Structure or Nurture? The Effects of Team-Building Activities and Team Composition on Team Outcomes

EMILY M. HASTINGS*, University of Illinois at Urbana-Champaign, USA

FARNAZ JAHANBAKHSH*, Massachusetts Institute of Technology, USA and University of Illinois at Urbana-Champaign, USA

KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA

DARKO MARINOV, University of Illinois at Urbana-Champaign, USA

BRIAN P. BAILEY, University of Illinois at Urbana-Champaign, USA

How can instructors group students into teams that interact and learn effectively together? One strand of research advocates for grouping students into teams with "good" compositions such as skill diversity. Another strand argues for deploying team-building activities to foster interpersonal relations like psychological safety. Our work synthesizes these two strands of research. We describe an experiment (N=249) that compares how team composition vs. team-building activities affect student team outcomes. In two university courses, we composed student teams either randomly or using a criteria-based team formation tool. Teams further performed team-building activities that promoted either team or task outcomes. We collected project scores, and used surveys to measure psychological safety, perceived performance, and team satisfaction. Surprisingly, the criteria-based teams did not statistically differ from the random teams on any of the measures taken, despite having compositions that better satisfied the criteria defined by the instructor. Our findings argue that, for instructors deploying a team formation tool, creating an expectation among team members that their team can perform well is as important as tuning the criteria in the tool. We also found that student teams reported high levels of psychological safety, but these levels appeared to develop organically and were not affected by the activities or compositional strategies tested. We distill these and other findings into implications for the design and deployment of team formation tools for learning environments.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**;

Keywords: Algorithms, CATME, Learning, Team formation, Psychological safety, Team-building.

1 INTRODUCTION

In higher education, many instructors are incorporating team-based learning into their courses to prepare students for careers in industry, where the ability to work in a team is highly valued [31]. However, instructors face a difficult challenge of deploying strategies for grouping students into teams and for helping the teams gel together. Choosing an appropriate strategy is essential, because

---

*The first two authors contributed equally and are listed alphabetically by last name.

an instructor's decisions regarding team formation can impact student team outcomes, learning, and satisfaction [14, 26].

Researchers in CSCW and related fields have investigated two independent strands of research from which instructors could draw to promote team outcomes and satisfaction. One strand highlights the need to form teams with "good" compositions. For example, prior work has shown that teams composed of members who have diverse skills [7, 15], balanced personality types [21], and balanced genders [17] outperform teams without these properties in several contexts. Instructors are increasingly leveraging algorithmic team formation tools such as CATME [20] to implement this strand of research in their courses as enrollments continue to grow. A challenge is that the effectiveness of a team formation tool depends in part on the instructor's understanding of the tool and his or her familiarity with the team composition literature [16]. For example, prior work has not tested whether the multi-criteria configurations that instructors typically define in practice are more effective than randomly grouping students into teams. Configurations defined in practice often synthesize criteria tested in multiple studies, e.g., combining task skills, gender, academic ability, and leadership orientation [4, 7, 21].

A separate strand of research investigates team-building activities that promote attributes of teamwork such as psychological safety (e.g., [34]). Psychological safety is the shared belief that the team is safe for interpersonal risk-taking [12]. A large-scale study in industry concluded that achieving high psychological safety contributes more to a team's success than the composition of the team does [11]. If this result transfers to student teams in learning environments, it might prompt instructors to prioritize the deployment of team-building activities to complement the use of team formation tools and possibly reduce the focus on configuring the "right" criteria in the tool. In addition, teams formed by algorithms can struggle with "cold starts" [16] because students assigned to the same team typically do not know one another and often differ along many of the criteria configured in the tool. The deployment of these tools may therefore be creating an increased need for team-building activities to help the teams overcome difficulties stemming from their lack of acquaintance.

In this work, we synthesize these two distinct strands of research to help instructors better manage the deployment of algorithmic team formation in their courses. In an experiment conducted in two university courses with team-based projects, we assigned students (N=249) to teams using two formation strategies: 1) creating teams using the criteria defined by the instructor in a tool, or 2) randomly. To ensure all teams perceived themselves as being on equal footing and capable of performing well, and due to the use of a criteria-based approach in prior instances of the courses, we informed students that all teams would be formed using the tool. However, students were not aware of the specific configuration that would be used. Teams were further assigned to perform team-building activities emphasizing the team (i.e., psychological safety) or the task (i.e., given outcomes). Teams performed activities from their assigned condition at the onset and midpoint of the teamwork. We collected several measures for each student, including project scores, and used surveys to measure psychological safety and aspects of team experience including perceived performance and satisfaction. We also analyzed essays written by students (N=92) to understand their perceptions of the team activities and formation strategy. The variables examined in our experiment are summarized in Figure 1.

Our results showed no statistical differences between the criteria-based and randomly-assigned teams on any of the measures. This result was surprising, given prior work showing that forming teams using subsets of the selected criteria relates to improved team performance (e.g., [7, 15, 21]). We hypothesize that the effectiveness of deploying a criteria-based team formation tool in this experiment is due in part to creating an expectation among team members that their team is capable
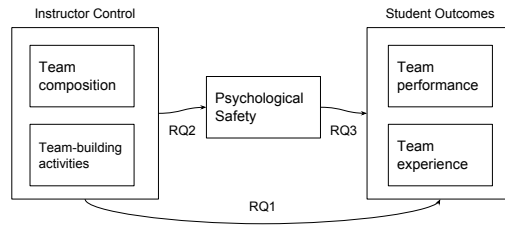
Fig. 1. The variables examined in our experiment. Each arrow represents a research question.

of performing well. For instructors, creating this effect when deploying a team formation tool to students may be as important as tuning the criteria in the tool.

In addition, most teams reported high levels of psychological safety. To the best of our knowledge, our study is the first to report the level of psychological safety achieved by student teams and extends prior work that has studied psychological safety in industry teams [12]. Psychological safety did not differ between the team- and task-focused activity conditions, or the two team formation conditions. Our results suggest that although students perceived practical benefits of both sets of activities tested, psychological safety may have developed organically regardless of the activities or formation strategies tested.

The main contribution of this work is a deeper empirical understanding of how algorithmic team formation, team building activities, and psychological safety impact team performance and satisfaction in authentic learning environments. We also offer practical implications for the design and deployment of team formation tools. For example, tool designers should include configuration exemplars that make it easier for instructors to mimic the criteria configurations shown to produce different positive outcomes in the literature.

2 RELATED WORK

We review the literature on team composition and team-building, and describe how our work synthesizes these strands. We also review studies of psychological safety and motivate the need to extend empirical knowledge of psychological safety to teams in learning environments.

2.1 Team Composition

In criteria-based approaches to team formation, instructors choose how to organize students into teams based on criteria such as students' skills, working styles, and demographics. Instructor choices can be informed by the literature on team composition. For example, research shows that skill diversity increases the performance of work teams [15]. Bear and Woolley show that placing more women on a team raises the team's collective intelligence [4]. Lykourentzou et al. report that balancing personality types within a team can foster performance and satisfaction [21]. Brickell finds grouping students based on academic ability and curricular interests improves team performance and satisfaction [7]. Wen et al. find that teams formed according to pairwise transactivity in a discussion activity showed greater knowledge integration than random teams in a simulated classroom on MTurk [44].

