LIFT: INTEGRATING STAKEHOLDER VOICES INTO ALGORITHMIC TEAM FORMATION

BY

EMILY MARIE HASTINGS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Advisers:

Professor Brian Bailey
Professor Karrie Karahalios

ABSTRACT

Team formation tools assume instructors should configure the criteria for creating teams, precluding students from participating in a process that affects their learning experience. We propose LIFT, a novel learner-centered workflow where students propose, vote for, and weigh team formation criteria, and the collective results serve 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. We found learners were capable of proposing novel criteria not part of existing algorithmic tools, like organizational style. Generally, learners avoided criteria frequently selected by instructors, including gender and GPA, and instead preferred those that promoted efficient collaboration. Second, LIFT led to team outcomes comparable to those achieved by the instructor-led approach, despite the differences in the configurations, and teams valued having control of the team formation process. We provide instructors and tool designers with a workflow and evidence supporting giving learners control of the algorithmic process used for grouping them into teams.

To my parents, Kevin and Gay Lynn, for their unending love and support.

ACKNOWLEDGMENTS

I would like to express the deepest gratitude to my advisor Professor Brian Bailey, without whose guidance and support this work would not have been possible. I would also like to thank my co-advisor Professor Karrie Karahalios, as well as Professor Darko Marinov, for their valuable input and support on this project. This work would also not have been possible without the contributions and support of my friends and collaborators Albatool Alamri, Andrew Kuznetsov, Christine Pisarczyk, and Farnaz Jahanbakhsh. I wish them the best of luck in all of their future endeavors.

I am deeply grateful to the students and instructors who participated in this study, the anonymous reviewers who offered valuable feedback on this work, and the other members of Professor Bailey's research group, who have always been happy to provide insights and support. I would also like to thank the Strategic Instructional Innovations Program at the University of Illinois, the National Science Foundation, the National Physical Science Consortium, and the National Institute of Standards and Technology for their financial support of me and this project.

Finally, I would like to express my love and gratitude to my family and friends. Their constant support and their belief in me have motivated me throughout the completion of this work and my life as a whole.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Instructors are increasingly utilizing algorithmic team formation tools such as CATME [1] 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 control an algorithm that will affect their team experience, learning, and grades [2].

In this paper, we introduce and empirically investigate a novel learner-centered workflow for identifying team formation criteria that are meaningful to students and deciding how to configure 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. The students then vote on whether they think each proposed criterion should be included in the team formation tool. Finally, the students collectively provide a weight for each selected criterion that is then entered into the tool. 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., [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]).

Giving students control of the inputs for algorithmic team formation has many potential benefits. For example, students have independent and localized knowledge of what makes a good team based on their own prior team experiences, of which the instructor may not be aware. In addition, involving students can prevent them from viewing the team formation tool as a "black box," which can lead to suspicions of favoritism and distrust of the instructor [15]. Research in algorithm transparency shows that increasing user knowledge of and control over algorithmic processes can increase satisfaction [16] and improve trust and acceptance of these systems [17, 18, 19]. Finally, increased control over the team formation process can prompt students to take greater ownership of group problems [20].

Given the prior work and these potential benefits, we hypothesized that students using LIFT would be capable of proposing rational criteria configurations that represent their collective preferences, and that they would have team outcomes at least as positive as those achieved by students using the traditional instructor-led process. To test these hypotheses and answer several additional research questions, we conducted a mixed-methods experiment in five university courses leveraging team-based learning (N=289 students). We compared LIFT to the traditional 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 many novel criteria including personal organization style and confidence in programming skills. In general, they favored criteria related to current skills and abilities, logistics, and other factors that could contribute to completing their project more efficiently. Interestingly, most students voted against or disregarded criteria frequently used by instructors, including gender, race, and GPA. This disagreement came from both male and female students. This finding is surprising because these criteria are commonly used by instructors and are supported by existing studies of team composition (e.g., [21, 22, 23, 24]). Our second hypothesis was partially supported: we show that 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, and students found it important to have a voice in configuring the team formation tool (Median=6.0 on a scale from 1 to 7, 7=most preferred). Finally, we also 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 that instructors can deploy to tap into the criteria that matter most to students in the context of their specific courses. While this paper focuses on deploying LIFT in face-to-face classrooms, the proposed workflow for team formation generalizes to online learning environments and other contexts, such as online labor markets and open design challenges. For example, the criteria discussion and voting phase could complement systems described in other work, such as the team dating procedure in [25, 26] or Hive's network rotation [27]. Finally, we share implications for how designers of team formation tools can give stakeholders more control over the algorithmic team formation process. For example, the tools might provide instructors with a graphical representation 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.

CHAPTER 2: 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.

2.1 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 removes the stress of having to form a team on one's own [2]. 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 like CATME [1] are increasingly being deployed to implement criteria-based team formation processes and to help instructors keep pace with increasing course enrollments.

