LIFT: Integrating Stakeholder Voices into Algorithmic Team Formation

Emily M. Hastings[*], **Albatool Alamri**[†*], **Andrew Kuznetsov**[§*], **Christine Pisarczyk**[¶*],
Karrie Karahalios[*], **Darko Marinov**[*], **Brian P. Bailey**[*]
University of Illinois at Urbana-Champaign[*], University of Jeddah[†],
Carnegie Mellon University[§], University of Maryland - College Park[¶]
[ehstngs2, kkarahal, marinov, bpbailey]@illinois.edu,
asalamri@uj.edu.sa, kuz@cmu.edu, psrczyk2@umd.edu

ABSTRACT

Team formation tools assume instructors should configure the criteria for creating teams, precluding students from participating in a process affecting their learning experience. We propose LIFT, a novel learner-centered workflow where students propose, vote for, and weigh the criteria used as inputs to the team formation algorithm. We conducted an experiment (N=289) comparing LIFT to the usual instructor-led process, and interviewed participants to evaluate their perceptions of LIFT and its outcomes. Learners proposed novel criteria not included in existing algorithmic tools, such as organizational style. They avoided criteria like gender and GPA that instructors frequently select, and preferred those promoting efficient collaboration. LIFT led to team outcomes comparable to those achieved by the instructor-led approach, and teams valued having control of the team formation process. We provide instructors and designers with a workflow and evidence supporting giving learners control of the algorithmic process used for grouping them into teams.

Author Keywords

Algorithms; CATME; Learnersourcing; Crowdsourcing; Learning; Team formation; Team composition.

CCS Concepts

•**Human-centered computing → Empirical studies in HCI; Computer supported cooperative work;** *Empirical studies in collaborative and social computing;*

INTRODUCTION

Instructors are increasingly utilizing algorithmic team formation tools such as CATME [37] to group students into teams in their courses for team-based learning. These tools are grounded in the literature on criteria-based team formation and enable instructors to group students into teams using criteria such as skills, work habits, and demographics.

Team formation tools make a critical assumption that the instructor should configure the criteria inputs to the team formation algorithm. Instructors can decide these inputs by considering the course learning goals, their prior teaching experience, and the literature. However, this assumption leaves those with the largest stake in the process, the *students*, with little to no opportunity to configure an algorithm that will affect their team experience, learning, and grades [29].

In this paper, we introduce and empirically investigate a novel learner-centered workflow for identifying team formation criteria that are meaningful to students and configuring these criteria in a team formation tool. In the LIFT (Learner Involvement in Forming Teams) workflow, students engage in an online activity to propose and discuss team formation criteria that they find meaningful. They then vote on whether they think each proposed criterion should be included in the tool. Finally, the students collectively provide a weight for each selected criterion. This approach is grounded in theories of crowdsourcing and collective intelligence, and inspired by prior successes of the use of crowdsourcing techniques in learning environments (e.g., [33, 36, 61, 8, 46, 35, 49, 60, 23, 24, 39, 63]).

Giving students more agency in algorithmic team formation has many potential benefits. Students have independent, localized knowledge of what makes a good team based on their own prior experiences, of which the instructor may not be aware. Involving students can also prevent them from viewing the team formation tool as a "black box," which can lead to suspicions of favoritism and distrust of the instructor [11]. Research in algorithm transparency shows that increasing user knowledge of and control over algorithmic processes can increase satisfaction [57] and improve trust and acceptance of these systems [19, 34, 38]. Finally, increased control over team formation can prompt students to take greater ownership of group problems [44], which can aid in setting goals, solving problems, and creating high-quality work [17, 15].

Given the prior work and these potential benefits, we hypothesized that (1) students using LIFT would be capable of proposing sensible criteria configurations that represent their collective preferences, and that (2) they would have team outcomes at least as positive as those achieved by students using the traditional instructor-led process. To test these hypotheses, we conducted a mixed-methods experiment in five university

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376797

courses leveraging team-based learning (N=289 students). We compared LIFT to the instructor-led process in terms of student team performance, satisfaction with the team assignment, and satisfaction with the team formation process, among other measures. We also conducted semi-structured interviews with 18 students and the 6 instructors of the courses to evaluate their perceptions of LIFT and, for the instructors, what they learned from the student criteria selections.

Our first hypothesis was supported. Students proposed novel criteria including personal organization style and confidence in programming skills. In general, they favored criteria related to skills, logistics, and other factors contributing to completing their project more efficiently. Interestingly, most students voted *against* or disregarded criteria frequently used by instructors, including gender, race, and GPA. This finding is surprising because these criteria are commonly used and are supported by existing studies of team composition (e.g., [9, 13, 28, 31]). Our second hypothesis was partially supported: LIFT led to team outcomes comparable to (but not significantly better than) those achieved by the instructor-led approach, despite the differences in the configurations. Students valued having a voice in configuring the team formation tool (Median=6.0 on a scale from 1 to 7, 7=most preferred). Finally, we found that LIFT gave instructors insight into creating criteria configurations that are more responsive to student preferences.

Our work makes three contributions to the HCI community. First, we provide deeper empirical understanding of the effectiveness of leveraging learners' collective choices to shape the algorithmic team formation process. Second, we describe a learner-centered workflow instructors can deploy to tap into the criteria that matter most to students in their specific courses. While we focus on deploying LIFT in face-to-face classrooms, the proposed team formation workflow could generalize to online learning environments, since it primarily leverages online tools. Extensions to other contexts, such as online labor markets and open design challenges, are also possible, but may require an initial pooling phase for workers who will eventually be formed into teams. Finally, we share implications for how designers of team formation tools can give stakeholders more control over the algorithmic team formation process. For example, tools might provide instructors with graphical representations of students' collective votes for the weight of each criterion in the configuration interface, in order to help them create configurations responsive to student preferences.

RELATED WORK

We ground our work in the prior literature on team formation methods and team composition. We also explain how our work contributes to the developing literature on learnersourcing.

Algorithmic Team Formation

There is growing support in the literature and in educational practice for the use of a criteria-based approach to team formation. The approach offers benefits such as providing a team formation experience perceived as fair and removing the stress of having to form a team on one's own [29]. In this approach, instructors group students into teams by considering how criteria such as skills, work habits, and demographics should factor into the team formation process. Algorithmic team formation tools (e.g., [37, 59, 27]) are increasingly being deployed to implement criteria-based team formation processes and to help instructors keep pace with growing course enrollments.