However, it remains an open question whether forming teams by specifying multiple criteria in parallel (e.g., to balance personality types, gender, and skills) would achieve the same empirical benefits of the individual criteria tested in prior work. For example, Zheng et al. studied algorithmic grouping with multiple criteria in the context of a MOOC, but found no conclusive evidence of a difference in drop-out rate or learning performance between algorithmically-assigned and

random teams [48]. Because most empirical knowledge of team composition has been derived through controlled tasks or settings, there is also a need to test the generalizability of criteria-based approaches in the context of authentic projects in traditional classroom settings.

Our work contributes to the literature on team composition by comparing outcomes between student teams formed using a team formation tool with multi-criteria configurations defined by instructors and teams formed randomly. Our experiment also tests these formation strategies in the context of authentic, multi-week team projects, and using a broader array of measures (i.e., both objective and subjective) compared to prior work.

2.2 Team-building Activities

Team-building refers to activities that focus on improving the team, its process, and its work [19]. Team-building is especially useful for teams formed by a third party such as the algorithm in a team formation tool. For example, students grouped using these tools have typically never met before and are often unaware of which members of the team identify with which task skills, working styles, or roles that were considered in the tool configuration. This unfamiliarity can cause teams to struggle to communicate at the onset of team work [16].

One way that team-building activities can address these types of problems is through cultivating interpersonal relations [37]. Activities that help teams foster relations can lead to improved team outcomes [6, 9, 18, 40, 41]. Supporting good relationships among team members may be especially important in teams formed using a criteria-based approach because the members typically differ from each other along the selected criteria. These individuals may struggle to find common ground with their teammates, and thus they may be in more need of additional support.

The knowledge of team-building has been produced almost entirely through field studies in organizations or controlled lab studies [19]. Our work is original because we study how team-building activities affect student teams in multi-week course projects. This study is also novel because we test how team-building activities interact with different strategies for team formation, and how the activities affect interpersonal relations as measured by psychological safety.

2.3 Psychological Safety

Psychological safety refers to a shared belief held by members of a team that the team is safe for interpersonal risk-taking. It has been shown to support learning behaviors and team performance in work teams [12], and can be encouraged through nurturing the interpersonal relationships within a team [9, 10, 38]. An industry study concluded that psychological safety is the main factor contributing to team performance in the workplace [11].

While most research on psychological safety has focused on teams working in industry, it is also important to examine its effect on team outcomes in educational contexts. Better understanding the relationship between psychological safety and team outcomes could allow instructors to improve student learning experiences through fostering psychological safety. A few studies have already begun to examine aspects of this relationship. For example, Adams and Ruiz report that psychological safety can explain perceived performance and attitudes toward teamwork in capstone project courses [3, 33].

Our work extends this prior literature in two ways. First, we expand current knowledge by relating psychological safety to objective measures of performance (project grades) in addition to the subjective measures studied in prior work. Second, our work sheds light on how different strategies for composing teams (criteria-based tools and random assignment) and different categories of team building activities impact the psychological safety reported by teams. To the best of our knowledge, we are also the first to report the quantified level of psychological safety achieved by student teams.

3 RESEARCH QUESTIONS

This work poses the following research questions:

RQ1: *How do composition and structured team-building activities affect student team performance and team experience?* We measure performance using team project grades, and team experience using surveys capturing perceived performance, satisfaction with team assignment, peer evaluations, cohesiveness, and conflict.

Hypothesis: We hypothesize that since the criteria-based teams are designed to have all the skills (among other complementary attributes) necessary to complete the group projects, these teams will outperform the randomly-assigned teams. In addition, this advantage may facilitate teamwork and lead to more positive experiences. We also expect that teams performing team-focused activities will have more positive experiences than those performing task-focused activities, due to an increased effort in developing interpersonal relationships among team members. Teams performing these activities may also exhibit different behaviors, which we anticipate observing through student essays.

RQ2: *To what degree do structured team-building activities and team composition influence psychological safety for student teams?*

Hypothesis: We anticipate that teams performing the team-focused activities will achieve higher psychological safety than those performing the task-focused activities, again due to the focus on building interpersonal relationships. We do not expect that team composition will affect psychological safety.

RQ3: *What is the relationship between psychological safety and team performance and experience? What level of psychological safety do student teams achieve?*

Hypothesis: Based on results from prior literature (e.g., [3, 11, 12, 33]), we expect that psychological safety will be correlated with performance and team experience measures.

Answers to these questions will provide instructors and CSCW researchers with a deeper empirical understanding of how to effectively group students into teams and how to nurture interpersonal relations within student teams.

4 METHOD

To answer our research questions, we conducted a full factorial between-participants experiment with two factors, *Composition* (criteria-based vs. random) and *Activity Focus* (team focus vs. task focus). The experiment was conducted in parallel in two computer science courses that incorporate team-based learning at a large public university. The study was approved by the IRB at our university.

4.1 Courses and Participants

The two courses involved in our study were User Interface (UI) Design and Software Engineering (SE) I. Both courses target upper-level undergraduate and beginning graduate students in computer science. There was very little overlap in student enrollment between these two courses. In total, 249 of the 304 students in these courses gave consent to use their data in our analysis.

4.1.1 UI Design Course. There were 176 students (50 female) in the UI design course. Approximately 25% of the students came from disciplines other than computer science, including engineering, psychology, and the arts. Topics included user research, prototyping, evaluation, and event-based programming. In the course, teams of 4-6 students worked to complete a nine-week user interface design project of their choice. Examples of projects include a personalized mobile news app and an online book rental interface. Teams submitted project deliverables in stages, including project proposals, user and task reports, low-fidelity prototypes, functional prototypes,

and user test results. Team projects comprised 40% of students' final course grades. There were 37 teams (18 random, 19 criteria-based) in this course.

4.1.2 SE Course. The SE course had 128 students (24 female). Topics included agile software processes, software configuration managements, refactorings, testing, metrics, and design patterns. Students completed a seven-week team project (6-8 students per team). The project required teams to build a configurable "pretty printer" for the K executable semantic framework [1]. For the project, there was an initial week of organization followed by three two-week iterations. Deliverables including user stories, source code, and written reports were submitted at each iteration for credit. For the first two iterations, teams implemented required aspects of the "pretty printer" such as adding brackets to improve human readability. Teams chose their own feature to implement for the final iteration. Examples included adding a consistency check and allowing interactive expansion of the content. The project was worth 40% of students' final course grade. There were 18 teams (9 random, 9 criteria-based) in this course.

4.2 Activity Focus

To determine the extent to which structured team-building activities can influence psychological safety, we developed two sets of exercises for teams to complete for the experiment. One set of exercises, used for the experimental condition, aimed at nurturing psychological safety within teams (*team focus* activities). The other set, used as a control condition, focused on producing outcomes (*task focus* activities). Teams completed these activities in addition to the other deliverables required for the team project. We randomly assigned half of the teams to perform the team focus activities, and the other half to the task focus activities. Two sets of activities were introduced to minimize the additional workload on teams, and were spaced at the onset and midpoint of the team work.