Researchers have studied how grouping people into teams based on different criteria affects team outcomes. For example, Woolley found that including more women in a team raises the team's collective intelligence [21, 28]. Lykourentzou et al. show that team performance and satisfaction can be increased by balancing personality types within a team [29]. Wen et al. found that teams formed according to pairwise transactivity in a discussion activity showed greater knowledge integration than random teams [30]. Team performance has also been shown to increase through including a diverse set of skills that are relevant to the project [23], by including moderate nationality diversity [31], by grouping according to academic ability and curricular interests [22], or by periodically modifying team membership according to tie strength [27].

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 in different contexts to give stakeholders control of the algorithmic processes used to group them into teams. This work is timely because many instructors, especially in the engineering disciplines, are implementing the algorithmic approach in their courses due to growing enrollments and increased diversity, and it is not clear how incorporating new mechanisms like giving students control over the algorithmic process could affect the deployment of this approach in authentic learning environments.

2.2 OTHER TEAM FORMATION APPROACHES

Self-selection and random assignment are some of the most common team formation approaches, especially since they are easy to implement by the instructor. These approaches can promote positive team experiences; for example, self-selection can increase team satisfaction [32] and encourage group members to take ownership over group interactions and conflicts [20]. Increased ownership over learning activities helps students to set goals, solve complex problems, and create high-quality work [33, 34].

However, algorithmic team formation has become more popular in part because it addresses some of the weaknesses of such approaches. For instance, self-selection may leave some students unable to find a team to join [35], and random assignment has been shown to lead to reduced team satisfaction [32]. In addition, these strategies often produce teams that lack the needed skill variety to accomplish the course tasks [36, 32].

Our contribution is to further strengthen algorithmic team formation by incorporating some of the strengths of these approaches, such as giving stakeholders increased ownership of the team assignments.

2.3 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 [37]. Crowdsourcing is increasingly being applied in learning environments, where learners often serve as the crowd. Learnersourcing has been defined as "a form of crowdsourcing in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves" [3]. For example, researchers have used learnersourcing to transcribe speech for online lecture videos [4], and to create a learning hierarchy for instructional media [5]. Learnersourcing has also been used to annotate texts [6] and to create assignments [38]. In addition, the method has been shown valuable for assessing learner solutions [8], and generating design feedback [9, 10] and problem solving advice [11, 12, 13, 14].

This 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 the theory of team formation and how it can impact their team experience and outcomes.

CHAPTER 3: 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 [39]. We chose this tool because it is used in many courses at our university and is grounded in the team composition literature [35].

In the typical CATME workflow, the instructor chooses from a set of predefined criteria or defines their own criteria based on learning goals for the course and the team composition literature. The tool provides 27 criteria by default, including demographics, schedules, and working styles. When the criteria have been selected, the tool creates a survey with questions related to the selections and distributes its link to the students in the course via email. When students have completed the survey, the instructor then reviews the responses and configures the weights for the criteria. Each criterion is assigned a weight 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 a weight of 5 to the schedule criterion strongly prefers groups where students report similar schedules. The tool then forms teams based on these weights using a randomized greedy algorithm, 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.

CHAPTER 4: THE LIFT WORKFLOW

Our proposed workflow consists of three main stages. First, students perform an online discussion activity prior to teams being formed, in which they discuss which formation criteria they think should be used in the course. Second, once the discussion is completed, students vote on which of the proposed criteria should be included in the team formation tool. Third, each student selects the weights for the criteria that she or he would prefer in the team formation tool, and these individual selections are averaged to create a configuration for the entire class. The goal of this workflow is to learn what criteria students think are important to consider when forming teams, and to use this knowledge to give students an increased sense of agency over the team formation process.

Criteria are proposed through a discussion because we wanted to elicit rich information about the similarities and differences in students' preferences on team formation. We believed that surveying individual students would not be as effective, since students would not be aware of their classmates' contributions and concerns, and would be unable to react to them. Therefore, they would likely generate repetitive criteria, and their responses would not provide as much insight as a dialogue between students would. However, freedom from the judgement of peers is crucial for the discussion to be truly reflective of student opinions. Therefore, the option of anonymity is necessary for students to be comfortable enough to propose and discuss criteria that may be more sensitive, such as race and gender. This decision is based on prior work which has shown that anonymity can promote increased and more egalitarian participation [40, 41], aid idea generation [42], and reduce status differences [43].

The voting stage takes place after the discussion is finished rather than continuously (e.g., "upvoting" posts as they are made). This choice gives students a chance to view all of the criteria that were proposed along with the associated conversations. Voting at this point allows students to form their own opinions prior to voting, and facilitates reaching consensus [7]. These votes are collected individually through a survey in order to prevent students' responses being influenced by seeing those of the majority [44].

Configuration of the team formation tool occurs at the time of students providing their information in the team formation survey. As students provide their responses for each selected criterion on the survey, they also specify the magnitude and sign of the weights for these criteria. 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.