Researchers have studied how different team formation criteria affect team outcomes. For example, Bear, Woolley, et al. found that including more women in a team raises the team's collective intelligence [9, 64]. Lykourentzou et al. show that team performance and satisfaction can be increased by balancing personality types within a team [40]. Team performance can also increase through including diverse skills relevant to the project [28], by including nationality diversity [7], by grouping according to academic ability and curricular interests [13] or pairwise transactivity in a discussion [62], or by modifying team membership according to tie strength [53]. Connerley and Mael began to identify which factors students find the most important, but did not test how forming teams according to these criteria affects team outcomes [18].

Our work contributes to this literature by reporting the criteria that students prefer for team formation, how these criteria compare to instructor preferences, and how these criteria choices impact team satisfaction and performance. We also contribute a novel workflow that can be deployed to give stakeholders control of the algorithmic process used to group them into teams. Our work is timely because many instructors, especially in engineering disciplines, are implementing algorithmic approaches in their courses due to growing enrollments and increased diversity, and it is unclear how incorporating new mechanisms like increasing student control could affect the deployment of this approach in authentic learning environments.

Other Team Formation Approaches

Self-selection and random assignment are common team formation approaches, especially since they are easy to implement. These methods can promote positive team experiences; for example, self-selection can increase satisfaction [16] and encourage group members to take ownership over group interactions and conflicts [44], which in turn can help students set goals, solve complex problems, and create high-quality work [17, 15]. However, algorithmic team formation has become more popular in part because it addresses the weaknesses of such approaches. For instance, self-selection may leave some students unable to find a team to join [21], and random assignment has been shown to lead to reduced team satisfaction [16]. In addition, both strategies often produce teams that lack the needed skill variety to accomplish course tasks [30, 16].

Our contribution is to further strengthen algorithmic team formation by incorporating strengths of these approaches, like giving stakeholders increased ownership of their work.

Learnersourcing

In crowdsourcing, complex work is decomposed into granular tasks and outsourced to a number of people who individually perform those tasks. The resulting partial solutions are then aggregated to complete the work [56]. Crowdsourcing is increasingly being applied in learning environments, where learners can serve as the crowd. Learnersourcing has been defined as crowdsourcing “in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves” [33]. For example, researchers have used learnersourcing to provide problem solving advice [23, 24, 39, 63] and to generate design feedback [49, 60], among other applications (e.g., [61, 8, 45, 35, 36]).

Our work builds upon these and similar successes and contributes a learnersourcing workflow for controlling the inputs to a team formation tool. By taking part in this process, students are both contributing novel content (criteria, weights, and rationales for these choices) which can be used to form teams in their own and future courses, as well as learning more about each other’s perspectives on how to form a good team.

THE TEAM FORMATION TOOL

The team formation tool we used in this study is the Comprehensive Assessment for Team-Member Effectiveness (CATME), which is a representative criteria-based tool [2]. We chose this tool because it is used in many courses at our university and is grounded in the team composition literature [37].

In the typical CATME workflow, the instructor chooses from a set of 27 predefined criteria or defines their own based on learning goals for the course and the team composition literature. The tool creates a survey with questions related to the selected criteria and distributes it to students via email. The instructor can then review the responses and configure the weights for each criterion. Weights range from -5 to 5, where negative weights indicate that students who have dissimilar responses for the associated attribute should be grouped together, and positive weights indicate that similar students should be grouped. The magnitude reflects the criterion’s impact relative to the other criteria in the configuration. For example, assigning "Schedule" a weight of 5 strongly prefers groups where students report similar schedules. The tool then forms teams based on these weights using a greedy randomized algorithm [1], and instructors can either accept the generated teams or rerun the algorithm to produce potentially different results. Finally, the tool notifies students of their team assignments and provides them with their teammates’ contact information.

THE LIFT WORKFLOW

Our proposed workflow consists of three stages. First, students engage in an online discussion prior to teams being formed, in which they discuss which formation criteria they think should be used in the course. Second, students vote on which of the proposed criteria should be included in the team formation tool. Third, each student selects their preferred weights for the criteria in the tool, and these selections are averaged to create a configuration for the entire class. The goal of the workflow is to learn what criteria students find important to consider when forming teams, and to use this knowledge to increase students’ sense of agency over the team formation process.

Criteria are proposed through a discussion to elicit rich information about students’ preferences on team formation. We believed surveying individuals would not be as effective, since students would not be aware of their classmates’ contributions, and would be unable to react to them. Thus, they would likely generate repetitive criteria, and their responses would not provide as much insight as a dialogue between students would. However, we deemed the option of anonymity necessary for students to be comfortable enough to propose and discuss criteria that may be sensitive, such as race and gender. This decision is based on prior work showing that anonymity can promote increased and more egalitarian participation [32, 55], aid idea generation [22], and reduce status differences [12].

The voting stage takes place after the discussion is finished rather than continuously (e.g., “upvoting” posts as they are made). This choice allows students to view all of the criteria that were proposed along with the associated conversations, and form their own opinions prior to voting, which facilitates reaching consensus [46]. These votes are collected individually through a survey in order to prevent students’ responses being influenced by seeing those of the majority [47].

Configuration of the team formation tool occurs when students provide their information in the survey. As students give their responses, they also specify the magnitude and sign of the weights for each selected criterion. These individual preferences are aggregated to produce the final configuration used for the whole class. Students provide weights at this stage because it is necessary for them to have seen which criteria were ultimately selected before trying to rank their importance.

RESEARCH QUESTIONS

This work addresses the following research questions:

RQ1: What team formation criteria do students select when given the chance, and how much agreement is there among students? How do student and instructor choices differ?

RQ2: How do students perceive their agency when they are allowed to have input into the team formation process?

RQ3: How does allowing students to select criteria affect their team performance, satisfaction, and other course experiences compared to having instructors select criteria?

RQ4: How do instructors perceive transferring agency in the team formation process to students, and what do they learn about student preferences?

Answering these questions will provide empirical knowledge of student preferences about team formation tools and how they relate to choices instructors make in practice, and help researchers, tool designers, and instructors develop and deploy tools that more closely consider student voices.

METHOD

To answer our research questions, we conducted a mixed-methods between-participants experiment examining the effects of one factor, *Criteria Selector* (Instructor vs. Learner), on team outcomes. The experiment was conducted in parallel in five project-based courses at a large public university. The study was approved by the IRB at our university.

Participants and Courses

Five university courses leveraging team-based learning were involved in our study: four engineering courses (Software Engineering I, Design for Manufacturability, Mechanical Design

II, Introduction to Statics) and one art course (Design Methods). See Table 1. In each course, approximately half of the teams were in each condition. There was little student overlap between courses. 289 of the 936 total students enrolled in these courses consented to participate in the experiment. With the exception of the Statics course, in which students completed weekly team assignments rather than a single large project, the projects for each course required students to submit multiple deliverables throughout the semester, including proposals, prototypes, and final demonstrations and reports.