4.2.1 Team Focus Activities. Teams in the *team focus* condition performed two scripted exercises. For the first exercise, completed just after team formation, teams met and shared their "About Me" page on their favorite social media platform and up to three recently posted photos that represented who they are as a person or an interesting aspect of themselves. If students did not have or want to share information from social media, they could share physical mementos that served a similar purpose. Teams then produced a written summary of what they learned about each other and submitted it to the course staff for credit.

The purpose of this activity was to nurture interpersonal relationships, as improved relations have been shown to correlate with feelings of psychological safety within teams [9, 10, 38]. By fostering an initial knowing of teammates and directing one's focus towards others [13, 28], we also wanted team members to begin to differentiate their respective behaviors, cognitions, and emotions and build connections with each other. Among many choices, we opted for an exercise involving social media because prior work has shown that it can be used to enable social perceptiveness between teammates [47]. In addition, research has shown that sharing photos can increase awareness about the contexts of others and encourage pro-social behavior towards teammates [24, 25, 42].

The second activity, performed halfway through project work, required each student to write an essay describing a few strengths of each of their teammates and a few of their own weaknesses in regards to their contribution to the project. Teams shared these individual essays with each other and wrote a document summarizing how the team could work better together for the remainder of the project. The purpose was to further nurture interpersonal relations and to foster respectful engagement [30]. This activity is similar to the teamwork intervention in [45], in which teams were asked to discuss how to work well together.

4.2.2 Task Focus Activities. Teams in the *task focus* condition also performed two exercises. For the first exercise, groups brainstormed team names, and selected the one that best fit the team's personality. Teams composed a document with the selected name and an explanation of it, and a list of at least twenty candidate names that were generated as evidence of productive brainstorming. Some of the names chosen were "Brogrammers," "The Fellowship of Designers," "4+2=7," "8 bits me," and "Will Code For Money." The purpose was to require teams to work together and produce an outcome. Moreover, the assignment prompted students to follow given principles of effective brainstorming. This epistemic script [43] was not present in the team focus condition. The task was also chosen because, at this early point in the course, students had not been exposed to enough lecture content to perform an activity related to the course content. In addition, it is consistent with activities previously performed by students in these courses and described in the literature (e.g., [36]).

For the second exercise, each team member wrote an individual essay describing how the *project* could be improved moving forward. The team members met and shared each other's essays, and wrote an action plan for the remainder of the project. This activity is very similar to the strategy intervention in [45], in which teams defined a strategy for their task, identified potential obstacles, and decided on how to best use resources.

The task-focused activities were designed to evoke mechanics of collaboration and communication (e.g., generating ideas, defining group processes, and reaching consensus), and emphasized producing an outcome (e.g., team name). In contrast, the team-focused activities were designed to evoke inter-personal communication between the members of a team rather than the mechanics of collaboration. For example, the kinds of interaction between members prompted by brainstorming a team name are much shallower and narrower than those experienced when sharing personal photos and mementos. These interactions are therefore likely to be less effective at building positive regard, and consequently psychological safety [9, 38].

The two sets of activities were performed in parallel between conditions. Students were aware that different versions of the activities existed, but were unaware of the specific assignments. To ensure a comparable experience and workload between conditions, students were instructed to spend an hour with their team on each activity. To encourage completion, students had to report the date, time, and duration of the meeting during which they completed the activity, as well as to submit the written results of the activity to the course staff for credit. These results were graded according to a rubric which had been made available to the students when the activities were assigned.

4.3 Team Composition

To determine how composition impacts team outcomes and how this effect relates to team-building activities and psychological safety, students were assigned to teams using two composition strategies: criteria-based and random. Half the students were randomly assigned to the criteria-based condition, the other half to the random condition.

4.3.1 Criteria-based. In the criteria-based condition, the instructors assigned students to teams using an online tool called CATME [2]. The instructors first surveyed students' working styles, skills, demographics, and other criteria by selecting from default criteria already available in the tool, or creating their own criteria. The tool generated a survey based on the criteria selected and emailed it to the students in the course.

After the student response period passed, the instructors reviewed the collected data and configured the weights for each criterion (see Table 1). These selections serve as input into an algorithm that optimizes team assignments. The weights are assigned on the scale [-5, +5], with 0 indicating

Table 1. The criteria configuration defined by the instructors to form teams in the criteria-based condition. Blank cells indicate the criteria was not included for that course. The two categories marked with asterisks are those the SE course used to assign teams in the random (control) condition.

Criteria	SE	UI
Schedule*	5	
Big-Picture vs. Detail-Oriented	-4	-5
Gender	4	4
GPA	-4	
Leadership Role	-4	-3
Race	3	
Writing Skills		-3
Weekend Meetings*	5	
Leadership Preference	1	-3
Commitment Level	-4	-2
English Skills	-3	
Software Skills	-5	
On-Campus Job	4	
Off-Campus Job	4	
Course skills (programming, graphic design, communication, etc.)		-5
Location (on-campus vs. online)		2

to ignore the criterion. Negative weights indicate that the algorithm should group students with dissimilar responses for that criterion, while positive weights indicate that students with similar responses should be grouped. For example, assigning the "Schedule" criterion a weight of +5 prefers teams with compatible schedules, while assigning the "Writing Skills" criterion a weight of -3 prefers teams with a mix of self-reported writing ability. The magnitude of the weight indicates the strength of the preference.

The criteria configurations in each course were different. Each instructor configured the criteria based on what s/he felt was most appropriate based on the course's learning goals and project requirements, their prior experiences with teams in the course, and the team formation literature. Both instructors had prior experience using CATME, and drew from this knowledge when configuring the tool. Once the instructors configured the criteria and specified the desired team size, the CATME algorithm generated teams.

4.3.2 Random. For the non-criteria condition, students in the UI design course were assigned to teams randomly, and the team assignments were manually entered in the tool. Students in the SE course were assigned to teams based solely on schedule compatibility criterion in the tool. The instructor of this course felt this was necessary because the teams would have 6-8 students, and finding meeting times for groups of this size was problematic in prior instances of the course.

To keep students blind to condition, all of the students completed the survey distributed by the team formation tool and were notified of their team assignments via the tool, even those who had been assigned randomly. We felt that it was most ethical to keep expectations consistent across conditions and ensure that students in the random condition would not perceive being disadvantaged by the condition. In addition, by having all students interact with the tool, we were able to utilize its peer evaluation instruments for all students at the end of the semester (see Measures).

4.4 Procedure

The instructors of each course knew the objectives of the study (the UI instructor is a member of the research team). However, as required by the IRB at our institution, we had a number of protocols in place to reduce any actual or perceived conflict of interest and potential bias. The team formation in each course was performed by one member of the course staff (who was neither the instructor nor a member of our research team), and the other course staff members did not know which team was formed in which condition (random vs. criteria-based) or the significance of the different activity types. The instructors in each course selected criteria that they believed to be the best for their own course. The research team was also blind to condition.

In both courses, members of our research team spent approximately 10 minutes during a lecture in the first few weeks of class describing the team formation tool. The tool had already been used in previous instances of both courses and there was a general expectation of its use. Students then had approximately one week to fill out the survey generated by the tool. Response rates were 91% for the SE course and 97% for the UI course. Team assignments were announced approximately one week after the surveys were collected, during course week 3 for the UI course and between weeks 4-5 for the SE course. Students in the criteria-based condition were not told of the specific weights used in the tool to organize them into teams.

