the

different members of the team) as well as perceptions of the evaluation process serve as more valuable measures of the quality of evaluations [9]. Although we might not be able to determine the accuracy of an evaluation, we can hopefully reach a common standard that makes scores consistent and is transparent about where the scores come from. Thus, we ask the following research questions:

RQ1: How does reviewing their team's activity traces affect participants' initial evaluations of team members from memory?

RQ2: How does reviewing their team's activity traces affect the consistency of evaluations that participants receive from their team members?

RQ3: How does reviewing their team's activity traces affect participants' attitudes towards the evaluations they performed and received?

Finally, to better understand how to incorporate activity traces effectively into peer evaluations in the real world, we ask:

RQ4: What are participants' perceived benefits and concerns for using activity traces for evaluating their team members?

These questions are not exhaustive but provide a useful starting point for examining the effects of using activity traces to support the peer evaluation process.

4 METHOD

To investigate our research questions, we conducted a between-subjects online experiment comparing participants' evaluations of their teammates' contributions to a collaborative task. The purpose of the collaborative task was to simulate a real-world collaborative decision-making task and to serve as a pretext for evaluating team members. We chose an ad writing task because it was open-ended and complex, did not require previous knowledge or background, and automatically generated activity traces. We first detail our study procedure below (Section 3.1), including discussion of our choice of task and experimental design. We then describe our participant recruitment process (Section 3.2), our measures (Section 3.3), and our data analysis methods (Section 3.4).

4.1 Experimental Flow

The experiment consisted of four primary steps: 1) demographic and availability survey, 2) scheduled collaborative activity, 3) peer evaluation survey, and 4) feedback survey. Figure 1 summarizes the experimental procedure.

4.1.1 Demographic and Availability Survey. After consenting to the study, participants completed a brief demographic survey where they were asked about their gender, age, English proficiency, education level, writing experience, and experience with collaboration and peer evaluations of team members. We also included open-ended questions in the survey that asked participants to describe the main challenges they have encountered with collaboration and peer evaluations. We used these questions to filter out facetious and bad-faith responses. Finally, participants provided their availability for the scheduled collaborative activity.

4.1.2 Collaborative Activity. Participants were grouped into teams for the collaborative activity based on their scheduling availability. Participants did not know each other prior to the activity. Thus, our results should be interpreted in the context of a newly-formed team. We discuss this aspect in more detail in the Discussion.

We assigned teams the task of creating an ad, based on prior work assessing creative outcomes in collaborative environments [26, 41]. The ad design task fulfilled key criteria. First, the task was open-ended and complex, involving processes common to collaborative projects such as decision-making, consensus-building, and communication while also allowing for individual creativity. Second, participants were not required to have any previous knowledge or background to complete

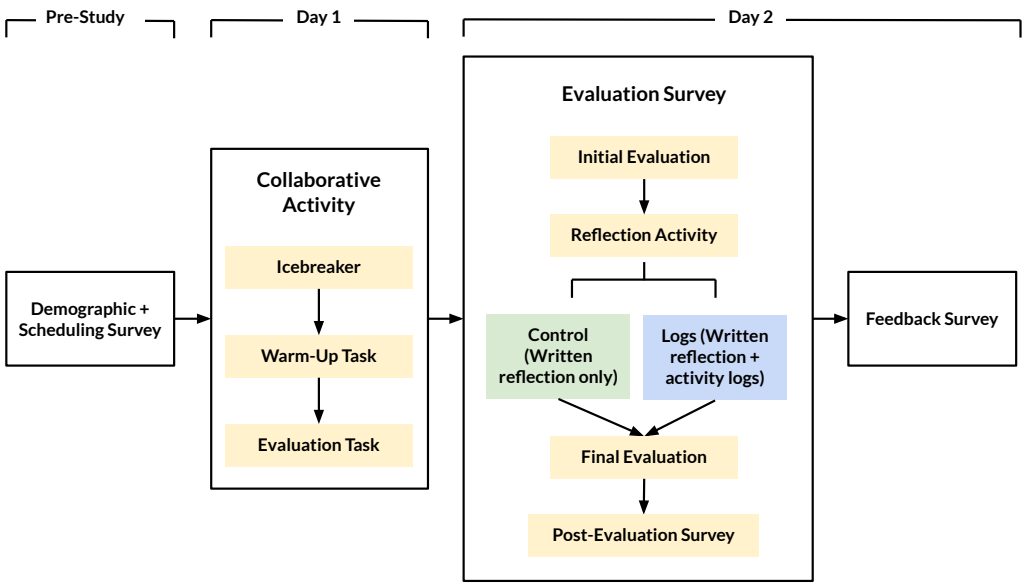
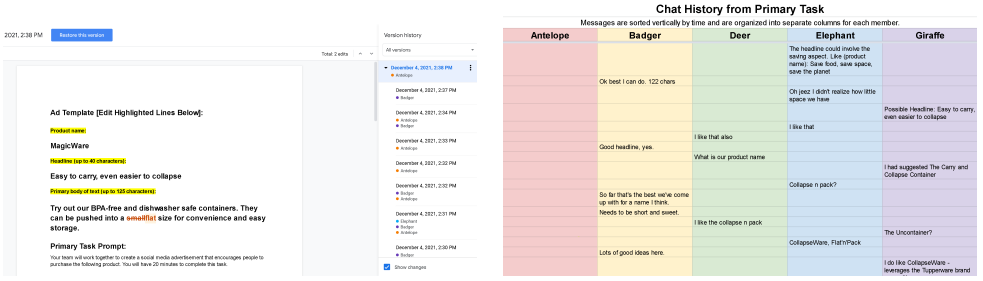


Fig. 1. Experimental flow

the task. Finally, the task fit naturally in the framework of Google Docs, which allowed participants to simultaneously edit the same document and communicate through the embedded chat feature. More importantly, these features generated detailed summaries of activity traces for the editing behavior and communication of each team member throughout the task.

On the first day, participants opened a survey at their scheduled time slot. To access the shared Google Doc in which participants would be working, the survey guided each participant to log into a Google account with a pseudonym such as Antelope or Badger to hide their identity and reduce the risk of racial or gender bias. Teams began with a warm-up task in order to become familiar with each other, the task instructions, and working in a shared document. For the warm-up, participants were linked to a Kickstarter campaign and asked to create an advertisement headline for the product in the campaign. The headline had a maximum character limit of 40 characters. After reading the instructions for the warm-up, participants performed a brief icebreaker where they introduced themselves to their teammates and shared an interesting or relevant piece of information they found from the product campaign. They then had 10 minutes to create the headline.

Following the warm-up, teams were directed to their primary task where they had 20 minutes to create a social media advertisement for a collapsible food storage container. We chose this product because it is a common household item that many participants are likely familiar with or could imagine using. Teams had to create an advertisement for the product, consisting of an ad headline and body. Character limits were imposed on the headline and body based on the requirements for Facebook advertisements (headline up to 40 characters and body up to 125 characters). Teams were also asked to come up with and include a name for the product in the ad. Participants were given only an image of the product and a short, two-line description. They were encouraged to perform background research on similar products in order to develop the advertisement. Participants were informed that their bonus payment would be determined based on the quality of their primary task outcome and how they are evaluated by their team members in the peer evaluation.



(a) Document version history (b) Chat history

Fig. 2. This figure shows the activity logs from one of the teams in our study.

4.1.3 Evaluation Survey. The day after completing the collaborative activity, participants were first sent a link to a 15-minute evaluation survey. We delayed the evaluation by a day in order to simulate real-world peer evaluation contexts where evaluations are often delayed from when the team actually performed the project or task being evaluated. We could then examine whether participants demonstrate limited or biased recall and how that interacted with the activity traces.

The evaluation survey asked participants to evaluate each of their team members' contributions to the primary task. Participants were informed that their evaluations could affect the bonuses that their team members receive and that their team members would be able to see the scores they receive. Belief that there are actual consequences to the evaluation results may increase participants' motivation to perform the evaluation [51]. At the same time, participants may feel more reluctant to submit low evaluations knowing that it may reduce other people's bonus payments and that their team members would be able to see the results [8]. These circumstances reflect the social context of many real-world peer evaluations.

Participants first rated their overall satisfaction with their team and then rated the percentage share of work that each team member contributed, including themselves. We chose a forced-distribution rating scheme because it was simple, efficient, and allowed participants to differentiate the relative contributions of each member [19]. Participants distributed a total score of 100 between each of their team members. For example, in a team of five where every member contributed equally, each member would receive a score of 20. Participants were also asked to provide explanations for each score that they allocated to a team member.

Participants were then asked to write a reflection on their team experience, thinking about the effectiveness of each contribution throughout the task. For the reflection, teams were randomly assigned to either the treatment condition or the control condition. All members within a team shared the same condition. Participants in the treatment condition were asked to review their team's activity traces in the form of the document version history and chat history while writing their reflection. Participants in the control condition were not informed about the activity traces and did not have access to them.

Activity traces from one of the teams in our study are shown in Figure 2. Because Google Docs does not automatically save the chat log once the document is closed, we saved all messages in a spreadsheet which we linked to participants. Based on pilot tests which found the chat logs to be difficult to parse, we added some visual organization to the spreadsheet by organizing messages by team member/column. However, the level of detail of information remained the same between the original chat log and the reformatted chat log. We confirmed that participants in the treatment

Table 1. Participant demographics

	Student (N=44)	**MTurk** (N=57)	**All** (N=101)
Gender			
Female	25 (56.8%)	21 (36.8%)	46 (45.5%)
Male	18 (40.9%)	36 (63.2%)	54 (53.5%)
No answer	1 (2%)	0 (0%)	1 (1%)
Age			
18-24	33 (75%)	1 (1.8%)	34 (33.7%)
25-34	9 (20.5%)	23 (40.4%)	32 (31.7%)
35-44	2 (4.5%)	21 (36.8%)	23 (22.8%)
45-54	-	8 (14%)	8 (7.9%)
55+	-	4 (7%)	4 (4%)
Education			
Less than high school	-	1 (1.8%)	1 (1%)
High school degree or equivalent	-	8 (14%)	8 (8%)
Some college but no degree	24 (54.6%)	9 (15.8%)	33 (32.7%)
Associates degree	-	5 (8.8%)	5 (5%)
Bachelors degree	20 (45.5%)	23 (40.4%)	43 (42.6%)
Graduate degree	-	11 (19.3%)	11 (10.1%)

condition accessed the activity traces by checking the log-in activity for their Google accounts and viewing activity for the documents.

Following the reflection activity, participants were allowed to revise their initial evaluations. If the participants revised their initial evaluations, they were asked to describe their revisions and explain why they made them. Through this evaluation-reflection-evaluation design, we were able to examine revisions that participants made to their initial evaluation. We incorporated the reflection activity in order to reduce the possible confound that participants in the treatment condition revised their evaluations simply because they had an additional opportunity to reflect on their team's contributions.

The final section of the survey asked about participants' attitudes towards the evaluations they gave and the overall evaluation process, including their confidence in the accuracy of the evaluations and their perceived difficulty with evaluating their team members. Participants could also provide an open-ended response if they wanted to appeal the evaluations they received. Participants in the treatment condition were asked additional open-ended questions about the potential benefits and concerns of using activity traces when evaluating team members in real-world collaborative situations.

4.1.4 Feedback Survey. After collecting all responses to the evaluation survey from a team, we aggregated the evaluation scores that each team member received, not including self-evaluations. Participants next responded to a final 5-minute survey which informed them of the evaluation scores that they received from their team members. Participants rated their perceived accuracy of and satisfaction with the evaluations they received. Once the final survey was submitted, participants were debriefed on the true purpose of the study and informed that they would receive the maximum bonus of $2

4.2 Participant Recruitment

To increase the generalizability of our findings, we recruited participants from two different populations: students and Amazon Mechanical Turk (MTurk) workers. Students frequently perform teamwork in the classroom and MTurkers are likely to collaborate with others on crowdwork platforms or in the workplace. The Mann-Whitney U-Test did not find significant differences between MTurk or student participants for either collaborative experience ($p > 0.6$) or peer evaluation experience ($p > 0.5$). Collaborative experience was rated highly across both populations (M=5, σ=1.85) (7-point Likert scale). The majority of participants (79%) had also completed at least one peer evaluation in a collaborative context. Further demographic information for both student and MTurk participants are shown in Table 1.

We also note that MTurkers are more financially motivated compared to participants recruited through other methods [35, 36]. Since we told participants that their bonuses related to their evaluations, we hypothesized that MTurkers' financial motivations might affect their evaluation scores, such as by inflating their own scores to maximize their bonuses. Thus we include the recruitment population (MTurk vs student) as a covariate in our statistical models.

Recruitment occurred separately for both populations such that students and MTurkers were not combined into the same teams. Students were recruited from a large Midwestern university in the United States using a student sampling service provided by the university. MTurk participants were recruited through the CloudResearch platform [23]. Our tasks on MTurk were open to workers who lived in Canada or the US, had successfully completed at least 1000 tasks on the platform, and had an approval rating of 99% or higher. We used these qualifications to ensure that our crowd workers were fluent in English and would be committed to showing up on time and completing the study. As not all of the participants who initially signed up for the study showed up for the collaborative activity at their scheduled time, teams varied in size between three to six members.