Criteria Selection

To determine which configurations students and instructors select, as well as how each perceive their agency in the team formation process, we utilized two methods to select and weight the criteria used in the team formation tool. In one version (LIFT), the configuration of the tool was crowdsourced to students, who discussed and voted on which criteria should be used as input to the tool (Learner condition). In the other version, acting as a control condition, the instructor configured the criteria, as in the traditional workflow (Instructor condition). Students were randomly assigned to one of the conditions. In courses divided into sections, we randomly assigned entire sections to a condition, in order to minimize the possibility of students becoming aware of the different conditions.

Learner Condition

Following the LIFT workflow, students in the Learner condition discussed which formation criteria they thought should be used in the course. We held this discussion on Piazza [3], an educational platform used in prior work on educational crowdsourcing [51], which provided a discussion environment restricted only to students in the course and the option of anonymity. Students were provided with a short description of the team formation tool and how it is configured, as well as a list of the default criteria available in the tool. They were asked to make at least three contributions to the discussion, where a contribution was either (a) a post identifying a criterion and explaining its importance for the course, or (b) a follow-up comment on another student's post discussing (dis)advantages of the criterion or suggesting enhancements. Students were told that criteria should be relevant to the course and come from the provided list or their own experiences and ideas. Students were able to post contributions that were anonymous to their peers, but not to the researchers (in order to track participation and discourage undesirable behavior).

After the activity, the research team examined the discussion and compiled a list for each course of all the criteria proposed by students, discarding duplicates and those few that would be infeasible to implement in the team formation tool. Those discarded include criteria with excessive answer choices (e.g., "Which student organizations are you part of?"), those that were ill-defined (e.g., "Equality"), and those that went against the spirit of criteria-based team formation (e.g., "Choosing own teammates"). A survey was prepared with the remaining criteria, which asked students to respond to the statement, "This criterion should be included in CATME" for each of the criteria using a 5-point Likert item (-2= Strongly disagree, 2=Strongly agree). Student responses were summed to create a score for each criterion that reflected the degree of support it received, and criteria were ranked from most to least popular.

Because students proposed many more criteria than are typically used in the tool, we considered two different selection thresholds for which of these criteria were actually included, in order to examine how including different numbers of criteria can impact outcomes. In the first approach (Learner-all), all criteria that had a total score above 0 (meaning they had more positive votes than negative votes) were included. For the other (Learner-strict), only the upper quartile (top 25%) of criteria receiving scores above 0 were included. Each course used only one threshold: Design for Manufacturability, Mechanical Design II, and Design Methods used Learner-all, while Software Engineering I and Introduction to Statics used Learner-strict.

Once students voted on the criteria, the team formation tool was configured according to student preferences. The final weight used in the system for each criterion was the floor of the mean of student weights (since weights cannot be fractional), with the sign that received the most support.

Instructor Condition

To maintain a consistent workload between conditions, students in the Instructor condition also performed a discussion activity prior to teams being formed. In this activity, they discussed their previous team experiences, or if they had none, what they expected to achieve working on a team in the course. While students in the other condition were voting on criteria, students in this condition completed a short survey asking them to describe their greatest takeaway from the discussion.

After the discussion activity, each instructor configured the criteria and weights in the team formation tool according to their own choices, as in the traditional workflow. These configurations were based on the course's learning goals and project requirements, instructors' prior experiences with teams in the course, and the team formation literature. The tool then distributed the team formation survey for students to complete.

All activities were performed in parallel between the two conditions. Students were aware that different versions of the discussion activity existed, but were not told the specific assignments other than their own. Additionally, they were not told that part of the class had been able to select their own formation criteria and weights while the rest had not.

Procedure

Students had one week at the beginning of the semester to participate in the online discussion and voting activity, after which the research team constructed the team formation survey in the tool. Students then had one week to complete this survey. Those who did not respond were placed onto a team randomly. At the end of the course, students were asked to complete a peer evaluation in the tool and a survey regarding their satisfaction with their team and the team formation process used in the course. A consent form was also distributed.

The study activities were required as part of regular course instruction or compensated with extra credit, depending on the course. See Table 2 for the response rate for these activities for students (N=289 of 936) who gave consent.

Table 1. Information about the courses involved in the study. For Statics and Design Methods, we list the number of female students as "N+" because this information was not available in the course rosters, but at least N students responded that they identify as female in our surveys.

Course	Students (Female)	Typical Level	Team Size	Teams	Project Length	% of Grade
Software Engr.	130 (12)	Senior-Grad	6-8	18	7 weeks	40%
Design for Manf.	148 (38)	Soph-Junior	4-6	30	13 weeks	25%
Mech. Design	59 (10)	Senior	4-5	16	7 weeks	35%
Statics	559 (33+)	Soph-Junior	2-4	154	Weekly	8%
Design Methods	40 (14+)	Junior	2-3	14	5 weeks (x2)	80%

Table 2. Response rates of consenting students for the study activities.

	Soft. Engr.	Dsgn. Manf.	Mech. Dsgn.	Stat.	Dsgn. Meth.
Discussion activity	70%	97%	88%	80%	97%
Formation survey	96%	92%	100%	89%	76%
Peer evaluation	75%	97%	100%	88%	32%
Post-survey	47%	85%	94%	98%	32%

Table 3. Breakdown of interview participants.

	Soft. Engr.	Dsgn. Manf.	Mech. Dsgn.	Stat.	Dsgn. Meth.
Learner condition	-	2	1	2	1
Instructor condition	2	3	3	2	2

We recruited N=18 students through an open email call to take part in semi-structured interviews for more detailed feedback regarding their experiences. All courses and conditions were represented in this sample; see Table 3. The questions in these interviews focused on student perceptions of the criteria chosen, as well as strengths and weaknesses of both LIFT and the instructor-led approach. We also interviewed the six instructors of the courses studied. Questions focused on their previous experience with the tool, their perceptions of the criteria chosen by students, and strengths and weaknesses of both approaches [1]. Interview participants completed an additional consent form and were compensated $10 for their time. Interviews lasted from 20-40 minutes and were audio-recorded and transcribed by the research team. Two researchers individually analyzed the interview data and iteratively formed conceptual categories and grouped statements into them. They then discussed each others' outcomes and iterated on the categories and grouping of the data until consensus was reached [52, 25].