After team assignments were announced, the newly formed teams had one week to complete the first team exercise. The second activity was due at the approximate midpoint of the project work, and again teams had one week to complete it. All teams submitted results for both activities.

In the final week of the course, students were asked to complete the summative peer evaluation in CATME, a written report detailing the contributions of each team member, a team satisfaction survey, and an essay reporting what lessons students learned about teamwork. Peer evaluations were required as part of the regular course instruction. We compensated survey and essay completions in the UI course with extra credit on the final exam. The SE instructor preferred not to offer extra credit, and survey completions in that course were compensated with $15.

At the end of the courses, a consent form was distributed to the students. Only the students (N=249 out of 304) who gave consent to use their data were included in our analysis.

4.4.1 Measures. The independent variables of our experiment were the two experimental factors: *Activity Focus* (with levels *team focus* and *task focus*) and *Composition* (with levels *criteria-based* and *random*). Our dependent variables were psychological safety, project grades, and measures of team experience.

Psychological safety. We assessed students' psychological safety using the CATME peer evaluation survey. This survey includes seven Likert items (each item uses a 7-pt scale) drawn from the validated instrument used in prior work [12]. Cronbach's alpha for this instrument for our student population was 0.70. We aggregated these seven items to produce a composite index for psychological safety, with 49 being the maximum score.

Project Grades. We assessed the performance of student teams using the grade they received on their project. This data was collected through regular course grading. The use of project grades as a dependent measure is consistent with prior work such as [5, 7, 27, 29].

Team experience. Students rated the experience they had with their project team by responding to statements represented as Likert items. The items measured satisfaction with the team ("How satisfied were you with your team assignment?", 1=Not satisfied, 5=Very satisfied, modified from [16]) and perceived performance ("My team produced a successful project outcome.", 1=Strongly Disagree, 7=Strongly Agree). Additional measures of team experience included team cohesiveness,

team conflict, and individual teamwork skills as rated by students' team members. These were collected using CATME's peer evaluation survey, which contains a series of questions covering these categories based on the team assessment literature [27]. Cronbach's alpha for the subdimensions of team cohesiveness (interpersonal cohesiveness, task commitment, and task attraction [8]) for our student population was 0.88, 0.69, and 0.82, respectively. For the subdimensions of team conflict (relationship conflict, task conflict, and process conflict [17]), alpha was 0.77, 0.74, and 0.78.

Lessons Learned Essay. Students were asked to write an essay addressing the following prompt: "Discuss what you have learned about teamwork in the course of interacting with your teammates." The only requirement specified was that students needed to demonstrate reasonable effort to receive the credit or compensation.

We performed qualitative analysis on the essays to understand student perceptions of their teams, and to determine whether there were differences in what students discussed between conditions. Because most students in the SE course gave very short essay responses or no response at all, we only considered the essays from the UI course in our qualitative analysis. Due to the scale of the data, we selected a random sample of 92 (approximately 55%) of the essays to analyze. We felt that this number was manageable, yet also representative of the full set of essays.

We partitioned these essays into idea units, where each idea unit represents a coherent unit of thought [39]. This partitioning resulted in 838 idea units. To categorize the idea units, we used the taxonomy of teamwork behaviors developed by Rousseau et al. [32]. Their work reviews 29 frameworks of teamwork behaviors and integrates them into a single taxonomy. We slightly modified this taxonomy by collapsing two subcategories (Collaborative Problem Solving and Team Practice Innovation) that were not clearly distinguishable in our data.

Two members of the research team discussed the application of the taxonomy and practiced categorizing small sets of idea units from the data set together until acceptable interrater reliability was reached. The two coders then split and individually categorized the remainder of the data set. Cohen's Kappa at the beginning and end of this process was 0.74 and 0.75, respectively, which indicates a high level of agreement. The full taxonomy and representative idea units are shown in Table 2. Note that students are anonymously identified with the string "S" + course number + three-digit student identifier.

5 RESULTS

To answer our research questions, we developed several linear mixed effect models to explain the outcome variables. Analysis was performed at the level of individual students nested within groups and classes, because an aggregate score for the group might not accurately represent the opinion of each individual within the group [35]. Unless otherwise noted, we constructed the models using the entire dataset (students from both the SE and UI courses). We considered class and team as random factors to account for the hierarchical structure of the data and because some variation in scores might result from the context of a particular group or class rather than our interventions. For any dependent variable following a normal distribution, we used the function "lmer" from the package "lme4" in R to define the model and fit it to our data. If the data failed the normality test, we defined a generalized linear mixed effect model using the function "glmer" from the same package to analyze the data. If a Wald Chi-Square Test on the fitted model revealed that any of formation, focus, or their interaction had a significant effect on the outcome variables, we performed post-hoc tests (Least Squares Means test with Bonferroni correction) to examine the differences between conditions. To determine whether we would see similar results in each of the individual courses, we also repeated our analysis for each course separately, removing the random effect for course.

Table 2. Correlation of psychological safety with our subjective measures of team experience. All but one of the measures had significant correlations. Adjusted significance threshold is p=0.01.

	Subcategory	Example
Regulation of team performance	Preparation of work accomplishment: team mission analysis, goal specification, and planning. (N=181)	*"[An imbalance of work] could have been avoided by discussing as a group what kind of organizational structure we liked best. Should there be one person acting as a sort of "boss" figure that has the final word on decisions? Or, should there be a vote? How should issues be raised? I don't think that this was something we ever discussed in detail, and I think this hurt our performance."* (S465104)
	Work assessment: performance and systems monitoring. (N=24)	*"Not everyone would admit their inabilities to do some types of works and it would [be] painful if we find out too late... We should really check progress early and swap jobs if some are too hard for some people."* (S465120)
	Task-related collaboration: coordination, cooperation, and information exchange. (N=246)	*"I've also learned how much more can be accomplished within a group than with just yourself. When we were stuck…we would sit down and throw around ideas and build on top of each other's ideas. That was definitely a rewarding experience because they would think of ideas I do not think I would have thought of by myself."* (S465134)
	Team adjustment: backing up behaviors, intrateam coaching, and team practice innovation. (N=58)	*"One of our team members was an online student. Initially we struggled to assign tasks and work effectively as a team. But then we created shared documents and assigned tasks to everyone. Therefore, by the end everyone made a significant contribution to the team."* (S465159)
Management of team maintenance	Psychological support. (N=114)	*"One of [my biggest weaknesses] is that I do not respect other peoples' ideas and want everybody work according to me…I thought I should change this habit since I did not want to offend anybody. It paid huge dividends, the project was better, the app was better and so was the team atmosphere. I gave respect and I received respect. I learnt how important being open to ideas and giving respect to their ideas and to them in turn is, how fruitful [it] is and how it makes the whole process go smoother."* (S465122)
	Integrative conflict management (N=27)	*"There should be no discrimination or hidden hostile undertones in the group. Every few weeks we should hold a session to release our thoughts on how the group is progressing so that we can get any project growth stunting emotions out of the way."* (S465124)

These results were consistent with those of the full model, so we limit our discussion here to the full model only.