112 participants joined our study for the collaborative activity (part 1 of the study), split into 27 teams of size 3 to 6. 104 participants went on to complete the peer evaluation survey the next day (part 2 of the study). We removed 3 participants from the study who did not follow instructions in the peer evaluation survey correctly. Out of the 101 remaining participants who completed the peer evaluation survey, 48 participants from 13 teams were assigned to the treatment condition and 53 participants from 14 teams were assigned to the control condition. As Bayesian analysis is less sensitive to the sample size, we did not see an issue having differences in sample sizes for each condition. Team sizes were also comparable across conditions (M=4, σ=0.95). Finally, 100 participants completed the feedback survey (part 3 of the study). 48 of these participants were from the treatment condition and 52 were from the control condition.

Participants received $12 as base compensation for completing the entire study, regardless of their performance on the collaborative task. In addition, they were informed at the beginning of the study that they may receive an additional bonus of up to $2 depending on the final outcome quality and how they are evaluated by their team members. In reality, all participants received the full bonus, regardless of the quality of the final outcome or the evaluations they received from their team members. Therefore, all student and MTurk participants were paid $14 in total for completing the entire study. After completing the study, participants were debriefed and sent the full bonus payment. Participation in the study spanned two days and took up to one hour in total.

4.3 Measures

4.3.1 Number, magnitude, and explanation for revisions (RQ1). To investigate how reviewing activity traces impact participants' initial evaluations from memory, we calculated the number of revisions that each participant made to their initial evaluation and the magnitude of those revisions. Given

that the average evaluation scores and maximum number of revisions vary by the size of the team, we control for team size in our analyses.

We supplement the quantitative metrics for revisions with open-ended explanations for why participants made those revisions. This allows us to determine whether reviewing the activity traces were the primary reason that participants in the treatment condition revised their evaluation and how the traces specifically influenced them to revise.

4.3.2 Evaluation consistency (RQ2). We operationalized the consistency of evaluations by using the standard deviation of the evaluation scores received by each participant. We define consistency as the inverse of the standard deviation of evaluation scores: as the scores of evaluators converge to the same value for a given team member, the standard deviation decreases and consistency increases.

4.3.3 Attitudes towards evaluation process (RQ3). After submitting their final evaluation, participants rated their agreement with various attitudes towards the evaluation: confidence in the accuracy of the evaluations they gave to team members (*"I am confident that the evaluations I gave are an accurate representation of each team members' contributions."*), their confidence in the accuracy of the evaluations they will receive from team members (*"I am confident that my other team members will be able to provide an accurate evaluation of my contributions."*), and their perceived difficulty in evaluating their teammates (*"Evaluating each team members' contributions was difficult."*). After they viewed their actual evaluation scores in the feedback survey, participants also rated their perceived accuracy of the evaluations (*"The evaluations that my team members gave me are an accurate representation of my contributions."*) and their satisfaction with the evaluations (*"I am satisfied with the evaluations that I received from my team members."*). All ratings were based on a 7-point Likert agreement scale – 1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, 4: Neither agree nor disagree, 5: Somewhat agree, 6: Agree, 7: Strongly agree. .

4.3.4 Attitudes towards activity traces (RQ4). Participants in the treatment condition were asked open-ended questions about what they perceive to be the potential benefits and concerns of using activity traces during the peer evaluation process in a real-world collaborative context.

4.4 Data Analysis

For our quantitative outcome measures, we used Bayesian analysis to compare the distributions of effects on the number and magnitude of revisions (RQ1), standard deviation of evaluation scores (RQ2), and attitude ratings (RQ3) between experimental conditions. While traditional statistics is a powerful tool in the hands of an experienced statistician, we were motivated to use Bayesian analysis for the following reasons. First, Bayesian models make transparent all assumptions in the model. The modeler does not need to cross-validate distributional assumptions with the data. Second, Bayesian analysis generally uses maximum entropy priors, which allow us to make the most conservative inferences given the evidence. We can encode skepticism towards larger effect sizes in small samples by using more conservative priors, making Bayesian analysis suitable for small-n studies. In our models, we use weakly informative priors to ensure that the priors do not dominate inference. Second, Finally, Bayesian models better facilitate the accrual of knowledge within the research community as prior outcomes can be used as informative priors in future research [29]. Kay et al. provide a more detailed overview of the advantages of Bayesian inference for statistical analysis of HCI research [20].

We formulate a hierarchical Bayesian model for each quantitative outcome measure. Full mathematical descriptions of each model are provided in the Appendix. Our data is naturally hierarchical as we analyze evaluations across multiple clusters, such as teams and population. Hierarchical

Bayesian models enable partial pooling, where information is shared across clusters to improve estimation. For RQ1, we define separate models for the number of revisions and the magnitude of the revisions. In all of the models, we control for the following covariates: team size, population (student vs. MTurk), and the participant's prior experience with collaboration and peer evaluations in teams. Figure 13 displays the full Directed Acyclic Graphs (DAGs) that we created to represent our beliefs about the causal relationships between the variables in our models. We do not derive the DAGs from the data. These DAGs are manifest in the Bayesian models that we ran in our analyses. We performed the Bayesian analysis using NumPyro [5, 39], a popular Bayesian inference framework. We used Markov Chain Monte Carlo (MCMC), a stochastic sampling technique to sample the posterior distribution $P(\theta|D)$, the distribution functions of the parameters in the likelihood function given the data observations D.

Finally, we performed thematic analysis [16] on all open-ended responses in the surveys. 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 to extract common themes for each survey question.

5 RESULTS

In this section, we present our quantitative findings for how reviewing activity logs affected participants' evaluations in terms of the revisions that they made (RQ1), the consistency within teams (RQ2), and their attitudes towards the evaluations (RQ3). When possible, we also report open-ended comments to contextualize the quantitative findings. Finally, we present a qualitative analysis of participants' perceived benefits and concerns on the usage of activity traces in the peer evaluation process (RQ4). For all of the Bayesian models described, the Gelman-Rubin statistic (a measure of MCMC convergence) for all parameters was around 1, indicating that the multiple sampling chains converged. Traceplots for the MCMC chains in all of the models are presented in the Appendix.

5.1 Effect on Revisions (RQ1)

About 54% of participants in the treatment condition revised their initial evaluations at least once, compared to 19% of participants in the control condition. Overall, we found that reviewing activity traces had a significant effect on both the number and magnitude of revisions that participants made to their initial evaluations. Below, we describe further details of our analysis and findings.

5.1.1 Number of Revisions. In our Bayesian analysis, we modeled the number of revisions as an ordinal variable and estimated the posterior distributions of the cumulative likelihood of each level of revision. Further details of this model can be found in Section 8.1.2 of the Appendix. The key outcome of interest is whether the odds of no revision is significantly lower in the treatment condition. To examine the effect of the treatment, we constructed the distribution of the difference between the cumulative odds of making zero revisions in the treatment condition and the cumulative odds of making zero revisions in the control condition, as shown in Figure 3. All plots are shown on the cumulative odds scale, though for zero revisions, the cumulative odds are simply equivalent to the odds of making zero revisions. If the odds of making zero revisions are significantly lower in the treatment condition (i.e. the difference is negative), this is equivalent to the odds of any revisions being greater in the treatment condition than in the control condition.

Our findings show that reviewing activity traces significantly decreased the odds of making no revisions to the initial evaluations. In other words, the odds of any revisions is greater in the treatment condition than in the control condition. The first column shows the cumulative odds difference when averaged over both the MTurk and Student populations. The 94% High-Posterior

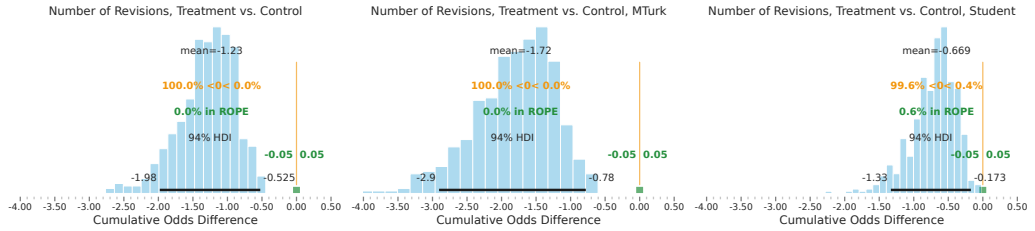


Fig. 3. Difference distributions between the control condition and the treatment condition for the odds of making no revisions to the initial evaluation. Lower odds of making no revisions is equivalent to increased odds of making one or more revisions. All plots are on the cumulative odds scale. The leftmost plot shows the distribution averaged over both MTurk and student participants. The center and right plots show the distributions for only MTurk and only student participants, respectively. Each plot shows an orange vertical line located at 0. This represents an odds difference of zero, meaning that participants in the treatment condition were not significantly more or less likely to make zero revisions compared to the control condition. **Main finding:** Our findings show that reviewing activity traces significantly decreased the odds of making no revisions to the initial evaluations compared to the control condition. In other words, the odds of making any revisions is greater in the treatment condition than in the control condition. This effect was primarily driven by MTurk participants.

Density Interval (HPDI) is [-1.98, -0.53], which lies outside a significant ROPE (Region of Practical Equivalence) of 0±0.05, where 0 indicates that the treatment had no effect. [1] Surprisingly, when we examine the two populations separately, we see that these differences primarily arise from MTurk participants. The 94% HPDI for the MTurk distribution is [-2.9, -0.78]. In contrast, for Student participants, the 94% HPDI is [-1.33, -0.17]. Although the HPDI still lies outside the ROPE, this indicates that the difference for students is only marginally significant.

We can better understand the effects of the treatment by visualizing the predicted distribution of outcomes for each condition under our model. These distributions are shown in the histograms in Figure 4. The first column shows the predicted distribution of number of revisions for MTurk participants across both conditions. The right column shows the predicted distribution of number of revisions for student participants across both conditions. The y-axis represents the number of samples in which each outcome value was predicted and is proportional to the probability of that outcome value. We can see that the treatment condition decreased the probability of making no revision in the simulated outcomes and increased the probability of making each successive number of revisions. The contrast was greatest for MTurk participants where the probability of making no revisions in the treatment condition was only about half the probability of making no revisions in the control condition.

5.1.2 Magnitude of Revisions. We estimated the posterior distribution of the mean magnitude of revisions that participants made to their initial evaluations. Further details of this model can be found in Section 8.1.1 of the Appendix. To contrast the mean magnitude of revisions between the control condition and the treatment condition, we constructed the distribution of the ratio of the

[1]Unlike non-Bayesian Statistics, where one can ask for example, if the two means for two treatments are different $P(\mu_1 \neq \mu_2)$, in Bayesian statistics, one asks if the HPDI interval of the distribution $P(\mu_1 - \mu_2)$, that is, the distribution of the difference of the means of the two treatments, excludes an interval where we can consider the two treatments equivalent. This equivalence interval is domain dependent. A posterior distribution HPDI that lies outside the ROPE is considered a significant result in Bayesian data analysis.

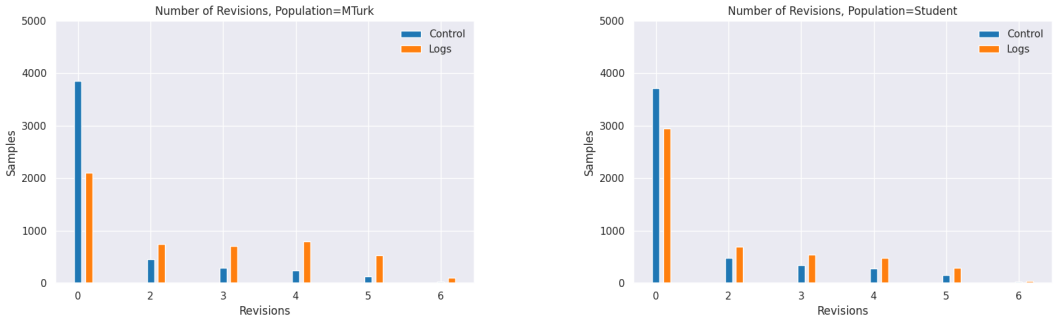


Fig. 4. This figure shows the predicted outcomes for the number of revisions that participants made to their initial peer evaluation over 5000 samples. Each plot shows how the distribution of predicted revisions varies by condition and population. **Main finding:** The treatment condition decreased the probability of making no revision in the simulated outcomes and increased the probability of making each successive number of revisions. The contrast was greatest for MTurk participants where the probability of making no revisions in the treatment condition was only about half the probability of making no revisions in the control condition.