Measures

The independent variable in our experiment was our experimental factor, Criteria Selector (with levels Learner and Instructor). Our dependent variables were project grades and measures of satisfaction, agency, and perceived learning.

Project Grades

We assessed student team performance by using the grade the team received on their project. This data was collected as a part of regular course instruction. The use of project grades is consistent with prior work on team outcomes [10, 13, 48, 50].

Team and Process Satisfaction

Students rated their team satisfaction by agreeing or disagreeing with two statements represented as 7-point Likert items (1=Strongly Disagree, 7=Strongly Agree). Statements focused on satisfaction with the team ("I was satisfied with the team assigned to me.") and perceived performance ("My team produced a successful project outcome."). Cronbach's alpha for a scale consisting of these two measures was 0.86.

Students rated their experience with the team formation process by agreeing or disagreeing with statements about their satisfaction with the approach ("What has been your experience with the approach used in this course?", 1=Very poor, 7=Excellent) and recommendation to repeat the approach ("I recommend repeating the approach to team formation I experienced in this course in the future."). Alpha for a scale consisting of these two measures was 0.82. Our satisfaction measures are consistent with prior work [4, 29, 26].

Agency and Perceived Learning

Students also reported their perceived agency ("I felt I had a voice in shaping how teams were formed in this class."), the importance of having input ("I believe it is important to have input into what information (which criteria) are considered when matching me with teammates in this class."), and their perceived learning about teamwork ("After this experience, I learned what makes an effective team.").

RESULTS

To answer our quantitative research questions, we developed a linear mixed effect model to explain each outcome variable. We considered course and team as random factors to account for the hierarchical structure of the data (students nested within courses and teams) and because some variation in scores might result from the context of a particular group or course rather than our conditions. Since none of our dependent variables follow a normal distribution, we used the function "glmer" from the package "lme4" in R to define the model and fit it to our data. We then used a Wald Chi-Square Test on the fitted model to determine whether the criteria selector had a significant effect on the outcome variables. Because we performed several regressions, we used a Bonferroni-adjusted significance threshold of p=0.05/8=0.006. In order to account for potential effects due to section groupings, we also performed our analyses including a random factor for section. There were no differences in the results using this model, however, so we present only the simpler model using course and team here.

We have limited our quantitative statistical analysis to three of the five courses (Software Engineering, Mechanical De-

[1] We have included the interview scripts as supplemental material.

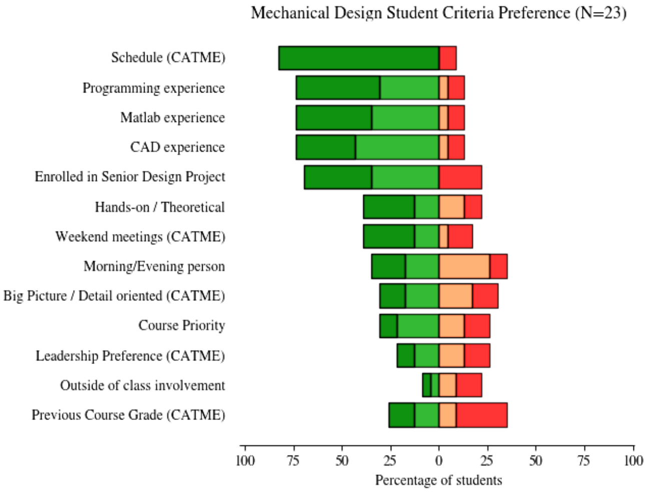

Figure 1. The distribution of votes for criteria discussed in the Mechanical Design course. For each criterion, the colored bars represent from left to right votes for "Strongly Agree", "Agree", "Disagree", and "Strongly Disagree" that the criterion should be included in the tool. For example, Schedule received strong support, while Big Picture/Detail-Oriented received roughly equal amounts of positive and negative votes.

sign, and Design for Manufacturability, N=132) because the group component in Statics was a relatively minor part of the course with no final project score, and in Design Methods some students changed teams after completing the first of the two course projects. We complement these statistical results with qualitative support from our interviews and the online discussion, for which we have data from all five courses. Note that participants are anonymously identified with the string "S" (student) or "I" (instructor) + a numerical identifier. The notation "S=n" indicates that n students gave similar responses.

We performed a power analysis with the R package "pwr" to assess our ability to correctly reject a false null hypothesis in our Chi-Square tests. The analysis revealed that we could detect a medium effect size (r=0.30) with a probability of 0.93, although the probability of detecting a small effect size (r=0.10) is lower (0.21). We believe this power is acceptable for our study, because in order for a difference between conditions to be of practical significance, it would need to be of medium to large effect size. A difference in project grades of less than a few points may not warrant the instructor effort required to implement a change in team formation process.

Student Criteria Choices (RQ1)

75 criteria in total were discussed across all courses, 48 of which (64%) were newly-proposed by students (i.e., not already present in the tool). The new criteria ranged from sensible to potentially irrelevant. One serious criterion was students' involvement in registered student organizations (RSOs). The rationale provided was that students in RSOs may have less time to devote to a team project, but may have more experience working in teams or being leaders. See Table 4 for a categorization of the list of criteria discussed (both new and existing), with examples of criteria falling under each category.

The voting phase eliminated all of the less serious criteria and kept only those which students found more relevant to the course. See Figure 1 for a visualization of the agreement and disagreement in criteria votes and Table 5 for a list of final student criteria selections and weights after the vote[2].

In general, the most popular criteria among students related to scheduling, skills, and work habits, while the least popular were related to aspects of students' past and identity they have no present control over, such as GPA and race. For comparison, see Table 6 for the criteria chosen by the instructors of each course. Note that all the criteria chosen by instructors were selected from the tool's built-in list of criteria, which is based in the team composition literature [37]. Interestingly, many of the weights selected by students were similar to those provided by instructors for criteria that were used in both approaches.

Student Perceptions of Agency (RQ2)

Students across conditions found it important to have a voice in the team formation process used in the course (median=6.0 on a scale of 1 to 7, s=1.17). This belief did not vary according to condition (Wald $\chi^2(1)$=0.07, B=-0.14, p=0.79).

We hypothesized that students in the Learner condition would report feeling more agency than students in the Instructor condition, since they played a greater role in the team formation process by selecting the configuration for the team formation tool. Interviewed students from this condition did express being pleased with their opportunity to contribute to the process:

> *"I thought that was one of the better parts of this course. I was really happy to see that they were taking our input this time around."* (S17)

The median agency score of the Learner condition was higher (Learner: median=5.0 vs. Instructor: median=4.0 on a scale of 1 to 7). However, the difference was not statistically significant (Wald $\chi^2(1)$=3.05, B= 0.77, p=0.08).