We performed a power analysis using the package "pwr" to detect the probability of correctly rejecting a false null hypothesis in our Chi-Square tests given our sample size, significance level (0.05) and the effect size we seek. The analysis revealed that we could detect a medium effect size (r=0.30) with a probability of 0.99, although the probability of detecting a small effect size (r=0.10) is
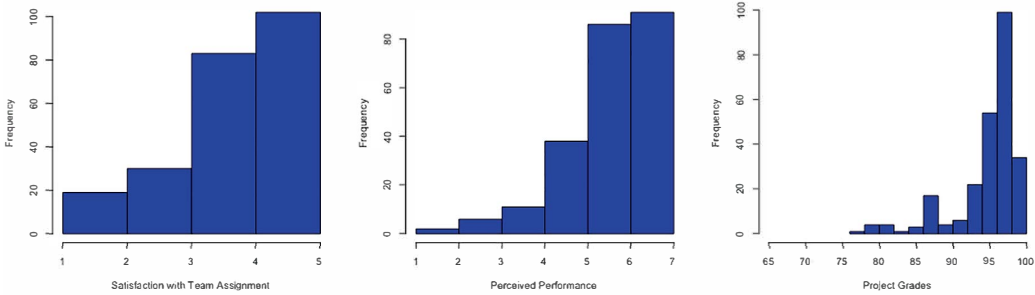
Fig. 2. Distributions of satisfaction (left) and perceived performance (middle) reported by individual students, and team project grades (right). For each of these measures, higher scores are more desirable.

Table 3. Effects of activity focus and team composition on our subjective measures of team experience.

	$\chi^2(3)$	p
Perceived performance	0.38	0.94
Satisfaction with team assignment	0.52	0.91
Peer evaluations	2.37	0.50
Cohesiveness	2.45	0.48
Conflict	2.84	0.42

lower (0.23). However, if a difference we are not able to detect exists between conditions, it would be too small to be of practical significance.

5.1 Effects of Composition and Activity Focus on Team Outcomes (RQ1)

Project grades (μ = 94.9 out of 100, s=4.5), students' satisfaction with their team (μ = 4.1 on a scale of 1 to 5, s=1.0), and their perceived performance (μ = 6.0 on a scale of 1 to 7, s=1.1) were high across all conditions. Distributions of these measures are shown in Figure 2.

The Wald test revealed no statistical effect of either team composition or activity focus on project grades (Wald $\chi^2(3)$ = 1.07, p=0.78). Similarly, activity focus and team composition did not affect perceived performance, satisfaction, or any other aspect of team experience. See Table 3.

These findings were surprising because prior work has reported that teams with specific compositions (e.g., teams with balanced genders) outperform teams lacking those compositions [47]. Possible explanations are described in the Discussion.

We examined the essays about teamwork written by students to provide more qualitative insights into how the activity focus affected team experiences and behaviors. For each student whose essay we coded, we counted the number of idea units pertaining to each of the two high-level categories of the taxonomy (Regulation of Team Performance and Management of Team Maintenance, which are task- and team-oriented behaviors, respectively). We then computed for each student the ratio of the idea units pertaining to each category to the student's total number of idea units. To determine whether students in certain activity conditions reported valuing behaviors that were more directed towards either of these two categories, we developed a model with this ratio as the dependent variable and activity focus as the independent variable. Team was included as a random effect. Recall that we only analyzed essays from the UI course, so we removed the random factor for class
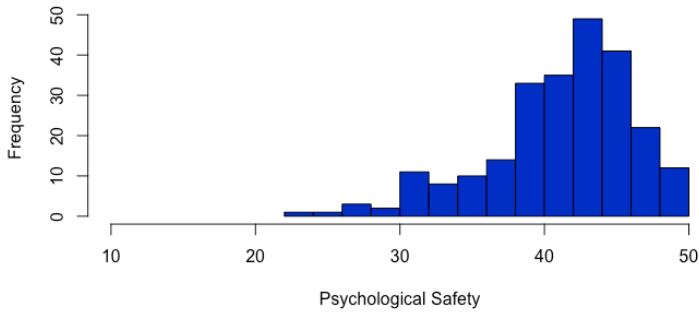
Fig. 3. Individual reports of psychological safety. Scores range from 7 to 49, with higher scores being more desirable.

that appeared in our other models. Wald tests revealed no significant effect of activity focus on the ratio of task-related (Wald $\chi^2(1)$ = 0.02, p=0.89) or team-related (Wald $\chi^2(1)$ = 0.01, p=0.94) idea units. However, we observed that the category students across all conditions reported learning the most about was Task-related Collaborative Behaviors (N=246 idea units). This category includes behaviors of coordination, cooperation, and information exchange, which are all important team skills [33].

5.2 Psychological Safety (RQ2 & RQ3)

The average psychological safety across all students was relatively high (μ = 41.70 on a scale of 7 to 49, s=5.02). The distribution of scores is shown in 3.

5.2.1 Effects of Composition and Activity. To assess whether our conditions affected psychological safety, we developed a model with psychological safety as the dependent variable and Composition and Activity Focus as the independent variables. Team and class were included as random effects.

Analysis revealed that psychological safety did not statistically differ between conditions, and no interaction effect was present (Wald $\chi^2(3)$ = 3.08, p=0.38). These results did not support our hypothesis that students performing the team focus activities would achieve higher psychological safety than those who performed the task focus activities. We provide several possible explanations for this in the Discussion.

5.2.2 Relation of Psychological Safety to Team Outcomes. To determine whether psychological safety is a predictor of the team outcome measures, we performed several linear regressions. We modelled each outcome variable with psychological safety as the predictor variable, and team and class as random effects. Because we performed several regressions, we adjusted the significance threshold for each to p=0.01.

The results indicated that psychological safety was not significantly related to project grades (β = 0.001, SE=0.02, p=0.94). However, it was correlated with all but one of our measures of team experience. See Table 4. This finding aligns with prior literature, which reports that psychological safety is related to satisfaction and perceived performance [3, 33].

Table 4. Correlation of psychological safety with our subjective measures of team experience. All but one of the measures had significant correlations. Adjusted significance threshold is p=0.01.

	β	SE	p
Perceived performance	0.37	0.15	0.01
Satisfaction with team assignment	0.43	0.13	<0.01
Peer evaluations	0.26	0.11	0.02
Cohesiveness	0.56	0.06	<0.01
Conflict	-0.38	0.07	<0.01

5.3 Additional Analysis

Because we examined the effects of team composition on team outcomes, it was important that the teams formed by the algorithmic tool had better compositions (based on the criteria defined by the instructors) than the teams formed using random assignment. Since randomly-assigned teams may have by chance achieved similar composition as teams assigned based on instructor-defined criteria, we examined the makeup of the teams and the "goodness" scores assigned to the teams by the tool based on the configurations set by the instructors. Note that we were only able to perform this analysis for the teams from the UI course, since the way the tool was configured for the SE course prevented us from obtaining scores for criteria other than Schedule for the teams in the "random" condition.