CHAPTER 5: 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 or disagreement is there among students? How do student- and instructor-chosen criteria differ?

RQ2: How does allowing students to select criteria affect their team performance, satisfaction, and other course experiences compared to having instructors select criteria?

RQ3: How do students perceive their agency when they are allowed to have input into the team formation process?

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 regarding the use of team formation tools, as well as how these preferences relate to selections instructors make in practice. This new understanding will help HCI researchers, tool designers, and instructors develop and deploy tools that more closely consider student voices.

CHAPTER 6: 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.

6.1 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 6.1. 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 plans and proposals, prototypes, and final demonstrations and reports.

6.2 CRITERIA SELECTION

To determine which team formation criteria students and instructors select, as well as how each perceive their agency in the team formation process, we utilized two different methods to select and weight the criteria used in the team formation tool. In one version (the LIFT workflow), 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, which acted 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 that were 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.

Table 6.1: 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%

6.2.1 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 [45], an educational platform which was able to provide a discussion environment restricted only to students in the course, as well as the option of anonymity. This platform has been used previously for educational crowdsourcing [46]. 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 then asked to make at least three contributions to the discussion, where a contribution was either (a) a post identifying a criterion and explaining why they thought it important for the course, or (b) a follow-up comment on another student's post discussing advantages or disadvantages of the criterion or suggesting enhancements to it. Students were told that the criteria they discussed should be relevant to the course and come either 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 due date for the activity passed, the research team examined the discussion and compiled a list of all the criteria proposed by students, discarding duplicates and those few criteria 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. The criteria were then ranked from most popular 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 (hereafter called 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 of these thresholds: 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.

6.2.2 Instructor Condition

To maintain a consistent workload between conditions, students in the Instructor condition also performed an online discussion activity prior to teams being formed. In this activity, they were asked to discuss their previous team experiences, or if they had never worked as part of a team, to describe 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, the instructor of each course 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.

Table 6.2: Response rates for the study activities.

	Soft. Engr.	Dsgn. Manf.	Mech. Dsgn.	Stat.	Dsgn. Meth.
Discussion activity	78%	97%	83%	59%	100%
Formation survey	92%	98%	95%	78%	77%
Midpoint survey	75%	74%	80%	69%	35%
Peer evaluation	74%	90%	96%	54%	30%
Post-survey	48%	51%	71%	53%	28%

6.3 PROCEDURE

Students had approximately 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 and distributed it to students. Students then had approximately one week to complete this survey. Those who did not respond to this survey were placed onto a team randomly. Two to three weeks after teams were formed (after teams had completed their first course assignment together), students completed a brief survey asking about their satisfaction with the team formation process and their first impressions of their team. In the final week of the course, students were asked to complete the summative 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 at the end of the course.

Completion of the surveys, peer evaluations, and online discussions was either required as part of regular course instruction or compensated with extra credit, depending on the course. See Table 6.2 for the response rate for these activities. To give a more accurate picture of class participation, the response rates in the table include all students who completed the activities. However, only those students (N=289 of 936) who gave consent to use their data were included in our analysis.

In addition, we recruited N=18 students to take part in semi-structured interviews for more detailed feedback regarding the team formation process they experienced in the course. The questions in these interviews focused on student perceptions of the criteria chosen, as well as strengths and weakness of both LIFT and the instructor-led approach. We also interviewed the six instructors of the courses studied. These questions focused on their previous experience with the team formation tool, their perceptions of the criteria chosen by students, and strengths and weaknesses of both approaches. 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 members of the research team then collaboratively coded each transcribed interview, categorized the codes, and refined these categories into themes [47].

6.4 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 with the team and the team formation process.

6.4.1 Project Grades

We assessed student team performance by using the grade the team received on the project they completed together. This data was collected as a part of regular course instruction.

6.4.2 Team Satisfaction

Students rated their satisfaction with their project team 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.").

6.4.3 Process Satisfaction

Students rated their experience with the team formation experience by agreeing or disagreeing with statements about: satisfaction with the approach ("What has been your experience with the approach used in this course?", 1=Very poor, 7=Excellent), perceived agency ("I felt I had a voice in shaping how teams were formed in this class.", 1=Strongly Disagree, 7=Strongly Agree), 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 recommendation to repeat the approach ("I recommend repeating the approach to team formation I experienced in this course in the future.").

CHAPTER 7: RESULTS

To answer the quantitative aspects of our 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=0.05/7=0.007.

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. Additionally, we have limited our statistical analysis to three of the five courses (Software Engineering, Mechanical Design, 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 performed a power analysis using 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 that this is an acceptable power 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. For example, a difference in project grades of less than a few points would likely not warrant the instructor effort required to implement a change in team formation process.

We complement our statistical analysis with support from our interviews and the online discussion activity, for which we have data from all five courses.