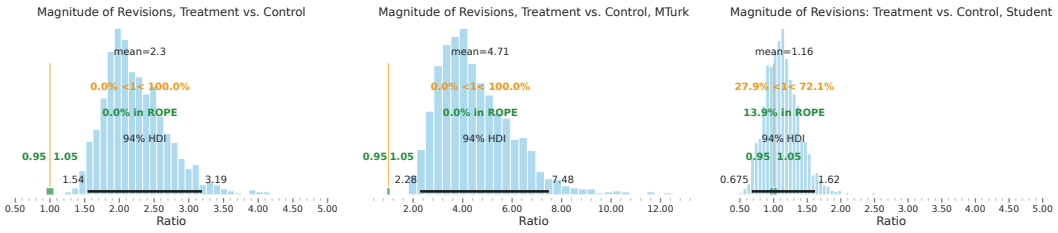


Fig. 5. This figure shows the ratio distributions of the treatment effects on the magnitude of revisions between the control condition and the logs condition. All plots are on the outcome scale. The rightmost plot shows the ratio distribution averaged over both MTurk and student participants. The middle and leftmost plots show the ratio distributions for only MTurk and only student participants, respectively. Each plot shows an orange vertical line located at 1. This represents that the magnitude of revisions was the same between conditions. **Main finding:** Reviewing activity traces significantly increased the magnitude of revisions that participants made to their initial evaluations. However, this difference was only significant for MTurk participants.

treatment effect in the treatment condition over the control condition, as shown in Figure 5. All plots are on the outcome scale.

Our analysis shows that viewing activity traces significantly increased the magnitude of revisions that participants made to their initial evaluations. The first column shows the ratio when averaged over both the MTurk and Student populations. The 94% HPDI is [1.54, 3.19], which lies outside the ROPE of 1±0.05, where 1 indicates that the treatment had no effect. This indicates there was a significant difference between the magnitude of revisions of the two conditions, where the treatment effect increased the magnitude of revisions. On average, the magnitude of revisions in the treatment condition is approximately 2.3 times the magnitude of revisions in the control condition. Similar to our findings for the number of revisions, we see that this difference again primarily comes from MTurk participants. The second column of the figure shows that the treatment effect had a stronger effect on MTurk participants. The average ratio is 4.71 and the 94% HPDI widens to [2.28, 7.48]. For

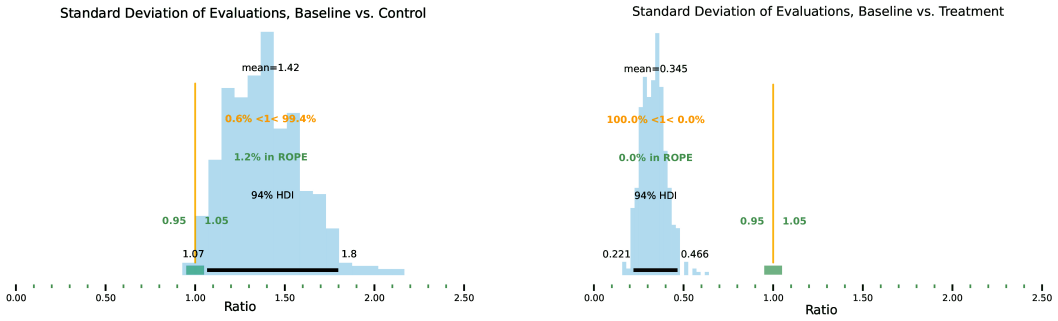


Fig. 6. This figure shows the ratio distributions of the treatment effects on the standard deviation of evaluation scores between the baseline condition (the initial evaluation) and each of the reflection conditions (control and treatment, respectively). All plots are on the outcome scale. The lefthand plot shows the control over baseline distribution and the righthand plot shows the treatment over baseline distribution. Each plot shows an orange vertical line located at 1. This represents that the standard deviation of evaluation scores was the same between conditions. A ratio lower than 1 indicates that standard deviation of scores increased after the reflection stage, i.e. consistency decreased. **Main finding:** Reviewing activity traces led to a significant decrease in the standard deviation of evaluation scores that each team member received.

Student participants, the average ratio is 1.16 and the 94% High-Posterior Density Interval (HPDI) shrinks to [0.68, 1.62], which overlaps with the ROPE of 1±0.05. About 13.9% of the entire posterior distribution for students is contained within the ROPE. This suggests that the treatment did not have a significant effect on magnitude of revisions for students.

5.1.3 Explanations for Revisions. From the open-ended explanations that participants provided for why they made revisions, we found that participants in the treatment condition who revised their evaluations frequently mentioned that the activity traces helped to scaffold their memory and clarify contributions from their team members (N=26). Participants often realized that they had underestimated (N=11) or overestimated (N=9) the contributions of their team members after reviewing the activity traces. One participant acknowledged experiencing a degree of memory bias: *"I recall liking certain peoples ideas more than others. I think this skewed my memory in remembering their work more than the others"* (P35, Student, Treatment). For participants in the control condition who revised their evaluations, the written reflection by itself led them to realize they had underestimated (N=5) or overestimated (N=1) other people's contributions. At the same time, three participants in the control condition stated that they only changed their evaluation ratings because they realized that their team members would see the feedback: *"I realized they were going to see what numbers I had given them, so I figured I might as well pretend to be nice"* " (P85, MTurk, Control).

5.2 Effect on Consistency (RQ2)

To examine how reviewing activity traces affected the consistency of evaluations, we modeled the evaluation scores that participants received and compared the standard deviations of evaluation scores. Further details of this model can be found in Section 8.1.3 of the Appendix. Higher standard deviation indicates that teams disagreed more about how much each member contributed. To contrast the changes in standard deviation between the control condition and the treatment condition, we first constructed the distribution of the ratio of standard deviation of the revised

evaluation in the treatment condition over the standard deviation of the initial baseline evaluation. This informs us about how much the standard deviation changed from the initial to revised evaluation in the treatment condition. We constructed the same ratio distribution for the control condition in order to examine whether the reflection activity on its own affected the consistency of evaluations. These distributions are shown in Figure 6. All plots are on the outcome scale.

We found that reviewing activity traces led to a significant decrease in the standard deviation of evaluation scores that each team member received. The right column shows that the 94% HPDI is [0.22, 0.47] for the ratio between the baseline evaluation and the revised evaluation in the treatment condition. This lies outside the ROPE of 1±0.05, demonstrating that there was a clear treatment effect on consistency. On average, the standard deviation of evaluation scores received by each participant was only 38% of the standard deviation of evaluation scores in the initial evaluation, decreasing by 62%. In contrast, the HPDI for the ratio in the control condition is [1.07, 1.8], which suggests that standard deviation actually increased from the initial evaluation. Although the HPDI does not overlap with the ROPE, this difference is only marginally significant. Overall, these results show that reviewing activity traces significantly increased the consistency of evaluation scores, but only in the treatment condition.

This increased consistency is supported by open-ended comments from participants in the treatment condition. Participants perceived activity traces as being a sort of "ground truth" about people's contributions. As one participant stated, *"With activity logs, we can be sure that at least there is one correct version of 'what has been going on'"* (P33, Student, Treatment).

5.3 Effect on Attitudes (RQ3)

We assessed participants' attitudes towards their peer evaluations both directly after they completed the evaluations as well as after they viewed the evaluations they received from their team in the final feedback survey. Distributions for the attitudes collected directly after the evaluation survey are shown in Figure 7 and distributions for the attitudes collected after the final feedback survey are shown in Figure 8. All attitude questions were rated on a 7-point Likert agreement scale. Median and standard deviation are reported using the notation M and σ, respectively. Overall, our analysis showed that reviewing activity traces significantly increased participants' perceived accuracy of the final evaluations they received. However, this effect varied between populations. All other attitudes were not significantly different between conditions. We first report descriptive findings about the attitude responses before delving into the results of our Bayesian analysis.

After revising their evaluations, participants in both conditions were confident that the evaluations they gave were accurate (M=6, σ=1.14). They were similarly confident that the evaluations they received would be accurate (M=6, σ=1.19). In general, participants did not find it difficult to evaluate their team members' contributions (M=4, σ=1.76). The median for perceived difficulty was actually higher for participants in the control condition (M=4, σ=1.89) compared to the treatment condition (M=3, σ=1.61). In the final feedback survey, participants viewed the evaluation scores they received from their team members and rated their perceived accuracy and satisfaction with the evaluations. Distributions for both attitude questions are shown in Figure 8. Overall, participants in both conditions were somewhat satisfied with the evaluations they received (M=5, σ=1.68). However, participants felt mixed about the accuracy of the evaluations (M=4, σ=1.67). Of note, participants in the treatment condition perceived the evaluations they received as being more accurate (M=5, σ=1.58) compared to the control condition (M=4, σ=1.69).

To examine whether these differences were statistically significant, we performed a similar analysis to our analysis for the number of revisions. We modeled the attitude ratings as an ordinal variable and estimated the posterior distributions of the cumulative likelihoods of the response values for each attitude rating scale. Further details of this model can be found in Section 8.1.4 of

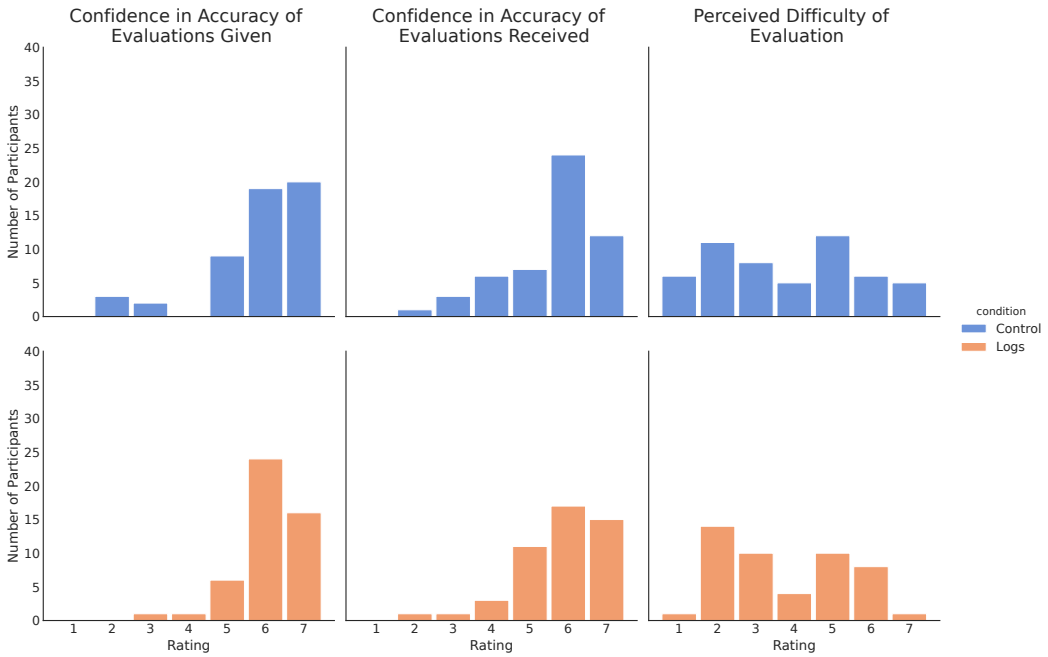


Fig. 7. Distribution of ratings for confidence in accuracy of the evaluations that participants gave, confidence in accuracy of the evaluations that participants would receive, and perceived difficulty of evaluation. Ratings were based on a 7-point Likert agreement scale – 1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, 4: Neither agree nor disagree, 5: Somewhat agree, 6: Agree, 7: Strongly agree. Survey takers on average felt confident that the evaluations they gave were accurate and that their teammates would also evaluate them accurately. Perceptions of difficulty were more mixed.

the Appendix. Our main question in the Bayesian analysis was whether the distribution of ratings was significantly different between conditions with respect to the neutral midpoint of the scale (*"Neither agree nor disagree"*). This would tell us whether participants were more likely to disagree with the attitude statement in one condition over the other. We constructed the distribution of the difference between the cumulative odds of a rating of 4 (the midpoint of the 7-point Likert scale) in the treatment condition and the cumulative odds of the rating in the control condition. Values that are negative and exclude the ROPE of 0±0.05 indicate that participants in the treatment condition had less odds of providing a neutral or negative response to the attitude question compared to the control condition.