We found that the team formation tool produced better teams with respect to several criteria such as distribution of gender (t(34.18)=3.14, p<0.01), commitment levels (t(34.89)=3.45, p<0.01), and skillsets (t(33.18)=4.96, p<0.01) compared to the randomly assigned teams. For example, we found that some of the randomly assigned teams lacked the set of task skills the instructor deemed necessary for the projects, such as programming, communication, and design skills; whereas the teams formed using the criteria-based approach typically possessed all of these skills.

However, we did observe high variation in the goodness scores of the randomly assigned teams (criteria-based variance=7.21; random variance=53.09), which is to be expected given the nature of random assignment. For example, 3 of the random teams received scores comparable to those earned by teams formed with the criteria-based approach. This inconsistency in scores is one reason why criteria-based approaches are preferred over random assignment, as they are better able to create a level playing field for all students [16].

This finding raised questions about whether the results we observed would still hold true when comparing teams that were superior in makeup to those were not, regardless of how they had been formed. We therefore repartitioned our data so that the teams with low scores were in one condition and those with high scores were in another. This involved relabeling 4 teams, and was done according to a natural cutoff observable in the data. We repeated analyses for RQ1-3 described above with the new conditions, and found similar results.

6 DISCUSSION

In this experiment, we investigated how algorithmic team formation and team-building activities affect team performance, psychological safety, and satisfaction in student teams. The results of our analysis have important practical implications for instructors and tool designers, as well as implications for the theory of team composition and psychological safety.

6.1 Team Composition

Teams with specific compositional characteristics have been reported in prior work to outperform teams lacking those compositions (e.g., [7, 46]). However, we did not find any difference in performance or satisfaction between the randomly assigned teams and those formed using the criteria defined by the instructors in our study. This is surprising, considering that the instructors had selected criteria for the tool based on the team formation literature and their years of experience teaching these courses.

One possible explanation for this finding is an expectation effect created by the instructors' description of the algorithmic tool in the courses studied. Students were briefly informed of the potential benefits of forming teams using the team formation tool during a lecture early in the semester. Although half of the teams were formed randomly, students believed that all teams were created using the tool. Thus, because they may have believed their teams to be as capable as teams formed algorithmically, randomly assigned teams were able to achieve similarly high project scores and satisfaction. Two students wrote:

> "I would say the project experience I have with [UI Design] this semester is one of the best…I would attribute the formation of our group to the CATME survey we did at the beginning of the semester" (S465106, random condition).

> "…this was a more positive experience than other group projects I have been in. I do believe part of this is due to the CATME survey at the beginning of the project to select which students would work best together" (S465095, random condition).

This explanation suggests that it is important for instructors to fully describe to students the rationale for the criteria selected for team formation and why the criteria should lead to more capable teams. Future work could further determine how this expectation effect impacts student outcomes. For example, researchers could conduct an experiment in which teams formed using the same approach are told they were formed using different methods.

A second explanation for this finding is that the context of our experiment is different enough from the prior work that the previous results do not generalize to our experiment. For example, many studies showing that teams with specific compositions outperform teams without those compositions have leveraged in-lab team exercises (e.g., [46]) or simulated classrooms on micro-task platforms (e.g., [44]); whereas our work leveraged multi-week team projects in an authentic learning environment. For the prior work that has examined authentic courses, it may be the case that the different disciplines (e.g., computer science vs. civil engineering [7]), project lengths, or teaching cultures prevent generalization.

A third potential explanation for the lack of statistical effect between the team formation conditions is that team composition may not relate to performance when multiple criteria are stacked together, as they often are in practice. Prior literature has typically studied team composition using only one or two criteria at a time. For example, Brickell et al. [7] studied student teams in a context similar to our experiment (upper-level engineering students working on team design projects worth roughly 40% of their course grade) and found that criteria-based teams earned higher project grades compared to self-selected teams. However, teams were formed according to only two criteria (grouping dissimilar GPA and similar curricular interests, and vice versa). In our study, criteria-based teams were formed using 8 criteria in the UI Design course and 13 in the SE course, and therefore had more complex compositions than the teams studied in prior work. It is possible that since some criteria were assigned higher weights (greater importance) in the tool, the algorithm may have satisfied certain criteria for a team at the expense of other criteria.

Finally, it may be that the criteria-based team formation tool used in the study was ineffective. Indeed, we did identify aspects of the tool that did not match instructors' mental models, especially

related to how they should specify the criteria for team composition. For example, since the instructors in both courses wanted teams to have a set of specific skills, they specified in the tool's user interface to group students with dissimilar skills (e.g., course skills in the UI course was set to -5 in the tool, see Table 2). However, this caused the tool to give low scores to teams whose members had overlapping skills, even if all skillsets were present in the team. For example, the tool would prefer a two-person team composed of a designer and good communicator over a similar team where both members also reported knowing how to program, even though this team would be unable to complete the projects in these courses. However, we find this explanation less convincing than those mentioned previously, because we confirmed that criteria-based teams did have an advantage (e.g., in skillsets) over the randomly-assigned teams.

The criteria-based teams produced in our study, and consequently the results we observed, are shaped by the specific criteria selected for team formation by instructors and by the distribution of the student population with respect to these criteria. These complexities of real-world application of team formation tools may potentially limit the generalizability of our results. However, we believe it is important to study the experiences of actual student teams formed using instructor-defined criteria configurations, to complement work done in a more controlled lab setting. In addition, most of the criteria studied here are present by default in the tool, and are likely to be selected by other instructors. Even in the two courses of our study, nearly one third of the list of criteria was shared by both courses. Finally, the fact that both courses in our study showed similar results, despite using different criteria configurations and having different student populations, suggests that our results may also generalize to other course contexts.

Our work can serve as a stepping stone for additional research that further clarifies under what conditions teams with specific compositions outperform randomly assigned teams in courses, and conducts these studies with the multi-criteria configurations that more closely align with the configurations that instructors define in practice.

6.2 Psychological Safety

Students reported high levels of psychological safety in the study, but we did not observe a statistical difference between the conditions. We interpret these results in several ways.

First, it could be that both the team- and task-focused activities helped develop interpersonal relationships in teams. In their essays, two students wrote:

> "The next lesson I can take away from this project was the importance of getting to know my team members before starting the main project work. For our project we had team activity #1 which was brainstorming team names. This exercise helped us learn some introductory facts about each other, such as our favorite hobbies, TV shows, and movies…the fact that we got to know each other on a more personal level made us feel more invested in the project. I will certainly seek to take this approach when working with teams from now on" (S465105, task focus condition / criteria condition).

> "The first couple of weeks were spent doing team bonding exercises, and at the time, they almost seemed like a waste of time. It was not until we actually started brainstorming ideas for our project that I realized the significance of the first couple of weeks. Because we had gotten to get to know each other a little bit, every one [sic] was quite comfortable with others in the group. There was no fear of being judged for throwing out random ideas, regardless of how feasible they were" (S465062, team focus condition / random condition).