7.1 STUDENT CRITERIA CHOICES (RQ1)

During the discussion activity, students advocated for both novel criteria and criteria that already exist in the tool (such as schedule and gender). There were 75 criteria discussed in

Table 7.1: A categorization of the criteria students discussed. Criteria with asterisks are newly-proposed and 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, Hands-on skills, Morning/evening person*, Big picture/detail-oriented, Confidence in programming skills*
	Commitment	Commitment level, Grade goal*, Extracurricular time commitments
Identity	Demographics	Race, Gender, Age
	Personality/Interests	MBTI,* Personal interests*

total among all five courses, 48 of which (64%) were newly-proposed by students. The new criteria students proposed ranged from rational and sensible to humorous and potentially irrelevant. For example, a serious criterion proposed was students' involvement in registered student organizations (RSOs). The rationale provided for this criterion was that students involved in these organizations may have less time to devote to a team project, but may have more experience working in teams or being leaders. See Table 7.1 for a categorization of the list of criteria discussed (both new and existing), with category descriptions and examples of criteria falling under each category.

The voting phase of the activity eliminated all of the less serious criteria and kept only those which were more relevant to the course in the view of the students. See Figure 7.1 for a visualization of the agreement and disagreement in criteria votes and Table 7.2 for a list of final student criteria selections and weights[1].

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

[1] For simplicity, we here present this information for only the Mechanical Design course as an example. See the Appendices for information on the other courses.

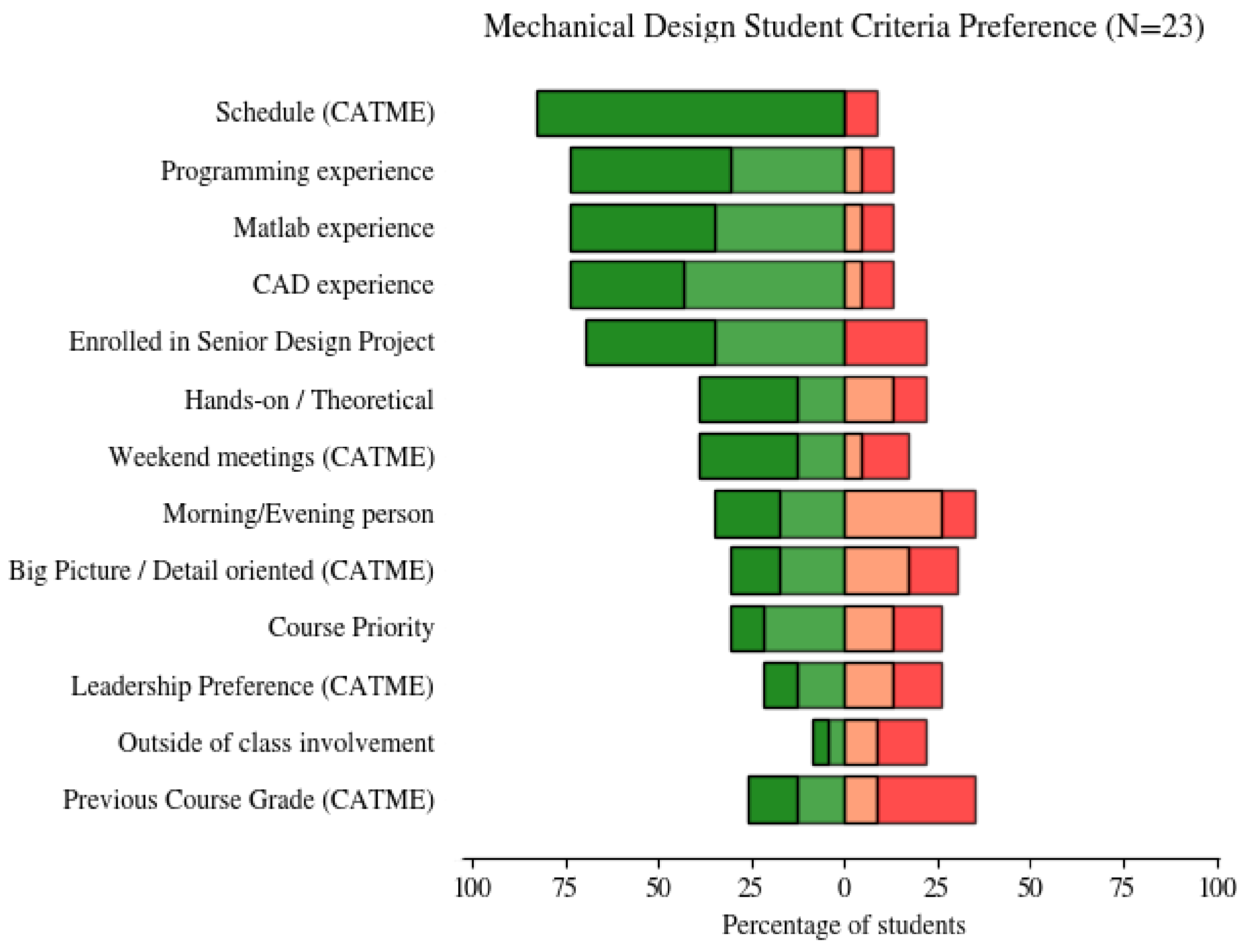

Figure 7.1: The distribution of votes for criteria discussed in the Mechanical Design course. For each criterion, the colored bars represent from left to right votes for “Strongly Agree”, “Agree”, “Disagree”, and “Strongly Disagree” that the criterion should be included in the tool. For example, Schedule received strong support, while Big Picture/Detail-Oriented received roughly equal amounts of positive and negative votes.