We found that participants' attitudes directly after completing the evaluations (confidence in the accuracy of evaluations given and received and perceived difficulty of evaluation) did not differ significantly between the control condition and treatment condition. Figure 9 shows the contrast distributions for the cumulative odds difference for each of the three attitudes. Each row represents a different attitude rating and each column represents the population selected for analysis. The HPDI for the cumulative odds difference overlapped with the ROPE of 0±0.05 in all cases, indicating that participants were not significantly more likely to disagree or agree in one condition over the other for these attitudes. Figure 10 shows the contrast distributions for the attitudes of participants once they received their scores in the final feedback survey (perceived accuracy and satisfaction

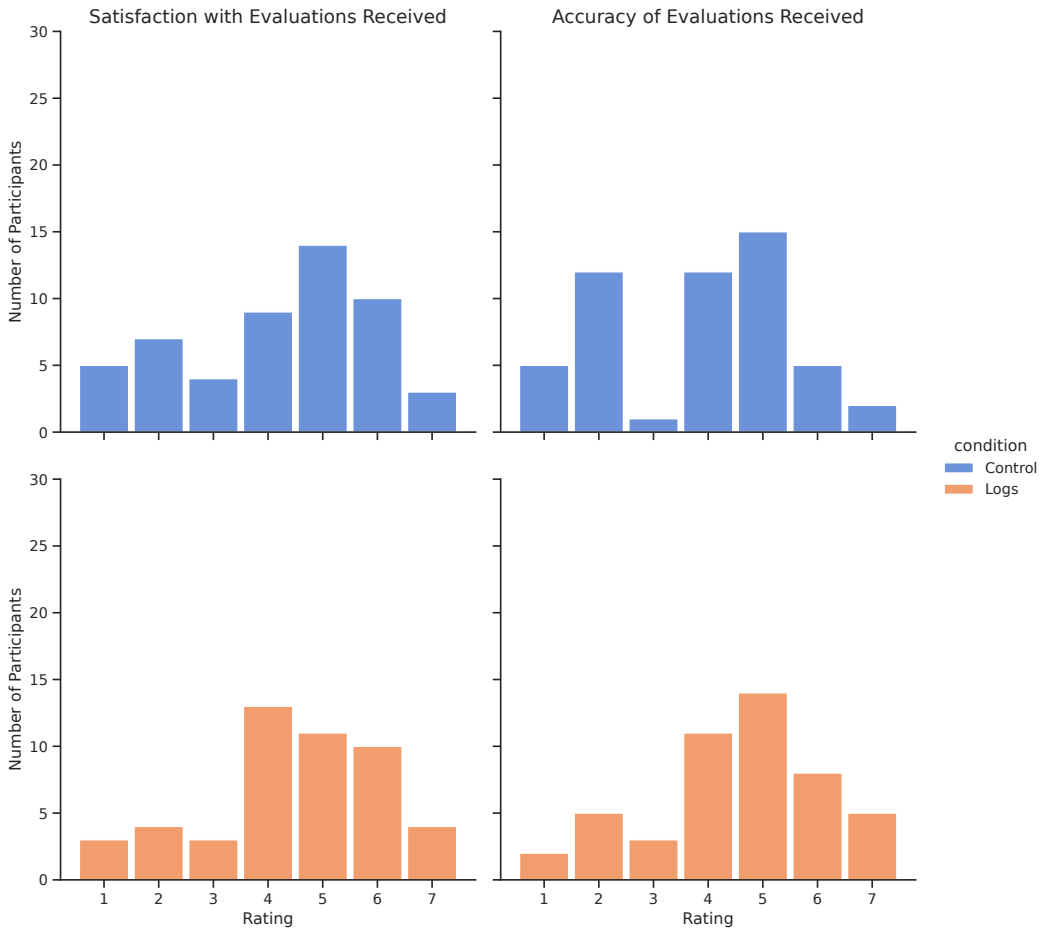


Fig. 8. Distribution of ratings for perceived accuracy and satisfaction with the evaluations participants received in the final feedback survey. Ratings were based on a 7-point Likert agreement scale – 1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, 4: Neither agree nor disagree, 5: Somewhat agree, 6: Agree, 7: Strongly agree. Participants were somewhat satisfied with the evaluations they received in both conditions. However, in the treatment condition, participants were more likely to perceive their evaluations as being accurate. .

with evaluations received). We did not find a significant difference in responses for participants' perceived satisfaction with the evaluations, as shown in the first row of Figure 10.

However, participants' perceived accuracy of the evaluations they received demonstrated a significant difference between conditions. The 94% High-Posterior Density Interval (HPDI) is [-1.07, -0.26], which lies outside a significant ROPE of 0±0.05. This indicates that participants in the treatment condition were significantly more likely to agree with the statement that the evaluations they received accurately represented their contributions. Surprisingly, this increase in perceived accuracy is primarily driven by student participants. The 94% HPDI for the student distribution is [-1.78, -0.37]. In contrast, for MTurk participants, 94% HPDI is [-0.67, 0.0], which overlaps with the ROPE and is not significant. Therefore, although reviewing activity traces did not significantly

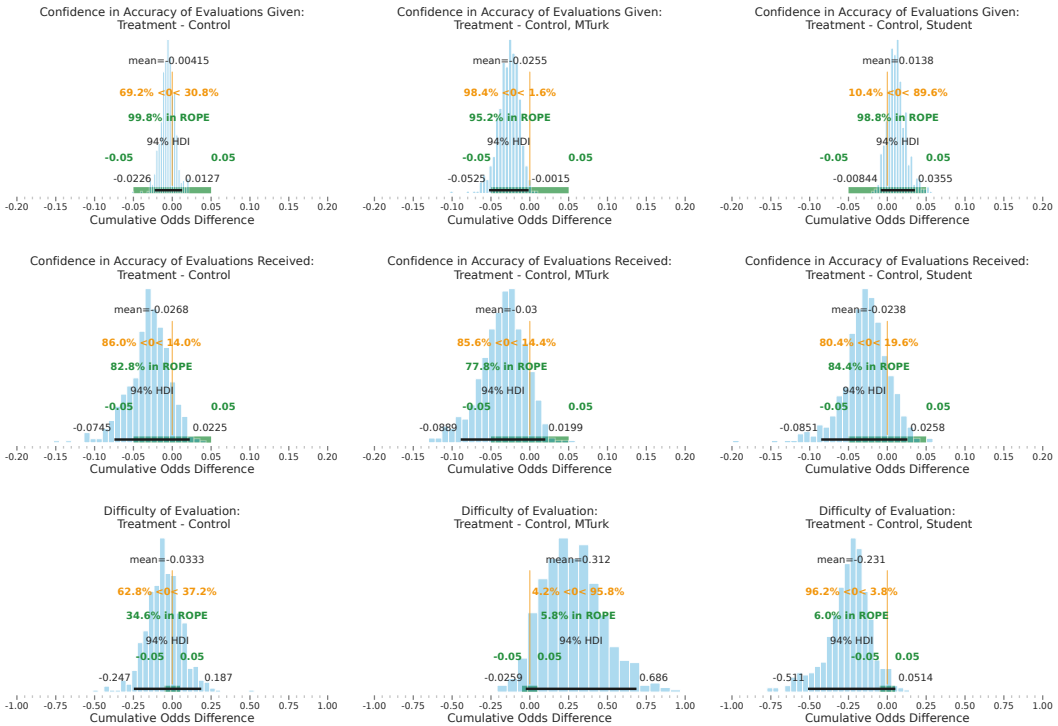


Fig. 9. Difference distributions between conditions for the cumulative odds of observing a neutral or lower rating for confidence in accuracy of evaluations given, confidence in accuracy of evaluations received, and difficulty of evaluation. All plots are on the cumulative odds scale. Each row represents a different attitude rating. The first column shows the distributions averaged over both MTurk and student participants. The second and third columns show the distributions for only MTurk and only student participants, respectively. Each plot shows an orange vertical line located at 1. This represents that there was no difference in responses between conditions. Note that the x-axis is not the same scale for all plots. **Main finding**: Reviewing activity traces did not have a significant effect on participants' confidence in the accuracy of their evaluations or perceived difficulty of evaluation.

impact the number or magnitude of revisions students made, it did increase the accuracy they perceived for their final evaluation scores.

We can again better understand the effects of the treatment on each attitude by visualizing the predicted outcomes for each condition under our model. The predicted outcomes for the evaluation survey attitudes are shown in Figure 11. The predicted outcomes for the feedback survey attitudes are shown in Figure 12. In both figures, the first column shows the predicted ratings for MTurk participants across both conditions. The right column shows the predicted ratings for student participants across both conditions. We can see that the distribution of predicted outcomes is very similar for most of the attitudes. When we examine perceived accuracy of evaluations received, shown in the first row of Figure 12, we find larger differences in the probability of each response value for students. The probability of responses on the negative end of the scale all decreased for the treatment condition while the probability of responses on the positive end of the scale all increased.

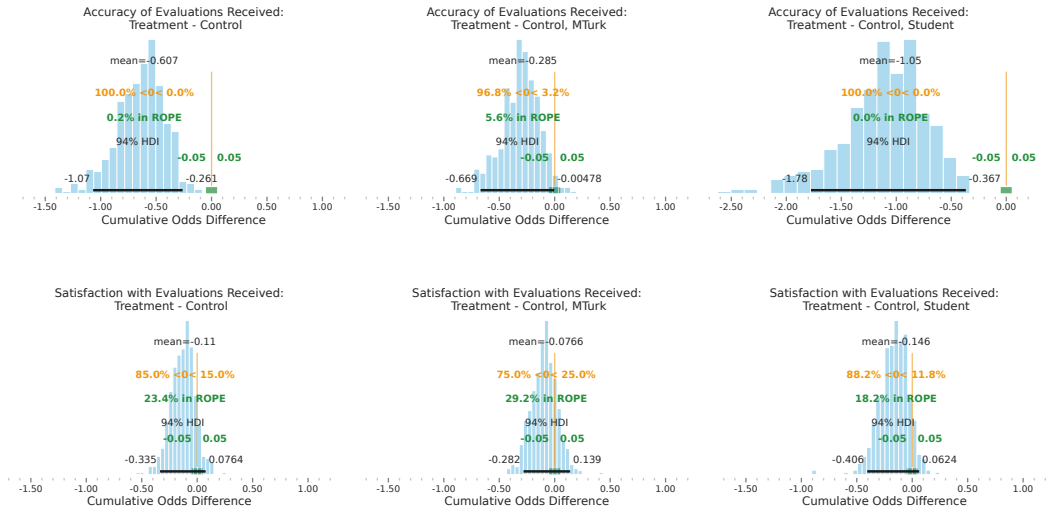


Fig. 10. Difference distributions between conditions for the cumulative odds of observing a neutral or lower rating for perceived accuracy of and satisfaction with evaluations received. All plots are on the cumulative odds scale. Each row represents a different attitude rating. The first column shows the distributions averaged over both MTurk and student participants. The second and third columns show the distributions for only MTurk and only student participants, respectively. Each plot shows an orange vertical line located at 1. This represents that there was no difference in responses between conditions. Note that the x-axis differs in scale for some plots. **Main finding:** Reviewing activity traces had a significant effect on participants' perceived accuracy of the evaluations they received, but only for student participants.

5.4 Benefits and Concerns (RQ4)

Participants in the treatment condition described multiple potential sources of value in using activity traces during peer evaluations in collaborative scenarios. Most predominantly, participants reported that activity traces can facilitate a better understanding of the individual contributions of everyone on the team (N=25). Participants discussed this benefit in the context of both short-term and long-term collaborations. For short-term, synchronous collaborations such as the ad-writing task in the study, participants stated that it can be difficult to keep track of what is happening and who is doing what during the task. For long-term collaborations, memory of events becomes further strained, making it easy to forget or misremember contributions that were made.

In both cases, activity traces were described as being helpful in refreshing or scaffolding participants' memory (N=9). P29, a student experienced this issue firsthand with the peer evaluation in the study: *"To put it bluntly, had I not reviewed the activity logs I would have absolutely screwed some people over. If I was someone's manager doing performance reviews, without the proof of how some people worked, I would've simply mixed up some people's work and underestimate or overestimate their total contribution."* Participants valued activity traces for being easy to refer back to at a later date (N=10) and as a verification tool to confirm their evaluations or correct possible errors (N=6). Participants (N=7) often attributed a sense of objectivity to the activity traces that allows them to *"rely on more than just their memories"* (P35, Student). P72 (MTurk) felt that this could help reduce perceptions of favoritism or bias in the evaluations. Two participants stated that that they could draw tangible examples from the activity traces as evidence for their evaluations.