The interpretation that both sets of activities helped students is bolstered by our qualitative analysis, which showed that the number of idea units related to team maintenance behaviors such as

psychological support and conflict management was not statistically different between the activity conditions. Another possible explanation is that the nature of the courses prompted students to achieve high levels of psychological safety. Because students in both courses were unfamiliar with their project topics, they may have perceived their teammates as equally inexperienced as they were and felt more willing to ask for help without the fear of appearing incompetent. In addition, both courses required students to submit many team deliverables, which forced them to meet frequently with the course staff and with each other during the course. Psychological safety may have developed organically through this repeated interaction and masked the effects of the activities. Future work is needed to further tease apart the effects of different team-building activities on psychological safety.

A recent *New York Times* article describing industry research advocated for psychological safety, not team composition, as the primary factor increasing team performance and satisfaction in the workplace [11]. Our results support the observation from the industry study that team composition does not predict the quality of team outcomes and experiences, as well as results from studies in learning environments showing that psychological safety correlates with subjective measures like perceived performance (e.g., [3, 33]). However, our findings extend prior work by showing that psychological safety's role in influencing team performance in the workplace does not appear to transfer to student teams in learning environments.

One possible explanation is that the context of our study is different enough from the prior literature that the prior results do not generalize. Industry features different incentive structures from learning environments (e.g., compensation vs. grades), and the consequences of risk-taking in student teams are generally less grave than in work teams, where members could be demoted or dismissed if they are perceived as not possessing the required competencies. In addition, members of work teams may be more likely to know each other before the beginning of teamwork, as opposed to the students in the courses studied. Finally, team projects in industry such as those described in [12] can last for years, whereas ours were constrained to a 16-week semester.

Another possible explanation for the lack of correlation between psychological safety and performance is that although high psychological safety can promote learning behaviors such as seeking help and feedback [12], it may conceivably put team members at ease to display certain behaviors that negatively influence their performance. For instance, team members can choose to perform less than their share of work if they do not fear retribution from teammates. We identified multiple students who reported high psychological safety but received low peer evaluation scores from their teammates, who cited a lack of contribution and a disregard for the norms of the group. This phenomenon may explain why we did not find a correlation between psychological safety and peer evaluations, despite its correlation with our other subjective measures. Prior work has also reported manifestations of negative behavior such as cheating in teams with high psychological safety [29].

Future work is needed to further explore the relationship between psychological safety and team performance in learning environments.

6.3 Implications for Instructors and Tool Designers

We found that student teams formed by the tool studied did not outperform the randomly assigned teams in the context of our study. This result suggests that instructors should not use team formation tools as a "black box," and assume that specifying a certain composition will automatically increase team performance. Instructors should rather explain to students what choices were made in configuring the inputs to the tool and how these choices and the automated approach will benefit them. Our findings show that by explaining the purpose of a team formation tool, team members have reason to believe that they are capable of being a good team and will perform to meet that

expectation. In addition, team formation tools have other important benefits that warrant their use such as more efficient, fair, and consistent team formation experiences [16].

Although our results may appear to lend credit to those instructors who desire to form teams randomly, we caution those who would make this assumption because of the possible expectation effect in our study. It is unclear how students who were informed that they were grouped into teams randomly would perform in comparison to criteria-based teams. Future work should investigate this distinction.

Current team formation tools offer an expansive set of criteria and make it easy for instructors to create configurations that involve multiple criteria. Our results show that these multi-criteria configurations do not achieve stacked benefits, and do not yield the same performance benefits for teams as the more focused selections of one or two criteria studied in the literature. Tool designers should consider implementing configuration templates that allow instructors to use configurations that directly link to the team formation literature. Additionally, tool designers could allow the community to contribute their own configuration templates to a searchable knowledge base. These templates could allow instructors to share the team criteria configurations, learning goals, course makeups, and measured outcomes that other instructors could consult when deciding how to compose teams in their own courses.

Our results also suggest that instructors could introduce either team-building activity described in this paper to help offset the initial unfamiliarity students experience when teams are formed algorithmically. However, instructors should not expect the activities to boost psychological safety or team performance.

7 LIMITATIONS AND FUTURE WORK

A limitation of our study is that we did not have a condition for the activity focus factor in which students did not complete any team-building activity. The inclusion of this condition would have allowed us to ascertain whether both sets of activities helped develop psychological safety in teams, or neither did. We chose not to include it in this experiment because adding an additional condition would have reduced the statistical power given the number of conditions already being tested, and we deemed differentiating the focus of the activities more important. Future studies could include a condition of this nature.

It would also be beneficial to measure how team performance, satisfaction, and psychological safety evolve over time as the members of a team get to know each other. This could be accomplished by collecting data for these measures at multiple points in time, rather than only at the end of the project work. These additional measurements could serve as a manipulation check, for example to help confirm whether the team-focus activities produced their intended effects (i.e., increasing psychological safety).

Our study was conducted in two courses taught within a specific university context. Future work could test whether our findings generalize to other courses and institutions with different teaching cultures, as well as courses using different combinations of criteria for team formation and those having different student populations (e.g., more junior students with less prior teamwork experience and domain knowledge). Future work could also test additional compositional strategies, such as those that include student preferences [22, 23].

Finally, since many students reported the maximum score for our team satisfaction and perceived performance measures, it is possible that a ceiling effect may have contributed to the lack of statistical effect. Future work could use different instruments with a larger scale to capture these measures and further test the generalizability of our results.

8 CONCLUSION

This work reported the results of an experiment comparing the effects of team formation strategies and team-building activities on team performance, satisfaction, and psychological safety in two university courses. Teams formed to achieve the composition defined by the instructor in each course did not show a statistical difference on any measure relative to teams formed randomly. This result suggests that instructors should be less concerned with fine-tuning the configuration in the tool, and should instead emphasize the value of the criteria-based approach, so that students perceive teams to be equally capable of performing well.

We also found that results from industry teams (e.g., [11]) regarding psychological safety's link to performance did not generalize to the student teams studied. This finding suggests that though activities aimed at fostering psychological safety may have practical benefits for student teams, instructors should not expect these activities to boost performance or psychological safety. More work is needed to further clarify what kinds of activities can improve psychological safety and in what contexts psychological safety impacts team performance.

REFERENCES

[1] 2017. K Framework Main Page. Retrieved 31 August 2018 from http://www.kframework.org/index.php/Main_Page
[2] 2018. CATME Smarter Teamwork. Retrieved 31 August 2018 from http://info.catme.org/
[3] S Adams, L Simon, and B Ruiz. 2002. A Pilot Study of the Performance of Student Teams In Engineering Education. In *2002 Annual Conference & Exposition, Montreal, Canada.*
[4] Julia B Bear and Anita Williams Woolley. 2011. The role of gender in team collaboration and performance. *Interdisciplinary science reviews* 36, 2 (2011), 146–153.
[5] Lisa Bender, Gursimran Walia, Krishna Kambhampaty, Kendall E Nygard, and Travis E Nygard. 2012. Social sensitivity and classroom team projects: an empirical investigation. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*. ACM, 403–408.
[6] Bret H Bradley, Bennett E Postlethwaite, Anthony C Klotz, Maria R Hamdani, and Kenneth G Brown. 2012. Reaping the benefits of task conflict in teams: The critical role of team psychological safety climate. *Journal of Applied Psychology* 97, 1 (2012), 151.
[7] Lt Col James L Brickell, Lt Col David B Porter, Lt Col Michael F Reynolds, and Capt Richard D Cosgrove. 1994. Assigning students to groups for engineering design projects: A comparison of five methods. *Journal of Engineering Education* 83, 3 (1994), 259–262.
[8] Sally A Carless and Caroline De Paola. 2000. The measurement of cohesion in work teams. *Small group research* 31, 1 (2000), 71–88.
[9] Abraham Carmeli, Daphna Brueller, and Jane E Dutton. 2009. Learning behaviours in the workplace: The role of high-quality interpersonal relationships and psychological safety. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research* 26, 1 (2009), 81–98.
[10] Abraham Carmeli and Jody Hoffer Gittell. 2009. High-quality relationships, psychological safety, and learning from failures in work organizations. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 30, 6 (2009), 709–729.
[11] Charles Duhigg. 2016. What Google learned from its quest to build the perfect team. *The New York Times Magazine* 26 (2016), 2016.
[12] Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly* 44, 2 (1999), 350–383.
[13] Donald G Gardner, Randall B Dunham, LL Cummings, and Jon L Pierce. 1987. Focus of Attention at Work and Leader-Follower Relationships. *Journal of Organizational Behavior* 8, 4 (1987), 277–294.