Table 7.2: The final criteria and weights selected by students in the Mechanical Design course.

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 7.3: The criteria configurations created by the instructors in each course.

	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	

they have no present control over, such as GPA and race. For comparison, see Table 7.3 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 [1]. Interestingly, many of the weights selected by students were similar to those provided by instructors for criteria that were used in both approaches.

7.2 EFFECTS OF CRITERIA SELECTOR ON OUTCOMES (RQ2)

Project grades and measures of team and process satisfaction were high across all conditions. See Figure 7.2 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 team and process satisfaction. See Table 7.4.

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

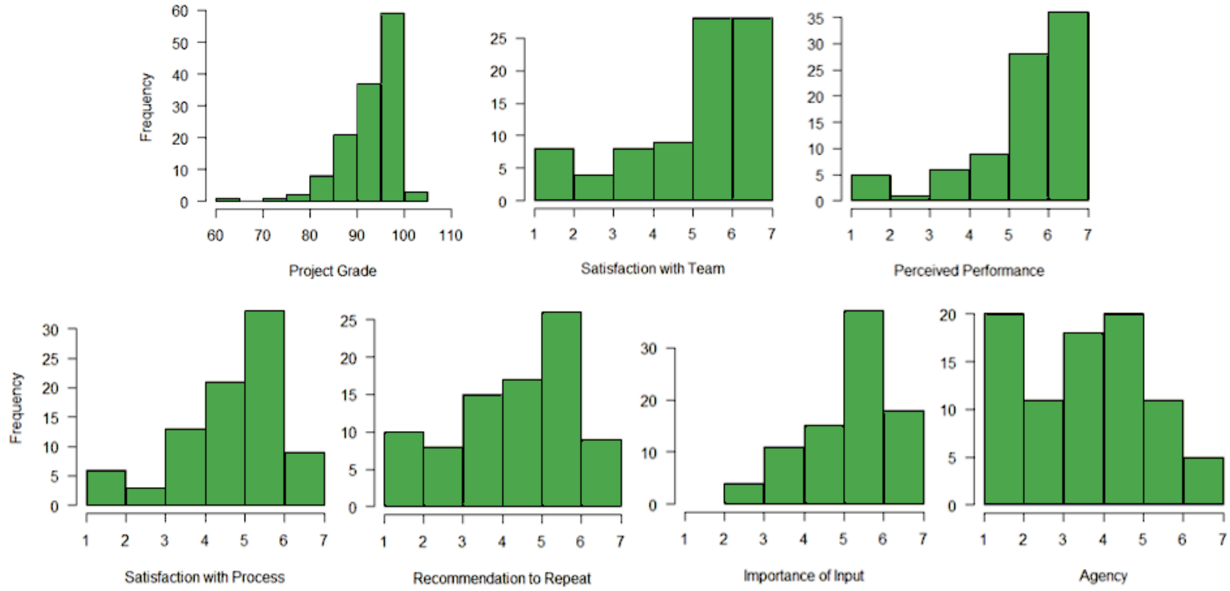

Figure 7.2: Distributions of the outcome measures. Project scores over 100 exist due to a few teams receiving extra credit.

Table 7.4: 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

7.3 STUDENT PERCEPTIONS OF AGENCY (RQ3)

Students across conditions thought that it was 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).

7.3.1 Strengths and Weaknesses of LIFT

During the interviews, we asked students what strengths and weaknesses they saw with the LIFT workflow. One strength students identified was that LIFT allows the instructor to gain deeper insights into how students actually function. Interviewed 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, N=10)

In addition, another benefit of LIFT that was reported by students was that it 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, N=5)

One drawback that was identified was that although students can offer a direct insight to their needs, they are not always experts on what makes a good team (N=6). Students are frequently unfamiliar with course goals and the team formation literature, and can only draw knowledge from their own experiences (N=2). Instructors know the goals of their course and what skills will be necessary to successfully complete the project:

"The 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, probably 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, N=8)

Additionally, allowing students to propose criteria may potentially allow them to manipulate the team formation process unfairly. Interviewed students raised concerns about others trying to propose certain criteria or weights in order to maximize their gains, often in the form of 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, N=5)

7.4 INSTRUCTOR PERCEPTIONS (RQ4)

To understand instructors' perception of LIFT, we conducted interviews with the instructors of the five participating courses. Note that the Design Methods course had two co-instructors. The interviews focused on learning about instructor's own criteria and rationale, thoughts on students' criteria, beliefs on the strengths and weakness of both approaches, and whether they would be willing to adopt students' criteria and LIFT.