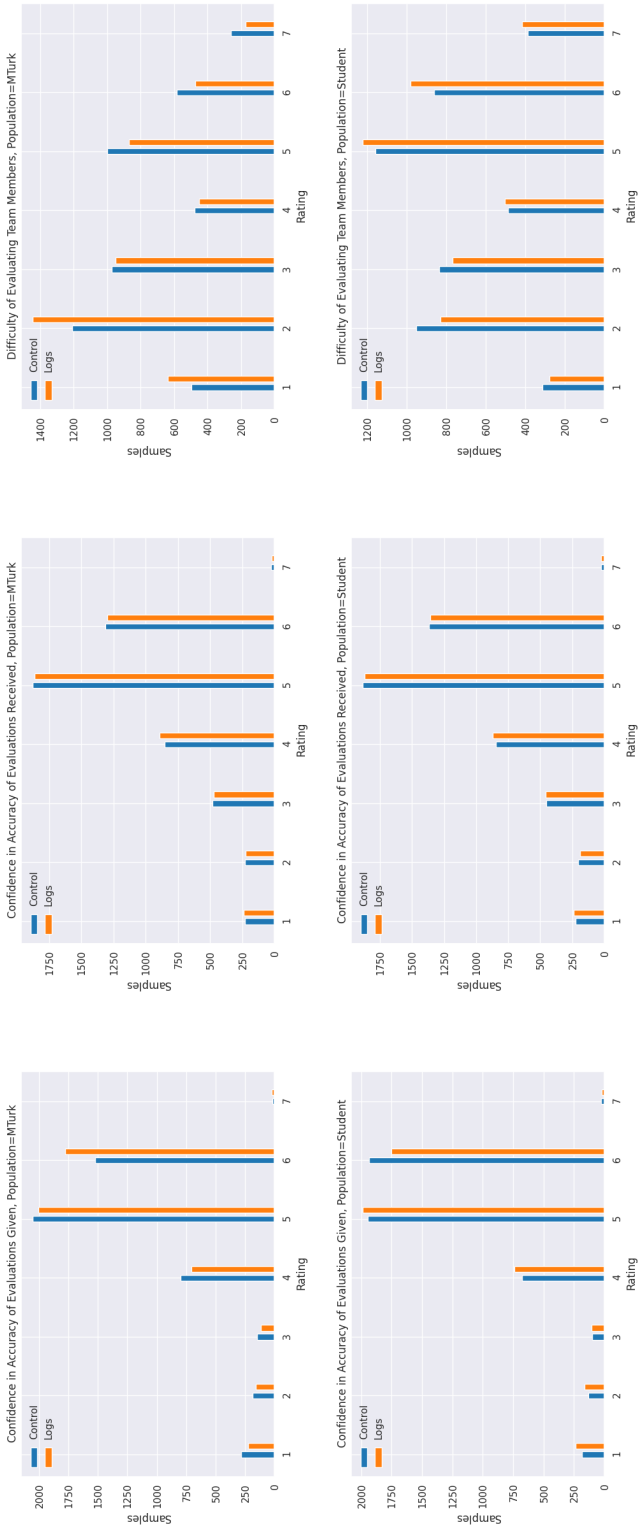


Fig. 11. Predicted outcomes for the attitude questions in the peer evaluation survey. Each plot shows how the distribution of predicted responses varies by condition and population. Each column represents a different attitude question. **Main finding:** Distributions of outcomes for the attitudes from the peer evaluation survey were similar across conditions.

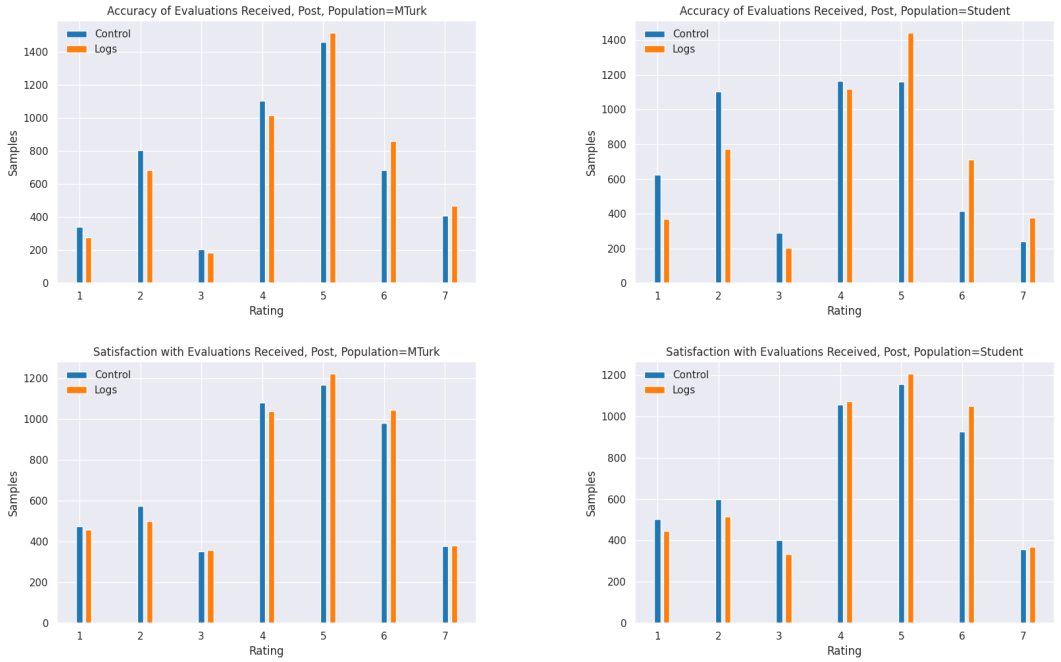


Fig. 12. Predicted outcomes for the attitude questions in the feedback survey. Each plot shows how the distribution of predicted responses varies by condition and population. Each column represents a different attitude question. **Main finding:** For students, reviewing activity traces decreased the probability of negative ratings for perceived accuracy and increased the probability of positive ratings.

At the same time, participants expressed concerns about lack of context, privacy, bias, and performativeness among other issues. The most frequent concern for participants was that activity traces do not provide a complete picture of the team and each members' contributions (N=18). This concern largely stemmed from participants feeling that activity traces do not necessarily convey the context of the information shown, such as the cognitive effort a team member is putting in: *"In exercises like these, there is both "visible" activity and "thought" activity and the logs tend to capture more of the "visible" activity (i.e., editing, comments, etc) but they don't really capture the "thought" activity - and it is that ideation that also leads to success and a better final deliverable - so there should be some way to capture and/or recognize that as well."* (P72, MTurk) Furthermore, several participants (N=7) also worried that activity traces might be biased towards certain types of roles or personalities on the team, such as those who are more extroverted or assertive. Participants (N=4) also cited other types of contributions such as offline interactions that might not be captured by activity traces.

Many participants (N=8) also had reservations about the privacy and security of the data being recorded in activity traces. These reservations mostly revolved around the type and granularity of information shown in the activity traces and how long they could be retained and viewed. P30 (Student) speculated that people might "dig too deep" into the activity traces and penalize their team members for past actions that do not reflect their current performance. Participants (N=8) were concerned that the usage of activity traces could aggravate social tensions within the team, making team members feel stressed about being monitored or resentful of their teammates. Finally,

several participants (N=6) felt that the usage of activity traces would encourage performative work. P68 (MTurk) stated, *"You are working hard only if other people reviewing the logs see you working hard."* In turn, this could lead to an overemphasis on quantity over quality of contributions.

6 DISCUSSION

In this section, we summarize and discuss our experimental findings in the context of the larger body of work on peer evaluations and collaborative awareness. We also address stakeholders that use peer evaluations and tool designers of peer evaluation systems by providing implications for leveraging activity traces effectively in the evaluation process. We integrate these implications with the potential benefits and concerns that participants perceived for using activity traces in the peer evaluation process.

6.1 Experimental Findings

In this study, we examined the effects of reviewing activity traces on peer evaluations along three primary dimensions: the number and magnitude of revisions to evaluations, the consistency of evaluations, and participants' attitudes towards the peer evaluation process. Our findings show that using activity traces can address social and cognitive biases with peer evaluations and increase the consistency among evaluators, without increasing the perceived cognitive workload.

6.1.1 Activity Traces Led to More and Greater Revisions. Our findings show that participants made significantly more revisions to their initial evaluations after reviewing their activity traces and that the size of these revisions was also significantly greater. Participants who reviewed their activity traces showed significantly greater odds of making a revision. The size of revisions made by participants in the treatment condition was on average more than twice the size of revisions made in the control condition. These revisions were contextualized by the open-ended explanations with many participants stating that the activity traces helped to clarify and correct their initial estimations of team members' contributions. This suggests that activity traces can address issues with limited or biased recall. While this study took place in the context of a short-term collaborative task, the benefits of activity traces for scaffolding memory become more relevant for peer evaluations that are performed between extensive periods of work, such as for annual reviews or at the end of semester-long projects. In these situations, memory biases are likely to be most problematic.

We note that the effect on revisions was stronger and, in the case of magnitude of revisions, only significant for MTurk participants compared to student participants. We consider two possible explanations for this. First, prior research has shown that MTurkers are more financially motivated compared to other recruiting methods [35, 36]. Thus, they may have been more likely to provide inaccurate evaluations initially in order to achieve the maximum possible bonus. The activity traces may then have had a stronger effect on their revised evaluations. Second, although we did not find significant differences between MTurk or student participants for either collaborative or peer evaluation experience, it is possible that students may have had more recent or frequent experience with evaluating peers through group projects and assignments in courses. Further investigation is needed to unpack these differences.

6.1.2 Activity Traces Led to More Consistent Evaluations. Teams in the treatment condition became significantly more aligned in how they evaluated each team member after revising their evaluations. The standard deviation of revised evaluation scores in the treatment condition was on average 38% of the standard deviation in the baseline evaluation. In contrast, performing the reflection activity on its own in the control condition did not improve the consistency of evaluations. This finding provides key evidence that reviewing activity traces improved the quality of peer evaluations. In the absence of a ground-truth, consistency between feedback sources (e.g. members of a team) is

one of the most important benchmarks for the quality of an evaluation [9]. An explanation for the increased consistency is that team members were able to use activity traces as a common reference point. This is supported by open-ended comments where participants perceived activity traces as being a sort of "ground truth" about people's contributions. As one participant stated, *"With activity logs, we can be sure that at least there is one correct version of 'what has been going on'"* (P33, Student). We note that this is only true when the collaboration involves artifact-based work, as was the case in our study where teams had to create product advertisements.

6.1.3 Perceptions of the Evaluation Process. Our findings on participants' perceptions of the evaluation process show that the treatment condition did not have a significant effect on attitudes measured directly after participants completed their self- and peer evaluations. We had speculated that participants may feel more confident in the accuracy of the evaluations when they were able to review their activity traces. However, we found that reviewing activity traces did not have a significant effect on participants' confidence in the accuracy of either the evaluations they gave to other members or the evaluations they would receive from their team members. One possible explanation for this is that there was a ceiling effect for perceived accuracy. Ratings were high for both accuracy of evaluations given and accuracy of evaluations received, with a median of 6 on a 7-point Likert scale for both questions. This could also suggest that participants in the control condition were over-confident in the accuracy of their evaluations. The comparative lack of revisions to initial evaluations in the control condition support this interpretation as well.

However, differences emerged between conditions once participants actually viewed their evaluation scores in the final feedback survey. Participants who reviewed their activity traces were more likely to perceive their evaluations as accurate representations of their contributions. Combined with our findings on increased consistency of evaluations, it is possible that team members who reviewed activity traces were able to converge on a common frame of reference against which they based their judgements of accuracy. When asked to elaborate on their accuracy rating, two participants (one MTurk participant and one student participant) in the treatment condition stated explicitly that the activity traces supported their estimations of accuracy. At the same time, these differences were only statistically significant for student participants. Furthermore, when given an opportunity to appeal the evaluation they received in the final survey, 9 out of 15 participants who suggested they would like to appeal were students, with 8 of them being from the control condition. Overall, these findings demonstrate that the presence of activity traces influence how team members evaluate each other as well as how they assess the quality of their evaluations. This is critical as perceptions of peer evaluation process can significantly affect their motivation to perform the evaluations.

Finally, we found that participants did not find it significantly more or less difficult to evaluate their team members between conditions, despite participants in the treatment condition needing to perform the extra task of reviewing the activity traces. This suggests that the cognitive effort of reviewing the activity traces may have somewhat mediated the cognitive effort of recalling each member's contributions without the data. This is promising as prior work has cautioned that greater activity awareness may increase cognitive load and stress for users [47].

6.2 When to Use Activity Traces?

Our findings demonstrate the value for leveraging activity traces in peer evaluations in the context of a specific online collaborative and evaluation task. The characteristics of the collaboration and evaluation will likely impact the applicability and value of activity traces. Most notably, activity traces are limited to capturing online contributions. In addition, activity traces are generally tied to artifacts of work. These artifacts can encompass many aspects of teamwork such as communication,

effort, and accountability. But collaborative work that primarily takes place in-person or involves more abstract outcomes will most likely not be appropriate contexts for using activity traces. Nevertheless, given the rapid advancements in online collaboration over the past few decades, we believe that the importance of activity traces will only increase in the future.

The most significant benefit that participants in our study perceived for using activity traces in the peer evaluation process was in helping to understand the individual contributions of team members. Therefore, activity traces are likely best-suited for evaluation contexts where it is important to differentiate the individual and relative contributions of a group, as was the case in the evaluation in our study. These types of evaluations are necessary when instructors or managers must make administrative decisions such as assigning individual grades or bonuses [12]. However, we believe that there is potential for leveraging activity traces to support other evaluation purposes as well. For example, activity traces can provide tangible examples for participants' evaluations. Team members can compose these examples in their peer evaluations to generate more specific and actionable feedback. Providing evidence in support of their evaluation scores could also make team members feel less apprehension about giving constructive feedback.

Although activity traces can be used to address social and cognitive biases in peer evaluations, they also introduce new concerns about privacy, surveillance, and performativeness. Greater transparency makes everyone's behavior more visible and therefore more open to judgement and criticism. Research on "evaluation apprehension" suggests that when people are worried about others' evaluations of their work, they sometimes make mistakes and learn less than when they are not watched [50]. This was echoed by participants in our study who expressed concern about the social and psychological effects of feeling monitored. Participants also worried that people may purposefully exhibit behaviors with the intention of being reflected more positively in the activity traces, rather than actually contributing to the team. Therefore, stakeholders should carefully consider the costs and benefits of using activity traces for their unique purposes and contexts. In addition, we caution against overly frequent usage of activity traces. Instead, activity traces could be used to form data-driven peer evaluations only at strategic milestones.