One strength of LIFT students identified in the interviews was that it allows the instructor to gain deeper insights into how students actually function. Students often expressed they felt that instructors are disconnected from the student team experience:

> *"The instructor maybe doesn't necessarily see the experience behind it but, if you're working in a group you might want some things that the instructor might not necessarily think about."* (S9, S=10)

Students also reported that LIFT contributed to an increased sense of ownership over the team and its functioning:

> *"I think then it makes the people more accepting of the teams because it's like, 'Well, I was sort of the one who thought we should be grouped like that."'* (S8, S=5)

One identified drawback was that although students can offer direct insight to their needs, they are not always experts on what makes a good team (S=6). Students are frequently unfamiliar with course goals and the team formation literature, and can only draw knowledge from their own experiences (S=2).

[2] Due to space constraints, these describe only the Mechanical Design course. See the supplemental material for the other courses.

Table 4. A categorization of the criteria students discussed across all courses. Criteria with asterisks did not previously exist in the tool.

Category	Subcategory/Theme	Example criteria
Team	Team Management	Schedule, Leadership role, Preferred workplace*
	Coordination Between Teams	Concurrently enrolled in [course]*
	Previous Teamwork Experiences	Teamwork experience*, Sports team experience, Involvement in RSOs*
Academics	Résumé	GPA, Major, Work history*
	Crystallized Knowledge	Software skills, Morning/evening person*, Confidence in programming*
	Commitment	Commitment level, Grade goal*, Extracurricular time commitments
Identity	Demographics	Race, Gender, Age
	Personality/Interests	MBTI personality type*, Personal interests*

Table 5. Criteria and weights selected by Mechanical Design students.

Criterion	Weight
Schedule	4
Morning vs. evening person	3
Theoretical vs. hands-on	-2
CAD skills	-3
Matlab skills	-2
Programming skills	-2
Weekend meetings	3
Enrolled in Senior Design Project	-3

Table 6. The criteria configurations created by the instructors.

	Soft. Engr.	Dsgn. Manf.	Mech. Dsgn.	Stat.	Dsgn. Meth.
Gender	4		2	5	
GPA	-4	-4	-2		
Schedule	5	5	5		
Big picture/detailed	-4		-2		-5
Shop skills			-2		
Race	3		-3	5	
Leadership pref.	1				-3
Leadership role	-4				-3
Commitment level	-4				-4
On-campus job	4				-2
Off-campus job	4				-2
Software skill	-5				
Weekend meetings	5				
English skills	-3				
Hands-on skills		-3			
Prev. course grade				-5	

Instructors know the goals of their course and what skills will be necessary to successfully complete the project:

> *"[Instructors] have the better idea of what they're trying to get out of the class, like what skills they're trying to make us learn, whereas we just want to think about other things... like what kind of grade we're going to get. So theirs is more holistic because they care about every person's skill and how they should improve while the students are only thinking about themselves."* (S4, S=8)

Students also raised concerns about others trying to propose certain criteria or weights in order to unfairly maximize their own gains (e.g., getting paired with their friends):

> *"One thing that I noticed that a lot of teams did... was they'd put that everyone was only free at 8am and they would all get the same group because that's the most important one. So they were able to form groups with their friends and a lot of us weren't aware of that until afterwards."* (S16, S=5)

Effects of Criteria Selector on Outcomes (RQ3)

Project grades and measures of learning and satisfaction were high across all conditions. See Figure 2 for distributions of these measures. The Wald test revealed no significant effect of criteria selector on either project grades or any of our measures of learning and satisfaction. See Table 7 for chi-square values, model coefficients, p-values, and average values for each measure across conditions.

Within the Learner condition, we examined whether selection threshold (All vs. Strict) had an effect on any of our measures by constructing mixed effect models using Threshold as the independent variable. Wald tests again revealed no significant effect of Threshold on any of the outcome measures.

Instructor Perceptions (RQ4)

For the criteria instructors selected, see Table 6. The instructors reported selecting these criteria based on the team formation literature (I3, I4), recommendations from colleagues or experts (I1, I2, I3), personal beliefs and experience (I5, I6), and the project requirements (all).

Prior to the interviews at the end of the semester, the instructors were not aware of the criteria their students selected, in order to prevent potential bias. When presented with the students' criteria, they expressed both their realization of students' perspective and doubts about student choices. For instance, I4, although surprised by the number and variety of criteria presented (42), still learned something about the students:

> *"What we don't do is... consider what they perceive to be their learning style or motive [for] learning... Individual vs. group style, big picture vs. detail oriented, course priority or grade goal are kind of things that might be reflective of different learning styles... "* (I4)

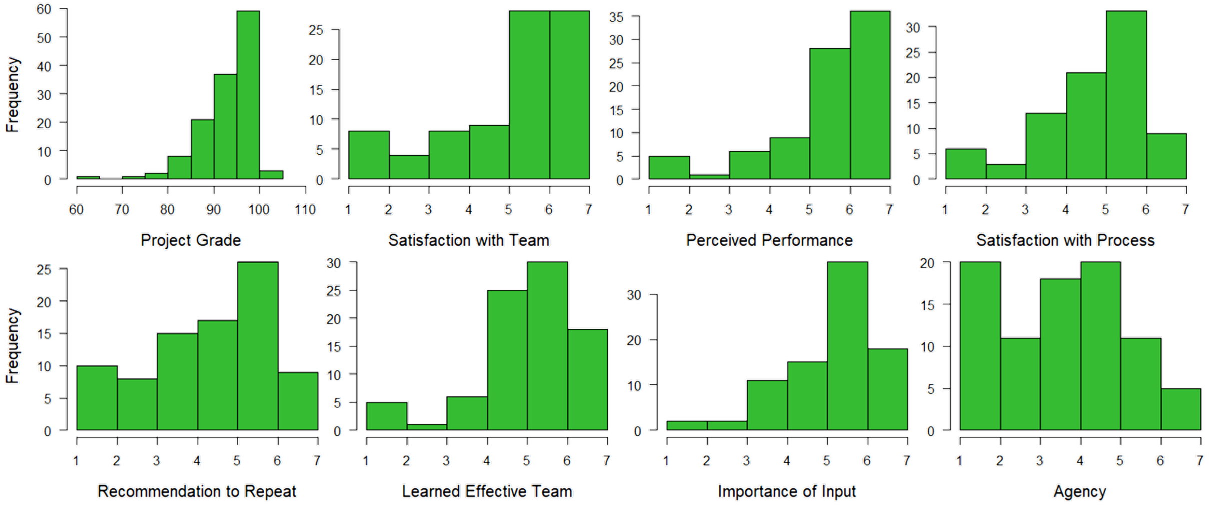

Figure 2. Distributions of the outcome measures. Project scores over 100 exist due to a few teams receiving extra credit.