For the criteria instructors selected, see Table 7.3. The instructors reported selecting these criteria based on sources of knowledge on team formation including the 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, which were conducted at the end of the semester, the instructors were not aware of the criteria their students selected, in order to prevent any potential bias. When they were presented with the students' criteria and began looking for similarities between theirs and the students', they expressed both their realization of students' perspective and doubts about student choices. For instance, instructor I4, although feeling overwhelmed by the number of criteria presented (42) and how varied they were, still came to learn something new about the students:

> *"I guess what we don't do is really consider what they perceive to be their learning style or motive [for] learning. I see for example 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)

Another instructor who previously had doubts about a criterion he had chosen (GPA), 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)

On the other hand, 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)

Furthermore, when asked about comparing the two approaches (instructor vs learner), 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... kind of [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, that's a lot of questions, but I can do the 4s and 3s, and adopt that for next semester. Absolutely." (I2)

On the other hand, I3 and I4 expressed their reluctance to adopt that approach, either because the students' selections neglect key criteria in the literature such as gender, race and GPA, or because the number of students in class is so large that taking their input could be difficult:

"That's a hard question to answer though. Part of me says... 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)

Despite the positive aspects of LIFT, I1 mentioned his concern that students may select criteria that maximize the 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)

CHAPTER 8: DISCUSSION

In this work 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. In addition, 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 rational criteria that would facilitate project work.

We also found that teams formed using the student-defined configurations performed as well as teams formed using the instructor-selected criteria, were no less satisfied with their teams, and felt more 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.

Examining the criteria that were selected by students and instructors, we observed several trends. The students in our study preferred configurations which focus on short term project success and satisfaction during the course of the semester. Students favored criteria related to current skills and abilities, logistics, and other immediate topics that could contribute to making the completion of their project more convenient or efficient. For example, Schedule was the most popular criterion in four of the five courses, and it was in the top 4 criteria in the remaining course (Design Methods). 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 [48] by grouping them with people different from themselves (in terms of academic achievement, work style, etc.). I2 takes an opposite stance:

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

Teams in the student- and instructor-defined conditions did not exhibit statistical differences in the measures of performance or satisfaction, 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 use of a criteria-based team formation approach and the explanation of its benefits to students may have contributed to the lack of statistical differences in the outcomes measured in the experiment [49].

One surprising result from our experiment was that students in the Instructor condition reported experiencing nearly as much agency as students in the Learner condition, even though they had had minimal input into the team formation process. A possible explanation for this finding is that students in the Instructor condition felt that filling out the survey in the team formation tool with their personal information was sufficient participation in the process. Presumably, if students in this condition had not been required to enter information in the tool (i.e., if the criteria used could be imported from the course roster or entered into the tool by the instructor), then those students 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, as students in the Learner condition did). Future work could further examine this distinction.

8.1 ADAPTING THE LIFT WORKFLOW

In this experiment, we implemented the LIFT workflow in three-stages: 1) students propose and discuss criteria, 2) students vote for which criteria should be used for team formation, and 3) students vote for how the criteria should be weighed by the team formation algorithm. Experienced instructors who have taught a course many times could simplify the workflow by eliminating the first two stages and only having students vote for how a given list of criteria should be weighed by the team formation algorithm. This simplified workflow only requires distributing a survey for choosing the criteria weights, and may make the process attractive for instructors who want to give students control of the algorithmic inputs without the cost of 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.

A second adaptation of LIFT that was suggested in several student interviews is to integrate both instructor- and student-chosen criteria into a single configuration. This adaptation could prove useful for instructors who may feel the need to refine students' selections when they become too many or too varied, or when they include criteria deemed irrelevant by the instructor such as skills not required for the team work in the course. In addition, the instructor can make sure that the criteria selected do not privilege the preferences of some students over others. For instance, minority students may have needs of which the other students are not aware. 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 team experiences for these students. Forming groups according to these criteria can also help students to learn to work with people who are different from themselves.

8.2 IMPLICATIONS FOR TOOL DESIGNERS

Designers of team formation tools should incorporate features that enable instructors to delegate additional control of the algorithmic inputs in the tool 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 these discussion posts could replace a separate voting survey, helping to automate the process and making it easier for instructors to identify the criteria students feel most strongly about and 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 the process of configuring these tools, concerns about possible manipulative behavior were raised by the participants in our study. We believe that 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, either intentionally or not, misrepresent their skill sets or other characteristics [50]. Tool designers should take steps to reduce this dependency, for example, by extracting skill data from prior coursework and grade history. In addition, student responses to the team formation survey could 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.

Team formation tools should also incorporate features that address the burgeoning needs of instructors to learn more about team formation. The interviewed instructors who had used the tool previously (I1, I2, I3, I4) were asked if they used the same criteria over time. All of them answered affirmatively and when asked why, they made statements such as:

> *"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)

Team formation tools could be augmented with configuration exemplars or searchable knowledge repositories, where the instructors using the tools 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.