6.3 Design Implications for Peer Evaluation Systems

In our study, we used existing activity traces automatically logged from a popular online collaborative tool and provided minimal guidance on how participants should interpret the traces. Although our findings demonstrate that there is already value in the activity traces available today, we also identified challenges and concerns that peer evaluation systems seeking to leverage activity traces should address. Presently, activity traces are automatically generated by online tools with little control or customization provided to the user. As our open-ended responses indicate, people care about how the activity traces might be interpreted and actuated by their peers, especially when these interpretations are being used to evaluate their contributions. Peer evaluation systems that intend to leverage activity traces should be designed to provide greater control and flexibility to users over how their contributions are represented in activity traces and scaffold the reflection process to identify quality contributions.

The representation of activity traces in peer evaluation systems should be configured based on the purpose of the evaluation, goals and preferences of the stakeholders, and the nature of the teamwork. For example, if the goal of evaluation is to focus on task behaviors, activity traces can be organized by tasks. If the goal is to differentiate individual contributions, then activity traces can be organized by person. Similarly, the type of content shown in the activity traces can be configured based on the dimensions of teamwork being evaluated. Peer evaluation systems should also provide opportunities for teams and individuals to contextualize the activity traces with information that cannot be automatically captured, such as the *"thought activity"* that one participant described

or significant offline interactions. This could be performed in-situ as team members work on a task or as part of a reflection activity where teams reflect on and annotate their activity traces. For example, perhaps a team member wrote the least text in a document but it was the most difficult or important section to write. A tool could allow them to annotate the contribution in the activity trace summary. Providing features to assess the value or impact of contributions in activity traces is a critical direction for the design of future tools.

Participants in our study worried that incorporating activity traces into evaluations may increase the occurrence of "performative work". Performativeness at work is not a new concept but it has become increasingly notable with the shift to online work in recent years [2]. For example, people might contribute more to public messaging channels that they know will be displayed in activity traces. Strategies should be developed to identify counterproductive behaviors and differentiate "quality work" from "performative work". During the evaluation process, team members can be instructed to identify these perceived behaviors in the activity traces through rubrics or other forms of messaging. We can draw inspiration from behaviorally anchored rating scales, which identify specific behaviors that represent varying levels of performance and use them to provide a frame of reference for evaluators [45]. Activity traces also serve as a repository from which to gather new distinctive online teamwork behaviors that can be used to guide team members in their evaluations.

Finally, our findings show that participants valued activity traces for scaffolding their memory and saw potential for using them in long-term collaborations. However, peer evaluation systems will inevitably encounter issues of scale when dealing with collaboration that occurs over extended periods of time. Real-world collaborations ebb and flow over time, with people's individual contributions varying based on their expertise, the tasks at hand, or spontaneous events. A team member's contributions may not be visible in one type of activity trace because their role requires them to contribute to a different outcome or stage of the project. The distributed nature of online collaboration also means that many different platforms and tools are used for a single project, further compounding these issues. A student team working together on a semester-long project could produce dozens of artifacts across different tools, each generating their own collection of activity traces. To effectively utilize activity traces, evaluation systems will need to incorporate new techniques for analyzing activity traces across different tools, granularities, and contexts as well as extracting the information that is most relevant to the evaluation. The CSCW community has already explored a range of different visualizations and representations for activity traces [15, 52]. Future work should build on these approaches in order to address the increased complexities of contexts such as long-term collaborations, multiplicity of platforms, and hybrid work.

6.4 Limitations and Future Work

Type and duration of collaboration: Our study examined peer evaluations in the context of a single short-term collaborative task. Future work is needed to understand how activity traces might affect peer evaluations of teamwork over a longer duration and for different types of collaborative work. Participants in our study also did not know each other before the task and did not have to continue working together after the task. Therefore, the social ties between members of a team in our study are likely not as strong as in a team that has been working together longer. Prior work has suggested that as familiarity and interpersonal contact grows within a team, performance may be increasingly interpreted through the lens of abstract impressions rather than concrete behaviors [7]. Therefore, it is likely that evaluation biases will accrue over time for longer-term projects. Future work can investigate whether activity traces may then have a stronger impact against these biases or may perhaps be overwhelmed by them.

Another consequence of a longer-term context is that peer evaluations are often conducted on a continual basis throughout the course of the teamwork in organizational and academic settings. As evaluators gain more experience with reviewing the activity traces, they may become more adept at interpreting the traces and extracting the information that is useful to them for evaluation. Future work can examine what strategies peers adopt in processing information from the activity traces when performing peer evaluations. These strategies can potentially be automated and incorporated into intelligent peer evaluation tools.

Evaluation scheme: For the peer evaluation in our study, we used a forced-distribution rating scheme because our goal was to encourage differentiation of individual contributions. This type of rating may not capture different dimensions of teamwork skills or contributions. Future work can explore the impact of activity traces on other types of evaluations that serve different purposes such as providing feedback for improvement or teaching teamwork skills. Furthermore, numeric ratings may suggest a bias towards quantitative interpretations of contributions over qualitative. Future work can examine qualitative schemas for interpreting activity traces.

Representation of activity traces: In this study, we used the document version history and chat log in Google Docs as a starting point for examining the effects of activity traces. There is minimal summarization of the work being performed in these activity traces. Future work should examine how different representations of activity traces might impact the findings in this paper and address issues of scale and interpretation.

Populations studied: Finally, our findings may be limited to the populations that we recruited in the experiment: students and MTurk workers. Other populations in alternative contexts should be studied in order to test the generalizability of our findings, such as teams that work together in traditional organizations.

7 CONCLUSION

In this paper, we examined the impact of using activity traces on peer evaluations of individual contributions to teamwork. Through a between-subjects experiment over two different participant populations and a combination of Bayesian and qualitative analysis, we provide empirical evidence that using activity traces can address social and cognitive limitations of peer evaluations, such as selective recall and biased standards, by improving consistency as well as perceptions of accuracy of evaluations received. Furthermore, our usage of authentic activity traces that already exist today on one of the most popular collaborative work platforms in the world demonstrates that anyone can begin leveraging activity traces in their own lives today. However, stakeholders who are interested in using activity traces should carefully consider their context and goals for collaboration and evaluation. The open-ended responses from our study suggest issues of scale, interpretation, and performativeness that must be addressed as activity traces continue to increase in relevance. We hope that our work motivates further exploration of the impact of activity traces and the design of new data-enhanced peer evaluation tools.

8 ACKNOWLEDGEMENTS

This work was supported in part by NSF award IIS-2016908. We would like to thank our colleagues and the anonymous reviewers for their invaluable feedback. Additional thanks to the TurkerNation community for their suggestions on recruitment and study design.

REFERENCES

[1] 2021. Are peer reviews the future of performance evaluations? https://hbr.org/2021/01/are-peer-reviews-the-future-of-performance-evaluations

[2] 2022. The rise of performative work. https://www.economist.com/business/2022/01/07/the-rise-of-performative-work
[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] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* 20 (2019), 28:1–28:6. http://jmlr.org/papers/v20/18-403.html
[6] 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
[7] Angelo S. DeNisi. 1996. *A Cognitive Approach to Performance Appraisal.* Routledge, London.
[8] Angelo S. DeNisi, Thomas P. Cafferty, and Bruce M. Meglino. 1984. A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance* 33, 3 (1984), 360–396. https://doi.org/10.1016/0030-5073(84)90029-1
[9] Angelo S. DeNisi and Kevin R. Murphy. 2017. Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology* 102, 3 (2017), 421–433. https://doi.org/10.1037/apl0000085
[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] Paul Dourish and Victoria Bellotti. 1992. Awareness and Coordination in Shared Workspaces. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work* (Toronto, Ontario, Canada) *(CSCW '92)*. Association for Computing Machinery, New York, NY, USA, 107–114. https://doi.org/10.1145/143457.143468
[12] 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
[13] 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
[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] 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
[16] Graham R Gibbs. 2007. Thematic coding and categorizing. *Analyzing qualitative data. London: Sage* (2007), 38–56.
[17] 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
[18] Chyng-Yang Jang, Charles Steinfield, and Ben Pfaff. 2002. Virtual team awareness and groupware support: an evaluation of the TeamSCOPE system. *International Journal of Human-Computer Studies* 56, 1 (2002), 109–126. https://doi.org/10.1006/ijhc.2001.0517
[19] Jeffrey S. Kane and Edward E. Lawler. 1978. Methods of peer assessment. *Psychological Bulletin* 85, 3 (1978), 555–586. https://doi.org/10.1037/0033-2909.85.3.555
[20] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. 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, 4521–4532. https://doi.org/10.1145/2858036.2858465
[21] J.F. Kihlstrom, Eric Eich, D. Sandbrand, and B.A. Tobias. 2000. Emotion and memory: Implications for self-report (with a critique of retrospective analyses). *The science of self-report: Implications for research and practice* (01 2000), 81–99.
[22] Diane Lambert. 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* 34, 1 (1992), 1–14. https://doi.org/10.1080/00401706.1992.10485228
[23] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (01 Apr 2017), 433–442. https://doi.org/10.3758/

s13428-016-0727-z

[24] 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

[25] 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

[26] Ioanna Lykourentzou, Robert E. Kraut, and Steven P. Dow. 2017. Team Dating Leads to Better Online Ad Hoc Collaborations. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2330–2343. https://doi.org/10.1145/2998181.2998322

[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] Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition* (2 ed.). CRC Press. http://xcelab.net/rm/statistical-rethinking/

[30] 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

[31] Kevin R. Murphy and Jeanette N. Cleveland. 1991. *Performance appraisal: An organizational perspective.* Allyn & Bacon, Needham Heights, MA, US. xiv, 349–xiv, 349 pages.

[32] Kevin R. Murphy, Carmen Martin, and Magda Garcia. 1982. Do behavioral observation scales measure observation? *Journal of Applied Psychology* 67, 5 (1982), 562–567. https://doi.org/10.1037/0021-9010.67.5.562

[33] Jo-Anne Murray and Sharon Boyd. 2015. A Preliminary Evaluation of Using WebPA for Online Peer Assessment of Collaborative Performance by Groups of Online Distance Learners. *International Journal of E-Learning & Distance Education* 30, 2 (2015), n2.

[34] 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

[35] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188. https://doi.org/10.1177/0963721414531598 arXiv:https://doi.org/10.1177/0963721414531598

[36] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.

[37] Christina Peterson and N. Peterson. 2011. Impact of Peer Evaluation Confidentiality on Student Marks. *International Journal for the Scholarship of Teaching and Learning* 5 (07 2011), 1–13. https://doi.org/10.20429/ijsotl.2011.050213

[38] 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

[39] Du Phan, Neeraj Pradhan, and Martin Jankowiak. 2019. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554* (2019).

[40] Richard Saavedra and Seog K. Kwun. 1993. Peer evaluation in self-managing work groups. *Journal of Applied Psychology* 78, 3 (1993), 450–462. https://doi.org/10.1037/0021-9010.78.3.450

[41] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2017. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1700–1713. https://doi.org/10.1145/2998181.2998300

[42] 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

[43] 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 (oct 2021), 20 pages. https://doi.org/10.1145/3479576

[44] 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

[45] Patricia Cain Smith and L. M. Kendall. 1963. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology* 47, 2 (1963), 149–155. https://doi.org/10.1037/h0047060

[46] Igor Steinmacher, Ana Paula Chaves, and Marco Aurélio Gerosa. 2013. Awareness Support in Distributed Software Development: A Systematic Review and Mapping of the Literature. *Computer Supported Cooperative Work (CSCW)* 22, 2 (01 Apr 2013), 113–158. https://doi.org/10.1007/s10606-012-9164-4

[47] H. Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. 2012. Social Transparency in Networked Information Exchange: A Theoretical Framework. 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, 451–460. https://doi.org/10.1145/2145204.2145275

[48] Lorne M. Sulsky and William K. Balzer. 1988. Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology* 73, 3 (1988), 497–506. https://doi.org/10.1037/0021-9010.73.3.497

[49] 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

[50] Lori Foster Thompson, Jeffrey D. Sebastianelli, and Nicholas P. Murray. 2009. Monitoring online training behaviors: Awareness of electronic surveillance hinders e-learners. *Journal of Applied Social Psychology* 39, 9 (2009), 2191–2212. https://doi.org/10.1111/j.1559-1816.2009.00521.x