Table 7. The results of the statistical analysis. χ^2 column shows Wald $\chi^2(1)$.

Measure	Mean (grade)/ Median (others)	Selector			Threshold		
		χ^2	B	p	χ^2	B	p
Project grade	93.33 out of 100	0.14	0.59	0.71	1.12	2.96	0.28
Satisfaction with team	6 on a scale of 1 to 7	0.16	-0.21	0.69	0.02	0.10	0.89
Perceived performance	6 on a scale of 1 to 7	0.91	-0.51	0.34	0.91	0.78	0.34
Satisfaction with process	5 on a scale of 1 to 7	0.01	-0.06	0.91	0.10	0.25	0.74
Recommendation to repeat	5 on a scale of 1 to 7	0.01	0.04	0.94	0.35	0.44	0.55
Learned effective team	6 on a scale of 1 to 7	0.02	-0.07	0.89	0.09	-0.24	0.76

Another instructor who previously had doubts about a criterion he had chosen was interested when his doubts were confirmed:

> *"Was GPA on? See GPA is not even on there! Gosh, see that! The students are smarter than me... See, I guess I wish [I had] heard or learned this earlier."* (I2)

However, some of them believed that although students' motivations are understandable, some criteria were irrelevant:

> *"The instructor can emphasize things in the course that the students might not know about because they're just entering the course. Like they had in here something about programming skills... [which is] not a big deal in this class at all and I didn't do any programming... I just see [that] as not relevant."* (I3)

When asked to compare the two approaches, LIFT was favored by three instructors (I1, I5, I6) who would use the process as is, as they thought it would make students more responsible, more motivated and give them a sense of ownership. A fourth instructor (I2) expressed his willingness to integrate the criteria given high weights by students into his configuration:

> *"They're just used to being assigned to teams or [picking teammates] on the spot. Having them setting the criteria for how teams form... [puts] it on them to make it work... It also I guess put some sort of ownership on everything with them. I think when they have more ownership of something they usually are more involved."* (I5)

> *"Yeah, I can totally adopt this. I don't know if I want to do this many... but I can do the 4s and 3s, and adopt that for next semester. Absolutely."* (I2)

I3 and I4 were more reluctant to adopt LIFT, either because the students' selections neglected certain key criteria, or because large classes could make student input overwhelming:

> *"That's a hard question... there's a lot of literature on gender and achievements and race, like we should really pay attention to that, but then again I don't know. I'm not the students, and I don't know what their biases are, if they have biases... all I know is literature so... I don't know. I don't know if I trust that much that they know themselves so well."* (I3)

> *"I think getting students' input is valuable in this process, but [I'm] not as inclined yet to say we're gonna try and satisfy a group of 600!"* (I4)

I1 was also concerned that students may select criteria that maximize individual gain instead of benefiting everyone:

> *"You have the students decide on the criteria, then... probably most of the students [only] care about maximizing their grade, so [they may] try to pick criteria*

> *in a smart way to maximize their grades, while the professors, we don't care about their grade as much as we care about the total learning, right?"* (I1)

DISCUSSION

We investigated LIFT, a learner-centered workflow for configuring the inputs to algorithmic team formation tools, and found that LIFT is a viable option for including student preferences in the team formation process. Students proposed novel criteria, like organizational style and confidence in programming skills, and selected from known criteria to collectively create configurations that were meaningful to them. All of the criteria individual students proposed that might be considered trivial or ineffective (such as astrological sign and favorite color) were ultimately voted against by the majority, who preferred reasonable criteria that would facilitate project work. Instructors could adopt these criteria, or use LIFT themselves to identify which criteria matter most to their own students.

Teams formed using the student-defined configurations performed no worse than teams formed using the instructor-selected criteria, were no less satisfied with their teams, and felt high levels of control over the team formation process. These results should offer instructors wishing to incorporate student preferences more confidence that they can do so without adversely affecting student grades or team experiences.

We observed several trends in the configurations selected by participants. Students favored criteria related to skills, logistics, and other immediate topics that could help them complete their project more conveniently. For example, schedule was the most popular criterion in four of the five courses. Weights were generally set to distribute skills and make finding meeting times easier. Conversely, students voted against or disregarded criteria related to previous academic performance or demographics, and other aspects of themselves they could not presently control. This trend included even criteria like GPA and gender that have been shown beneficial in prior work [7, 10]. The comments by I1 and S4 also fit with this interpretation of students' goals as maximizing short-term utility.

On the other hand, instructors tended to prioritize student learning and long-term success over minimizing present conflict. They created configurations that included more of the criteria students opposed (such as GPA and gender), sometimes to the exclusion of the logistical criteria like schedule (as in the Statics course). There was, however, some disagreement in whether teammates should be similar or dissimilar with respect to certain criteria. I3 explained that she tries to place students in their zone of proximal development [58] by grouping them with people different from themselves (in terms of academic achievement, work style, etc.). I2 takes an opposite stance:

> *"High achievers may need to be in teams with other high achievers so that they have this sort of conflict… [and] can work through a disagreement with another student. I think it is a wonderful opportunity for growth."* (I2)

Teams in the student- and instructor-defined conditions did not exhibit statistical differences in our outcome measures, despite the differences in the criteria selected. This result argues that the specifics of the configuration may not be the most important factor for team outcomes, at least in the context of the present experiment. Instead, the explanation of the benefits of criteria-based team formation to students may have created an expectation effect contributing to the lack of statistical differences in the outcomes measured in the experiment [26].

One surprising result was that students in the Instructor condition reported experiencing nearly as much agency as students in the Learner condition, despite having minimal input into the formation process. A possible explanation is that students in the Instructor condition found filling out the survey in the tool with their personal information to be sufficient participation. If students in this condition had not been required to enter this information (i.e., if the criteria used could be imported from the course roster or entered into the tool by the instructor), then they might have reported experiencing less agency. However, this result suggests that our agency survey item may not have captured clearly the distinction between "participation" and "choice" (i.e., students in the Instructor condition participated in the process via the survey but did not have a choice in which criteria were on the survey, or how the criteria would be weighed). Future work could further examine this distinction.