[14] J Richard Hackman. 1990. *Groups that work and those that don't.* Number E10 H123. Jossey-Bass.

[15] Sujin K Horwitz and Irwin B Horwitz. 2007. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management* 33, 6 (2007), 987–1015.

[16] Farnaz Jahanbakhsh, Wai-Tat Fu, Karrie Karahalios, Darko Marinov, and Brian Bailey. 2017. You Want Me to Work with Who?: Stakeholder Perceptions of Automated Team Formation in Project-based Courses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 3201–3212.

[17] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly* 44, 4 (1999), 741–763.

[18] Tony Karahalios and Karrie Karahalios. 2009. Social mirrors as social signals: transforming audio into graphics. *IEEE computer graphics and applications* 29, 5 (2009), 22–32.

[19] Cameron Klein, Deborah DiazGranados, Eduardo Salas, Huy Le, C Shawn Burke, Rebecca Lyons, and Gerald F Goodwin. 2009. Does team building work? *Small Group Research* 40, 2 (2009), 181–222.

[20] Richard A Layton, Misty L Loughry, Matthew W Ohland, and George D Ricco. 2010. Design and Validation of a Web-Based System for Assigning Members to Teams Using Instructor-Specified Criteria. *Advances in Engineering Education* 2, 1 (2010), n1.

[21] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P Dow. 2016. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 260–273.

[22] Ioanna Lykourentzou, Robert E Kraut, and Steven P Dow. 2017. Team Dating Leads to Better Online Ad Hoc Collaborations.. In *CSCW.* 2330–2343.

[23] Ioanna Lykourentzou, Shannon Wang, Robert E Kraut, and Steven P Dow. 2016. Team dating: A self-organized team formation strategy for collaborative crowdsourcing. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 1243–1249.

[24] Jennifer Marlow and Laura Dabbish. 2011. Photo sharing in diverse distributed teams. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work.* ACM, 317–320.

[25] Jennifer Marlow and Laura Dabbish. 2012. Designing interventions to reduce psychological distance in globally distributed teams. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion.* ACM, 163–166.

[26] Barbara Oakley, Richard M Felder, Rebecca Brent, and Imad Elhajj. 2004. Turning student groups into effective teams. *Journal of student centered learning* 2, 1 (2004), 9–34.

[27] 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.

[28] Jennifer R Overbeck and Bernadette Park. 2001. When power does not corrupt: superior individuation processes among powerful perceivers. *Journal of personality and social psychology* 81, 4 (2001), 549.

[29] Matthew J Pearsall and Aleksander PJ Ellis. 2011. Thick as thieves: the effects of ethical orientation and psychological safety on unethical team behavior. *Journal of Applied Psychology* 96, 2 (2011), 401.

[30] Lakshmi Ramarajan, Sigal G Barsade, and Orah R Burack. 2008. The influence of organizational respect on emotional exhaustion in the human services. *The Journal of Positive Psychology* 3, 1 (2008), 4–18.

[31] Marcel M Robles. 2012. Executive perceptions of the top 10 soft skills needed in todayâĂŹs workplace. *Business Communication Quarterly* 75, 4 (2012), 453–465.

[32] Vincent Rousseau, Caroline Aubé, and André Savoie. 2006. Teamwork behaviors: A review and an integration of frameworks. *Small group research* 37, 5 (2006), 540–570.

[33] Bianey C Ruiz Ulloa and Stephanie G Adams. 2004. Attitude toward teamwork and effective teaming. *Team Performance Management: An International Journal* 10, 7/8 (2004), 145–151.

[34] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2016. Huddler: Convening stable and familiar crowd teams despite unpredictable availability. *arXiv preprint arXiv:1610.08216* (2016).

[35] Jeroen Schepers, Ad de Jong, Martin Wetzels, and Ko de Ruyter. 2008. Psychological safety and social support in groupware adoption: A multi-level assessment in education. *Computers & Education* 51, 2 (2008), 757–775.

[36] Debra Smarkusky, Richard Dempsey, J Ludka, and Frouke de Quillettes. 2005. Enhancing team knowledge: instruction vs. experience. In *ACM SIGCSE Bulletin*, Vol. 37. ACM, 460–464.

[37] Neville Anthony Stanton, Alan Hedge, Karel Brookhuis, Eduardo Salas, and Hal W Hendrick. 2004. *Handbook of human factors and ergonomics methods.* CRC press.

[38] John Paul Stephens, Emily Heaphy, and Jane E Dutton. 2011. High quality connections. *The Oxford handbook of positive organizational scholarship* (2011), 385–399.

[39] Anselm L Strauss. 1987. *Qualitative analysis for social scientists.* Cambridge University Press.

[40] Scott I Tannenbaum, Rebecca L Beard, and Eduardo Salas. 1992. Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. In *Advances in psychology*. Vol. 82. Elsevier, 117–153.

[41] Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 459–468.

[42] Jennifer Thom-Santelli and David R Millen. 2009. Learning by seeing: photo viewing in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2081–2090.

[43] Armin Weinberger, Bernhard Ertl, Frank Fischer, and Heinz Mandl. 2005. Epistemic and social scripts in computer–supported collaborative learning. *Instructional Science* 33, 1 (2005), 1–30.

[44] Miaomiao Wen, Keith Maki, Xu Wang, Steven Dow, James D Herbsleb, and Carolyn Penstein Rosé. 2016. Transactivity as a Predictor of Future Collaborative Knowledge Integration in Team-Based Learning in Online Courses.. In *EDM*. 533–538.

[45] Anita Williams Woolley. 1998. Effects of intervention content and timing on group task performance. *The Journal of Applied Behavioral Science* 34, 1 (1998), 30–46.

[46] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[47] Anita Williams Woolley and Nada Hashmi. 2013. Cultivating collective intelligence in online groups. In *Handbook of Human Computation*. Springer, 703–714.

[48] Zhilin Zheng, Tim Vogelsang, and Niels Pinkwart. 2015. The impact of small learning group composition on student engagement and success in a MOOC. In *Proceedings of the 8th International Conference of Educational Data Mining*. 500–503.