CHAPTER 9: LIMITATIONS AND FUTURE WORK

Our experiment was conducted in primarily engineering courses in the context of a specific university. Future work could examine whether our findings generalize to other disciplines or academic institutions with different teaching cultures. In addition, our work considered the effects of student involvement on a limited set of outcome measures, such as team satisfaction and performance. Future work could investigate the impact on a broader range of outcomes, such as individual learning measures (including learning about team composition), classroom inclusiveness, and climate [51], and on patterns of team communication and conflict. The criteria proposed by the students in this study also present opportunities for further experimentation on team composition.

We limited student involvement in this study to the configuration of criteria in a team formation tool. This is only one approach to incorporating student input into the team formation process. Future work could explore additional strategies, such as those where students meet potential teammates and rate candidate partners [25, 26] or explicitly select classmates with whom they would like to work [52].

CHAPTER 10: 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. We found that students generally proposed rational 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 into teams using the instructor-led approach. These results question the existing assumption that instructors alone should configure team formation tools.

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.

REFERENCES

[1] R. A. Layton, M. L. Loughry, M. W. Ohland, and G. D. Ricco, "Design and validation of a web-based system for assigning members to teams using instructor-specified criteria." *Advances in Engineering Education*, vol. 2, no. 1, p. n1, 2010.

[2] F. Jahanbakhsh, W.-T. Fu, K. Karahalios, D. Marinov, and B. Bailey, "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, 2017, pp. 3201–3212.

[3] J. Kim et al., "Learnersourcing: improving learning with collective learner activity," Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[4] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham, "Real-time captioning by groups of non-experts," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 23–34.

[5] S. Weir, J. Kim, K. Z. Gajos, and R. C. Miller, "Learnersourcing subgoal labels for how-to videos," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 405–416.

[6] A. Basharat, "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, 2016, pp. 92–97.

[7] R. Z. Moghaddam, B. P. Bailey, and C. Poon, "Ideatracker: an interactive visualization supporting collaboration and consensus building in online interface design discussions," in *IFIP Conference on Human-Computer Interaction*. Springer, 2011, pp. 259–276.

[8] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, "Peer and self assessment in massive online classes," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 6, p. 33, 2013.

[9] A. Pavel, D. B. Goldman, B. Hartmann, and M. Agrawala, "Vidcrit: video-based asynchronous video review," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 517–528.

[10] H. Wauck, Y.-C. G. Yen, W.-T. Fu, E. Gerber, S. P. Dow, and B. P. Bailey, "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, 2017, pp. 5580–5591.

[11] E. L. Glassman, A. Lin, C. J. Cai, and R. C. Miller, "Learnersourcing personalized hints," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016, pp. 1626–1636.

[12] E. L. Glassman and R. C. Miller, "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, 2016, pp. 37–40.

[13] S.-W. D. Li and P. Mitros, "Learnersourced recommendations for remediation," in *Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference on*. IEEE, 2015, pp. 411–412.

[14] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan, "Axis: Generating explanations at scale with learnersourcing and machine learning," in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 2016, pp. 379–388.

[15] P. Blowers, "Using student skill self-assessments to get balanced groups for group projects," *College Teaching*, vol. 51, no. 3, pp. 106–110, 2003.

[16] K. Vaccaro, D. Huang, M. Eslami, C. Sandvig, K. Hamilton, and K. Karahalios, "The illusion of control: Placebo effects of control settings," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 16.

[17] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-Adapted Interaction*, vol. 18, no. 5, p. 455, 2008.

[18] R. F. Kizilcec, "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, 2016, pp. 2390–2395.

[19] M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish, "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, 2015, pp. 1603–1612.

[20] J. A. Mello, "Improving individual member accountability in small work group settings," *Journal of management Education*, vol. 17, no. 2, pp. 253–259, 1993.

[21] J. B. Bear and A. W. Woolley, "The role of gender in team collaboration and performance," *Interdisciplinary science reviews*, vol. 36, no. 2, pp. 146–153, 2011.

[22] L. C. J. L. Brickell, L. C. D. B. Porter, L. C. M. F. Reynolds, and C. R. D. Cosgrove, "Assigning students to groups for engineering design projects: A comparison of five methods," *Journal of Engineering Education*, vol. 83, no. 3, pp. 259–262, 1994.

[23] S. K. Horwitz and I. B. Horwitz, "The effects of team diversity on team outcomes: A meta-analytic review of team demography," *Journal of management*, vol. 33, no. 6, pp. 987–1015, 2007.

[24] K. A. Jehn, G. B. Northcraft, and M. A. Neale, “Why differences make a difference: A field study of diversity, conflict and performance in workgroups,” *Administrative science quarterly*, vol. 44, no. 4, pp. 741–763, 1999.