Adapting the LIFT Workflow

In this experiment, we implemented LIFT in three stages. Instructors who have taught a course many times could simplify the workflow by having students simply vote for how a given list of criteria should be weighed by the team formation algorithm. This reduced workflow only requires distributing a survey to choose the criteria weights, and may make the process attractive for instructors who want to give students control of the algorithmic inputs without implementing the full workflow. Instructors who are new to teaching or to a particular course might first use the full workflow to determine what criteria matter most to students, and then use the simplified workflow when teaching subsequent instances of the course.

Another possibility is to integrate instructor- and student-chosen criteria into a single configuration. Instructors could refine students' selections when they become too many or too varied, or when they include irrelevant criteria such as skills not required for the course. In addition, instructors can ensure that the criteria selected do not privilege the preferences of some students over others. For instance, minority students may have needs of which other students are not aware, and often struggle to have their voices heard [54]. It may be advisable for instructors to include criteria like gender and race even if students do not select them, in order to promote good experiences for these students. However, instructors should then explain their rationale for these inclusions to students in the course, who might not have been aware of the value of these criteria.

Implications for Tool Designers

Designers of team formation tools should incorporate features for instructors to delegate additional control of algorithmic inputs to their students. At a minimum, the tool could distribute a survey to students to collect and aggregate their individual opinions for the criteria weights, and then show instructors the distribution of these responses adjacent to each criterion in the configuration interface. Instructors could then consider the student input when deciding the configuration. Tools could also link to existing discussion forums such as Piazza, or incorporate their own forums, in order to facilitate student discussion of criteria. "Upvotes" on posts could replace a separate voting survey, helping to automate the process and make it easier for instructors to identify the criteria which students most feel should enter the next stage of the workflow. The surveys and online discussions could be augmented with background knowledge about team composition and resources where students could further learn about the some of the criteria.

Despite the potential benefits of involving students in configuring these tools, participants raised concerns about possible manipulative behavior. We believe it is difficult for students to collude to be placed on the same team due to the complex set of criteria in use (in terms of both number and student similarity or dissimilarity for each criterion). However, the dependency on data self-reported by students remains a weakness of these tools, because students may, intentionally or not, misrepresent their skill sets or other characteristics [5]. Tool designers could take steps to reduce this dependency, for example, by extracting skill data from prior coursework and grade history. Student responses to the team formation survey could also be collected prior to revealing or soliciting the weights. This additional precaution would prevent students looking to game the system from knowing in advance which criteria will have the greatest impact on they way teams are formed.

Tools should also incorporate features that address the burgeoning needs of instructors to learn more about team formation. The instructors who had used the tool previously (I1, I2, I3, I4) all indicated that they used the same criteria over time:

> *"I guess I've always used the standard ones because I don't know any better. . . I figure somebody who is smarter than me has studied this a lot more than I have. You don't mess with the defaults unless you know what you're doing, and I don't claim enough understanding."* (I2)

Tools could be augmented with configuration exemplars or searchable repositories, where instructors could share criteria configurations defined by either themselves or their students, the type and size of class they are teaching, and course makeup. Such features would provide instructors, especially those new to a particular course or to teaching in general, guidance on how to form teams in their courses, and could also aid students contributing to the configurations (e.g., via LIFT).

LIMITATIONS AND FUTURE WORK

Our experiment was conducted in the context of a specific university. Future work could examine whether our findings generalize to other disciplines or institutions with different teaching cultures. Future work could also investigate the impact of student involvement on a broader range of outcomes, such as classroom inclusiveness, climate [6], and patterns of team communication and conflict [20, 14, 31]. In addition, future work could use different survey instruments (e.g., with larger scales) in order to investigate whether a ceiling effect contributed to the lack of statistical differences we observed in this study, or whether students interpreted survey items in ways different from our intent.

The criteria proposed by students also present opportunities for further experimentation on team composition. For example, it is unclear how teams formed according to these criteria (or others gathered using LIFT) would perform relative to those formed according to other criteria in the literature, or teams formed without using an algorithmic tool (i.e., randomly or through self-selection). Future work could also explore predicting outcomes of interest from these and existing criteria, using techniques such as regression or decision trees.

Finally, we limited student involvement in this study to the configuration of criteria in the tool. Future work could explore other strategies, such as those where students meet potential teammates and rate candidate partners [41, 42] or explicitly select classmates with whom they would like to work [43].

CONCLUSION

We reported the results of an experiment evaluating a learner-centered workflow (LIFT) for implementing algorithmic team formation in courses leveraging team-based learning. Following LIFT, students propose and discuss criteria that they deem important, vote on whether these criteria should be included in the team formation tool, and collectively configure the weight for each criterion in the tool. Students generally proposed criteria related to team management, academics, and personal identity, and ultimately voted to include skills, logistics, and other criteria that could contribute to completing their project more efficiently. They tended to vote against certain criteria recommended in the literature such as gender, race, and GPA. In addition, students grouped into teams using LIFT achieved project grades and satisfaction comparable to students grouped using the instructor-led approach. Through semi-structured interviews, we evaluated student and instructor perceptions of LIFT and of what they learned during the team formation process. Students were appreciative of having their voices heard, and instructors reported gaining new insight into the team formation criteria preferred by students, as well as a willingness to use LIFT in the future. These results strongly suggest that instructors can (and should) deploy mechanisms such as LIFT to gather student input for team formation criteria, rather than only asking students about their individual attributes.

ACKNOWLEDGEMENTS

This project is part of a larger work supported in part by the Academy for Excellence in Engineering Education at the University of Illinois at Urbana-Champaign. We would like to thank Priscilla Ferronato and Profs. Eric Benson, Alison Dunn, Grigore Rosu, Bruce Flachsbart, and Mariana Kersh for their support of the study and help with data collection. We also thank Farnaz Jahanbakhsh for her invaluable feedback.

REFERENCES

[1] 2017. Team-Maker Algorithm Detail. (2017). Retrieved 4 April 2019 from `https://www.catme.org/faculty/help#TeamMakerScoring`

[2] 2018. CATME Smarter Teamwork. (2018). Retrieved 31 August 2018 from `http://info.catme.org/`

[3] 2018. Piazza Homepage. (2018). Retrieved 20 September 2018 from `https://piazza.com/signup`

[4] S Adams, L Simon, and B Ruiz. 2002. A Pilot Study of the Performance of Student Teams In Engineering Education. In *2002 American Society for Engineering Education Annual Conference & Exposition, Montreal, Canada*.

[5] Albatool A. Alamri and Brian P. Bailey. 2018. Examination of the Effectiveness of a Criteria-based Team Formation Tool. In *Frontiers in Education*. IEEE.