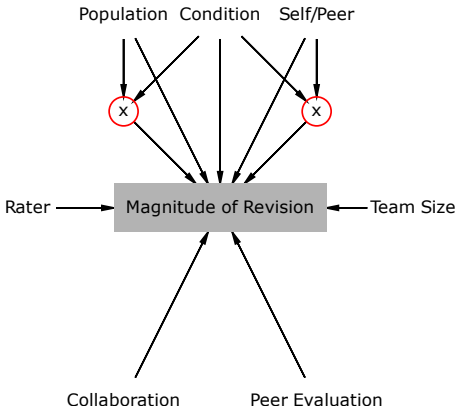
[51] Aharon Tziner, Kevin R. Murphy, Jeanette N. Cleveland, Guy Beaudin, and Sylvie Marchand. 1998. Impact of Rater Beliefs Regarding Performance Appraisal and Its Organizational Context on Appraisal Quality. *Journal of Business and Psychology* 12, 4 (01 Apr 1998), 457–467. https://doi.org/10.1023/A:1025003106150

[52] 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

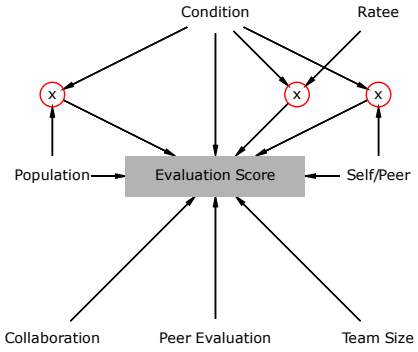
[53] Suzanne Weisband. 2002. *Maintaining awareness in distributed team collaboration: Implications for leadership and performance.* Boston Review, Cambridge, MA, US, 311–333. https://doi.org/10.7551/mitpress/2464.001.0001

[54] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (nov 2018), 27 pages. https://doi.org/10.1145/3274465

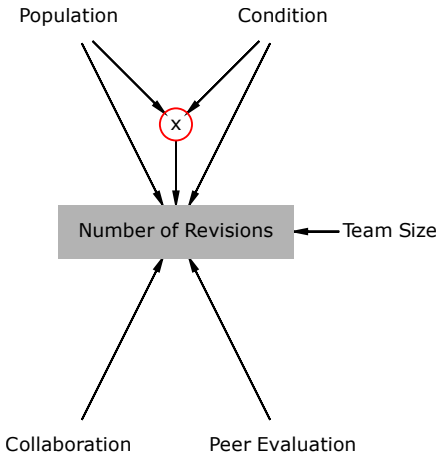
[55] Bo Zhang and Matthew W. Ohland. 2009. How to Assign Individualized Scores on a Group Project: An Empirical Evaluation. *Applied Measurement in Education* 22, 3 (2009), 290–308. https://doi.org/10.1080/08957340902984075



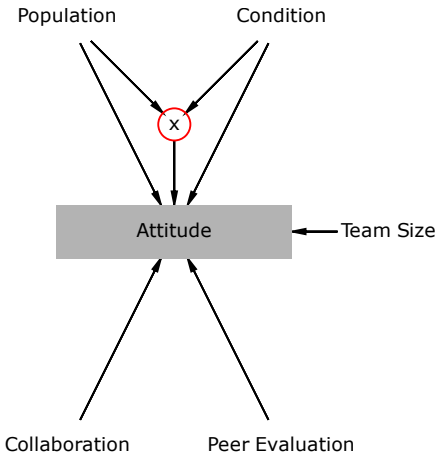
(a) Model 1: Magnitude of Revisions



(b) Model 3: Evaluation Score



(c) Model 2: Number of Revisions



(d) Model 4: Attitudes

Fig. 13. Causal DAGs for each of the four Bayesian models in our analysis. DAGs represent causal relationships between variables, with the arrows determining the direction of causal effect ($x \rightarrow y$ means x affects y). The symbol 'x' circled in red represents an interaction effect between the two variables that point to it. We note that these DAGs represent our beliefs about the causal relationships in our experiment. We do not derive these relationships from the data.

A APPENDIX

A.1 Model Definitions

Now, we discuss the Bayesian formulation for each of the models in our analysis. In a Bayesian formulation, we first need to define a likelihood function to model the outcome variable under the

control condition and logs condition. Consistent with McElreath [29], we posit that the likelihood function represents the modeler's view of the data, and not a claim about the world. In general, the likelihood function is a parametric formulation and we treat each model parameter as a random variable, drawn from another distribution with parameters (its prior distribution). In this way, Bayesian formulations are conducive to hierarchical or multilevel models. Bayesian researchers typically recommend "weakly informative" priors — conservative priors which allow for all possible values of the parameter but are chosen in a manner that promotes fast convergence.

Figure 13 displays the full Directed Acyclic Graphs (DAGs) manifest in our Bayesian models. We emphasize that these DAGs represent our beliefs about the causal relationships between the variables in our models. They are not derived from the data. Our main predictor variable of interest is the experimental condition. We also include the following covariates in all of our models: population (student vs. MTurk), team size, and collaboration and peer evaluation experience. As we discussed earlier in the paper, MTurk participants are more financially motivated compared to other recruitment populations. This may impact how they respond to the experimental treatment. Therefore, we include a varying slope for the interaction between population and experimental condition in our models. Below, we describe the variables that are included in all of our models. Other variables specific to an individual model are described in the respective model definition.

β_{s_i}: Varying intercept for the effect of the experimental condition s of rater i
P_{p_i}: Varying intercept for the effect of the population p (MTurk or student) of rater i
ϵ_{p_i,s_i}: Varying slope for the interaction effect between the population p of rater i and the condition s of rater i
T $\sum_{n=0}^{t[i]-1} \delta_t$: Effect for team size of rater i modeled as an ordinal predictor.
C $\sum_{n=0}^{c[i]-1} \delta_c$: Effect for collaboration experience of rater i modeled as an ordinal predictor.
E $\sum_{n=0}^{e[i]-1} \delta_e$: Effect for peer evaluation experience of rater i modeled as an ordinal predictor.

A.1.1 RQ1: Magnitude of Revisions (Model 1). There is one outcome variable: $c_{i,j}$, the magnitude of revision between the initial evaluation score and final evaluation score of participant i towards participant j. We aim to fit a distribution for the mean (i.e. the expected) magnitude of revision between initial evaluation scores and final evaluation scores.

We expected many participants would not make revisions to their initial evaluations and that the lack of revision could be due to different factors (i.e. initial calibration was accurate, lack of effort, etc.). Therefore, we use a zero-inflated Poisson (ZIP) distribution to characterize the mean magnitude of revision in both conditions (control and logs). A ZIP distribution is a mixture model that accounts for an excess of zeroes in the data by using a mixture of a Poisson distribution and a zero-probability distribution [22]. The ZIP distribution has two parameters: the probability of a 0 (ϕ) and the mean of the Poisson distribution, also known as the rate of change (λ). Both parameters are random variables, and we need to define likelihood functions for them. Our Bayesian model:

$$c_{i,j} \sim \text{ZIP}(\phi_{i,j}, \lambda_{i,j}) \quad (1)$$

$$\log(\lambda_{i,j}) = \alpha_\lambda + \gamma_{r[i]} + \beta_{s[i]} + P_{p[i]} + \epsilon_{p[i],s[i]} + O_{w[i,j]} + o_{w[i,j],s[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + C \sum_{n=0}^{c[i]-1} \delta_c + E \sum_{n=0}^{e[i]-1} \delta_e \quad \text{Linear model for } \lambda \ (2)$$

$$\text{logit}(\phi_{i,j}) = \alpha_\phi + \beta_{s[i]} + P_{p[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + o_{w[i,j],s[i]} \quad \text{Linear model for } \phi \ (3)$$

$$\epsilon_{p,s} \sim \text{MVNormal}([\bar{\mu_p}, \bar{\mu_s}], S_p) \quad \text{Prior for varying slope (4)}$$

$$o_{w,s} \sim \text{MVNormal}([\bar{\mu_w}, \bar{\mu_s}], S_w) \quad \text{Prior for varying slope (5)}$$

$$\gamma_r \sim \text{N}(\bar{\mu_1}, \sigma_1) \quad \text{Prior for each rater (6)}$$

$$\beta_s \sim \text{N}(\bar{\mu_2}, \sigma_2) \quad \text{Prior for each condition (7)}$$

$$P_p \sim \text{N}(\bar{\mu_3}, \sigma_3) \quad \text{Prior for each population (8)}$$

$$O_w \sim \text{N}(\bar{\mu_4}, \sigma_4) \quad \text{Prior for self/peer (9)}$$

$$T \sim \text{N}(\bar{\mu_5}, \sigma_5) \quad \text{Prior for team size (10)}$$

$$C \sim \text{N}(\bar{\mu_6}, \sigma_6) \quad \text{Prior for collaboration (11)}$$

$$E \sim \text{N}(\bar{\mu_7}, \sigma_7) \quad \text{Prior for peer evaluation (12)}$$

$$\delta_t \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_t \ (13)$$

$$\delta_c \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_c \ (14)$$

$$\delta_e \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_e \ (15)$$

$$\alpha_\lambda \sim \text{N}(0, 1) \quad \text{Prior for intercept of } \alpha_\lambda \ (16)$$

$$\alpha_\phi \sim \text{N}(1, 0.5) \quad \text{Prior for intercept of } \alpha_p \ (17)$$

Equation (1) describes that the magnitude of revision is modeled as a ZIP distribution with zero-probability p and Poisson mean λ. There are two linear models and two link functions, one for each parameter of the ZIP distribution. Equation (2) gives the linear model for the Poisson mean and is modeled on the log scale. This predicts the mean magnitude of revision between the initial and final evaluations participant i gave to participant j. Equation (3) gives the linear model for the probability of zero and is modeled on the logit scale. This predicts the probability of zero revision between the initial and final evaluations participant i gave to participant j. The rest of the model definition lists the priors of the parameters of the two linear models. In addition to the variables that are common to all of our models, we include the following in the linear models:

α_λ: Base magnitude of revision.
α_ϕ: Base probability of no revision.
$\gamma_{r[i]}$: Varying intercept for the effect of rater i.
$O_{w_{i,j}}$: Varying intercept for self/peer (i.e. whether or not for evaluation $[i,j]$, i=j). This represents whether the evaluation was a self or peer evaluation.
$o_{w_{i,j},s_i}$: Varying slope for the interaction effect between self/peer of evaluation $[i,j]$ and condition of rater i.

We note the inclusion of self/peer as another predictor variable in our model which represents whether the evaluation was a self or peer evaluation.

We set the priors in our model to be weakly informative. For all $\bar{\mu}$ parameters, we set priors to be N(0, 0.25). For all σ parameters, we use exponential priors i.e. $\sigma \sim$ Exponential(1). In addition, for the varying slopes priors, we note that each S in the varying slopes represents the covariance matrices specific to the population or self/peer category, respectively. They can each be factored into their own σ parameters and correlation matrix R.

A.1.2 RQ1: Number of Revisions (Model 2). There is one outcome variable: n_i, the number of team members' evaluations that participant i revised, where each revised evaluation represents one revision. We model the number of revisions as an ordinal outcome because each successive level of revision implies and is greater than the previous level of revision. For example, a participant cannot make four revisions to their evaluations without having also made three revisions.

We use an Ordered Logistic distribution to characterize the number of revisions. This distribution uses the cumulative link function to guarantee the ordering of the outcome variable. The Ordered Logistic distribution has two parameters: a linear predictor (ϕ) and a vector of cutpoints for the response values (κ). Both parameters are random variables, and we need to define likelihood functions for them. Our Bayesian model:

$$n_i \sim \text{Ordered-logit}(\phi_i, \kappa) \tag{1}$$

$$\phi_i = \beta_{s[i]} + P_{p[i]} + \epsilon_{p[i],s[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + C \sum_{n=0}^{c[i]-1} \delta_c + E \sum_{n=0}^{e[i]-1} \delta_e \quad \text{Linear model for } \phi \tag{2}$$

$$\kappa_k \sim \text{N}(0, 1.5) \quad \text{Common prior for each intercept } k \tag{3}$$

$$\epsilon_{p,s} \sim \text{MVNormal}([\bar{\mu_p}, \bar{\mu_s}], S_p) \quad \text{Prior for varying slope} \tag{4}$$

$$\beta_s \sim \text{N}(\bar{\mu_1}, \sigma_1) \quad \text{Prior for each condition} \tag{5}$$

$$P_p \sim \text{N}(\bar{\mu_2}, \sigma_2) \quad \text{Prior for each population} \tag{6}$$

$$T \sim \text{N}(\bar{\mu_3}, \sigma_3) \quad \text{Prior for team size} \tag{7}$$

$$C \sim \text{N}(\bar{\mu_4}, \sigma_4) \quad \text{Prior for collaboration} \tag{8}$$

$$E \sim \text{N}(\bar{\mu_5}, \sigma_5) \quad \text{Prior for peer evaluation} \tag{9}$$

$$\delta_t \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_t \tag{10}$$

$$\delta_c \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_c \tag{11}$$

$$\delta_e \sim \text{Dirichlet}(2) \quad \text{Prior for } \delta_e \tag{12}$$

Equation (1) states that the number of revisions is modeled as an Ordered Logistic distribution with linear predictor ϕ and a vector of cutpoints κ. Equation (2) gives the linear model for ϕ and is modeled on the cumulative logit scale. This predicts the log-cumulative-odds of each number of revisions when subtracted from the cutpoint of that number. Equation (3) gives the vector of cutpoints associated with each number of revisions. All other parameters have been described previously in the Appendix. We set priors in the same way as for the magnitude of revisions model.

A.1.3 RQ2: Consistency of Revisions (Model 3). There is one outcome variable: $r_{i,j}$, the evaluation score participant i received from participant j. In answering RQ2, we are interested in the standard deviation of evaluation scores received by participants. We use a Normal distribution to characterize the evaluation scores. As we only assume a mean and variance, the Normal distribution is the most conservative distribution we can use to model the data [29]. We also expected that most

participants would receive an average evaluation score, with fewer exceptional cases of over- or under-contributors. The Normal distribution has two parameters: the mean (μ) and standard deviation (σ). Both parameters are random variables, and we need to define likelihood functions for them. Our Bayesian model:

$$x_{i,j} \sim \mathrm{N}(\mu_{i,j}, \sigma_{i,j}) \tag{1}$$

$$\mu_{i,j} = \alpha_\mu + Z_{r[i]} + \beta_{s[i]} + P_{p[i]} + \epsilon_{p[i],s[i]} + O_{w[i,j]} + o_{w[i,j],s[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + C \sum_{n=0}^{c[i]-1} \delta_c + E \sum_{n=0}^{e[i]-1} \delta_e + \zeta_{r[i],s[i]} \quad \text{Linear model for } \mu \tag{2}$$

$$\log(\sigma_{i,j}) = \alpha_\sigma + \beta_{s[i]} + P_{p[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + O_{w[i,j]} \quad \text{Linear model for } \sigma \tag{3}$$

$$\epsilon_{p,s} \sim \mathrm{MVNormal}([\bar{\mu_p}, \bar{\mu_s}], S_p) \quad \text{Prior for varying slope} \tag{4}$$

$$o_{w,s} \sim \mathrm{MVNormal}([\bar{\mu_w}, \bar{\mu_s}], S_w) \quad \text{Prior for varying slope} \tag{5}$$

$$\zeta_{r,s} \sim \mathrm{MVNormal}([\bar{\mu_r}, \bar{\mu_s}], S_r) \quad \text{Prior for varying slope} \tag{6}$$

$$Z_r \sim \mathrm{N}(\bar{\mu_1}, \sigma_1) \quad \text{Prior for each ratee} \tag{7}$$

$$\beta_s \sim \mathrm{N}(\bar{\mu_2}, \sigma_2) \quad \text{Prior for each condition} \tag{8}$$

$$P_p \sim \mathrm{N}(\bar{\mu_3}, \sigma_3) \quad \text{Prior for each population} \tag{9}$$

$$O_w \sim \mathrm{N}(\bar{\mu_4}, \sigma_4) \quad \text{Prior for self/peer} \tag{10}$$

$$T \sim \mathrm{N}(\bar{\mu_5}, \sigma_5) \quad \text{Prior for team size} \tag{11}$$

$$C \sim \mathrm{N}(\bar{\mu_6}, \sigma_6) \quad \text{Prior for collaboration} \tag{12}$$

$$E \sim \mathrm{N}(\bar{\mu_7}, \sigma_7) \quad \text{Prior for peer evaluation} \tag{13}$$

$$\delta_t \sim \mathrm{Dirichlet}(2) \quad \text{Prior for } \delta_t \tag{14}$$

$$\delta_c \sim \mathrm{Dirichlet}(2) \quad \text{Prior for } \delta_c \tag{15}$$

$$\delta_e \sim \mathrm{Dirichlet}(2) \quad \text{Prior for } \delta_e \tag{16}$$

$$\alpha_\mu \sim \mathrm{N}(3, 2) \quad \text{Prior for intercept of } \alpha_\lambda \tag{17}$$

$$\alpha_\sigma \sim \mathrm{N}(0, 1) \quad \text{Prior for intercept of } \alpha_p \tag{18}$$

We note the inclusion of self/peer as another predictor variable in our model, as described in the model for magnitude of revisions. In addition, we include the following:

α_μ: Base evaluation score.
α_σ: Base standard deviation.
$Z_{r[i]}$: Varying intercept for the effect of ratee i.
$\zeta_{r[i],s[i]}$: Varying slope for the interaction effect between the ratee i and their condition

We set the priors in our model to be weakly informative. For all $\bar{\mu}$ parameters, we set priors to be N(0, 1). For all σ parameters, we use exponential priors i.e. $\sigma \sim$ Exponential(1). In addition, for the varying slopes priors, we note that each S in the varying slopes represents the covariance matrices specific to the population or self/peer category, respectively. They can each be factored into their own σ parameters and correlation matrix R.

A.1.4 RQ3: Attitudes (Model 4). There is one outcome variable: a_i, the attitude rating given by participant i. We model attitude ratings as an ordinal outcome since attitudes are rated on a 7-point

Likert scale. We use an Ordered Logistic distribution to characterize the ratings. The Bayesian model for attitude ratings is identical to the model for number of revisions, so we do not repeat the full model definition. Our Bayesian model:

$$a_i \sim \text{Ordered-logit}(\phi_i, \kappa) \tag{1}$$

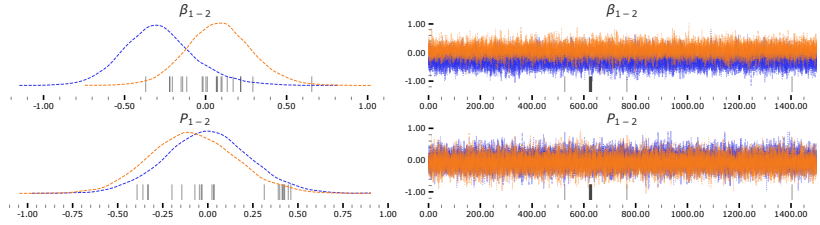
$$\phi_i = \beta_{s[i]} + P_{p[i]} + \epsilon_{p[i],s[i]} + T \sum_{n=0}^{t[i]-1} \delta_t + C \sum_{n=0}^{c[i]-1} \delta_c + E \sum_{n=0}^{e[i]-1} \delta_e \quad \text{Linear model for } \phi \tag{2}$$

$$\kappa_k \sim \text{N}(0, 1.5) \quad \text{Common prior for each intercept } k \tag{3}$$

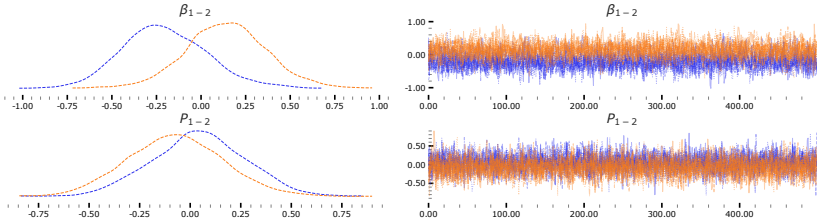
A.2 Model Convergence

We applied non-centered parameterization to all of the models in order to increase convergence. For all of the Bayesian models described, the Gelman-Rubin statistic (a measure of MCMC convergence) for all parameters was around 1, indicating that the multiple sampling chains converged. Traceplots for the MCMC chains in all of the models are shown in Figures 14- 16. We only show the traceplots for coefficients of condition and population, since these were the most significant variables in our models. The traceplots for the other parameters show that the models are equally well-behaved and can be added to the extended version of the paper.

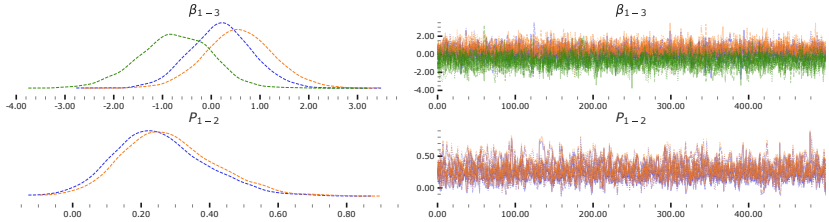
Received July 2022; revised October 2022; accepted January 2023



(a) Traceplot for magnitude of revision model.

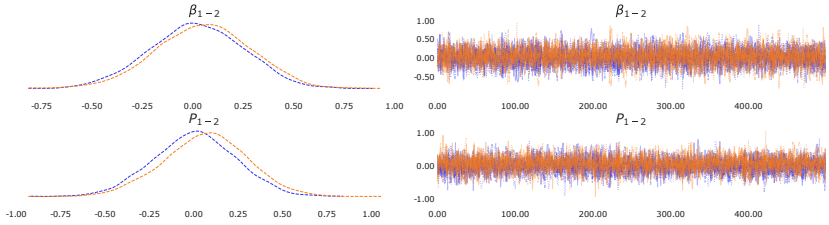


(b) Traceplot for number of revisions model.

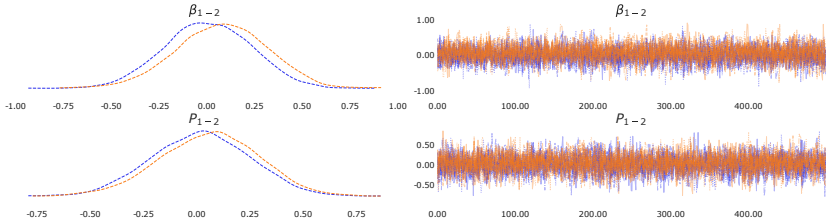


(c) Traceplot for consistency model (standard deviation of scores).

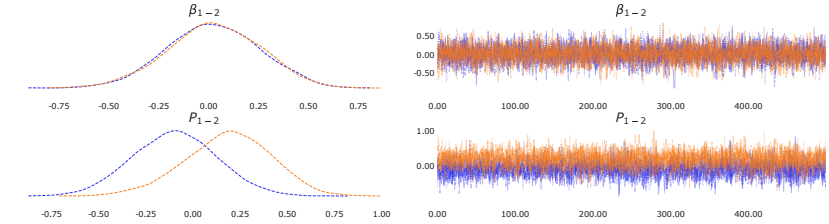
Fig. 14. Traceplots showing the results of the MCMC estimation for the magnitude of revisions model, the number of revisions model, and the consistency model (RQ1-2). The left column shows the posterior distributions for β_{1-2} and P_{1-2}, the condition and population effects, respectively. Note that for the consistency model, we can see that there are three distributions for β. This is because we wanted to compare the change in distribution from the treatment to initial evaluation and the control to initial evaluation. The right column shows the corresponding sampling traces. The color mappings for the β_{1-2} plots are: blue – Control, orange – Treatment. The color mappings for the β_{1-3} plots are: blue – Baseline, orange – Control, green - Treatment. The color mappings for the P_{1-2} plots are: blue – MTurk, orange – Student.



(a) Traceplot for confidence in accuracy of evaluations given model.

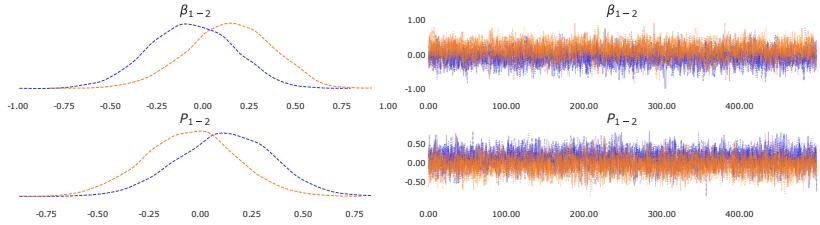


(b) Traceplot for confidence in accuracy of evaluations received model.

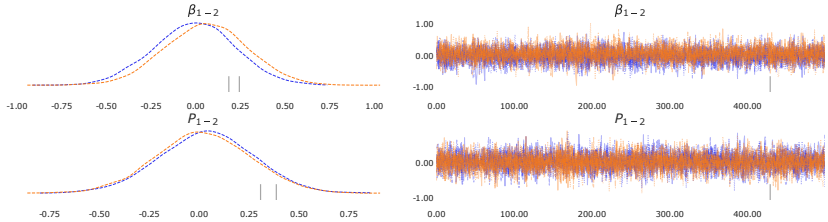


(c) Traceplot for perceived difficulty of evaluation model.

Fig. 15. Traceplots showing the results of the MCMC estimation for the post-evaluation attitude ratings (confidence in accuracy of evaluations given and received and perceived difficulty of evaluation) (RQ3). The left column is the posterior distributions for β_{1-2} and P_{1-2}, the condition and population effects. The right column shows the corresponding sampling traces. The color mappings for the β_{1-2} plots are: blue – Control, orange – Treatment. The color mappings for the P_{1-2} plots are: blue – MTurk, orange – Student.



(a) Traceplot for perceived accuracy of evaluations received model.



(b) Traceplot for satisfaction with evaluations received model.

Fig. 16. Traceplots showing the results of the MCMC estimation for the post-feedback attitude ratings and for perceived difficulty of evaluation (RQ1-2). The left column shows the posterior distributions for β_{1-2} and P_{1-2}, the condition and population effects. The right column shows the corresponding sampling traces. The color mappings for the β_{1-2} plots are: blue – Control, orange – Treatment. The color mappings for the P_{1-2} plots are: blue – MTurk, orange – Student.