[6] Elliot Aronson. 2002. Chapter 10 - Building Empathy, Compassion, and Achievement in the Jigsaw Classroom. In *Improving Academic Achievement*, Joshua Aronson (Ed.). Academic Press, San Diego, 209 – 225. DOI: `http://dx.doi.org/https://doi.org/10.1016/B978-012064455-1/50013-0`

[7] Donald R Bacon, Kim A Stewart, and Sue Stewart-Belle. 1998. Exploring predictors of student team project performance. *Journal of Marketing Education* 20, 1 (1998), 63–71.

[8] Amna Basharat. 2016. Learnersourcing Thematic and Inter-Contextual Annotations from Islamic Texts. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 92–97.

[9] 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.

[10] 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.

[11] Paul Blowers. 2003. Using student skill self-assessments to get balanced groups for group projects. *College Teaching* 51, 3 (2003), 106–110.

[12] Dale E Brashers, Mark Adkins, and Renée A Meyers. 1994. Argumentation and computer-mediated group decision making. *Group communication in context: Studies of natural groups* (1994), 263–282.

[13] 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.

[14] Sally A Carless and Caroline De Paola. 2000. The measurement of cohesion in work teams. *Small group research* 31, 1 (2000), 71–88.

[15] Paula E. Chan, Kristall J. Graham-Day, Virginia A. Ressa, Mary T. Peters, and Moira Konrad. 2014. Beyond Involvement: Promoting Student Ownership of Learning in Classrooms. *Intervention in School and Clinic* 50, 2 (2014), 105–113. DOI: `http://dx.doi.org/10.1177/1053451214536039`

[16] Kenneth J Chapman, Matthew Meuter, Dan Toy, and Lauren Wright. 2006. Can't we pick our own groups? The influence of group selection method on group dynamics and outcomes. *Journal of Management Education* 30, 4 (2006), 557–569.

[17] David T Conley and Elizabeth M French. 2014. Student ownership of learning as a key component of college readiness. *American Behavioral Scientist* 58, 8 (2014), 1018–1034.

[18] Mary L Connerley and Fred A Mael. 2001. The importance and invasiveness of student team selection criteria. *Journal of management education* 25, 5 (2001), 471–494.

[19] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455.

[20] Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly* 44, 2 (1999), 350–383.

[21] Susan Brown Feichtner and Elaine Actis Davis. 1984. Why some groups fail: A survey of students' experiences with learning groups. *Organizational Behavior Teaching Review* 9, 4 (1984), 58–73.

[22] R Brent Gallupe, Lana M Bastianutti, and William H Cooper. 1991. Unblocking brainstorms. *Journal of applied psychology* 76, 1 (1991), 137.

[23] Elena L Glassman, Aaron Lin, Carrie J Cai, and Robert C Miller. 2016. Learnersourcing personalized hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1626–1636.

[24] Elena L Glassman and Robert C Miller. 2016. Leveraging Learners for Teaching Programming and Hardware Design at Scale. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, 37–40.

[25] Beth Harry, Keith M Sturges, and Janette K Klingner. 2005. Mapping the process: An exemplar of process and challenge in grounded theory analysis. *Educational researcher* 34, 2 (2005), 3–13.

[26] Emily M. Hastings, Farnaz Jahanbakhsh, Karrie Karahalios, Darko Marinov, and Brian P. Bailey. 2018. Structure or Nurture? The Effects of Team-Building Activities and Team Composition on Team Outcomes. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. ACM.

[27] Tyson R Henry. 2013. Creating effective student groups: an introduction to groupformation. org. In *Proceeding of the 44th ACM technical symposium on Computer science education*. ACM, 645–650.

[28] 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.

[29] 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.

[30] Irving Lester Janis. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes*. Vol. 349. Houghton Mifflin Boston.

[31] 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.

[32] Sara Kiesler, Jane Siegel, and Timothy W McGuire. 1984. Social psychological aspects of computer-mediated communication. *American psychologist* 39, 10 (1984), 1123.

[33] Juho Kim. 2015. *Learnersourcing: improving learning with collective learner activity*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[34] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[35] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 6 (2013), 33.

[36] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. ACM, 23–34.

[37] 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.

[38] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

[39] Shang-Wen Daniel Li and Piotr Mitros. 2015. Learnersourced recommendations for remediation. In *Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference on*. IEEE, 411–412.

[40] 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.

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

[42] 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.

[43] Sakthi Mahenthiran and Pamela J Rouse. 2000. The impact of group selection on student performance and satisfaction. *International Journal of Educational Management* 14, 6 (2000), 255–265.

[44] Jeffrey A Mello. 1993. Improving individual member accountability in small work group settings. *Journal of Management Education* 17, 2 (1993), 253–259.

[45] Piotr Mitros. 2015. Learnersourcing of complex assessments. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 317–320.

[46] Roshanak Zilouchian Moghaddam, Brian P Bailey, and Christina Poon. 2011. Ideatracker: An interactive visualization supporting collaboration and consensus building in online interface design discussions. In *IFIP Conference on Human-Computer Interaction*. Springer, 259–276.

[47] Richard Nadeau, Edouard Cloutier, and J-H Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review* 14, 2 (1993), 203–213.

[48] 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.

[49] Amy Pavel, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: video-based asynchronous video review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 517–528.

[50] 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.

[51] Scott Rice and Margaret N Gregor. 2016. *E-Learning and the Academic Library: Essays on Innovative Initiatives*. McFarland.

[52] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.

[53] Niloufar Salehi and Michael S Bernstein. 2018. Hive: Collective Design Through Network Rotation. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 151.

[54] James B Shaw. 2004. A fair go for all? The impact of intragroup diversity and diversity-management skills on student experiences and outcomes in team-based class projects. *Journal of Management Education* 28, 2 (2004), 139–169.

[55] Jane Siegel, Vitaly Dubrovsky, Sara Kiesler, and Timothy W McGuire. 1986. Group processes in computer-mediated communication. *Organizational behavior and human decision processes* 37, 2 (1986), 157–187.

[56] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[57] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 16.

[58] Lev Vygotsky. 1987. Zone of proximal development. *Mind in society: The development of higher psycological processes* 5291 (1987), 157.

[59] Dai-Yi Wang, Sunny SJ Lin, and Chuen-Tsai Sun. 2007. DIANA: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. *Computers in Human Behavior* 23, 4 (2007), 1997–2010.

[60] Helen Wauck, Yu-Chun Grace Yen, Wai-Tat Fu, Elizabeth Gerber, Steven P Dow, and Brian P Bailey. 2017. From in the Class or in the Wild?: Peers Provide Better Design Feedback Than External Crowds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5580–5591.

[61] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 405–416.

[62] 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.

[63] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.

[64] 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.