[25] I. Lykourentzou, R. E. Kraut, and S. P. Dow, “Team dating leads to better online ad hoc collaborations.” in *CSCW*, 2017, pp. 2330–2343.

[26] I. Lykourentzou, S. Wang, R. E. Kraut, and S. P. Dow, “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, 2016, pp. 1243–1249.

[27] N. Salehi and M. S. Bernstein, “Hive: Collective design through network rotation,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 151, 2018.

[28] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, “Evidence for a collective intelligence factor in the performance of human groups,” *science*, vol. 330, no. 6004, pp. 686–688, 2010.

[29] I. Lykourentzou, A. Antoniou, Y. Naudet, and S. P. Dow, “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, 2016, pp. 260–273.

[30] M. Wen, K. Maki, X. Wang, S. Dow, J. D. Herbsleb, and C. P. Rosé, “Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses.” in *EDM*, 2016, pp. 533–538.

[31] D. R. Bacon, K. A. Stewart, and S. Stewart-Belle, “Exploring predictors of student team project performance,” *Journal of Marketing education*, vol. 20, no. 1, pp. 63–71, 1998.

[32] K. J. Chapman, M. Meuter, D. Toy, and L. Wright, “Cant we pick our own groups? the influence of group selection method on group dynamics and outcomes,” *Journal of Management Education*, vol. 30, no. 4, pp. 557–569, 2006.

[33] D. T. Conley and E. M. French, “Student ownership of learning as a key component of college readiness,” *American Behavioral Scientist*, vol. 58, no. 8, pp. 1018–1034, 2014.

[34] P. E. Chan, K. J. Graham-Day, V. A. Ressa, M. T. Peters, and M. Konrad, “Beyond involvement: Promoting student ownership of learning in classrooms,” *Intervention in School and Clinic*, vol. 50, no. 2, pp. 105–113, 2014. [Online]. Available: https://doi.org/10.1177/1053451214536039

[35] S. B. Feichtner and E. A. Davis, “Why some groups fail: A survey of students’ experiences with learning groups,” *Organizational Behavior Teaching Review*, vol. 9, no. 4, pp. 58–73, 1984.

[36] I. L. Janis, *Groupthink: Psychological studies of policy decisions and fiascoes.* Houghton Mifflin Boston, 1982, vol. 349.

[37] J. Surowiecki, *The wisdom of crowds.* Anchor, 2005.

[38] P. Mitros, "Learnersourcing of complex assessments," in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale.* ACM, 2015, pp. 317–320.

[39] "Catme smarter teamwork," 2018. [Online]. Available: http://info.catme.org/

[40] S. Kiesler, J. Siegel, and T. W. McGuire, "Social psychological aspects of computer-mediated communication." *American psychologist*, vol. 39, no. 10, p. 1123, 1984.

[41] J. Siegel, V. Dubrovsky, S. Kiesler, T. W. McGuire et al., "Group processes in computer-mediated communication," *Organizational behavior and human decision processes*, vol. 37, no. 2, pp. 157–187, 1986.

[42] R. B. Gallupe, L. M. Bastianutti, and W. H. Cooper, "Unblocking brainstorms." *Journal of applied psychology*, vol. 76, no. 1, p. 137, 1991.

[43] D. E. Brashers, M. Adkins, and R. A. Meyers, "Argumentation and computer-mediated group decision making," *Group communication in context: Studies of natural groups*, pp. 263–282, 1994.

[44] R. Nadeau, E. Cloutier, and J.-H. Guay, "New evidence about the existence of a bandwagon effect in the opinion formation process," *International Political Science Review*, vol. 14, no. 2, pp. 203–213, 1993.

[45] "Piazza homepage," 2018. [Online]. Available: https://piazza.com/signup

[46] S. Rice and M. N. Gregor, *E-Learning and the Academic Library: Essays on Innovative Initiatives.* McFarland, 2016.

[47] J. Saldaña, *The coding manual for qualitative researchers.* Sage, 2015.

[48] L. Vygotsky, "Zone of proximal development," *Mind in society: The development of higher psycological processes*, vol. 5291, p. 157, 1987.

[49] E. M. Hastings, F. Jahanbakhsh, K. Karahalios, D. Marinov, and B. P. Bailey, "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, no. CSCW. ACM, 2018.

[50] A. A. Alamri and B. P. Bailey, "Examination of the effectiveness of a criteria-based team formation tool," in *Frontiers in Education.* IEEE, 2018.

[51] E. Aronson, "Chapter 10 - building empathy, compassion, and achievement in the jigsaw classroom," in *Improving Academic Achievement*, ser. Educational Psychology, J. Aronson, Ed. San Diego: Academic Press, 2002, pp. 209 – 225. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780120644551500130

[52] S. Mahenthiran and P. J. Rouse, "The impact of group selection on student performance and satisfaction," *International Journal of Educational Management*, vol. 14, no. 6, pp. 255–265, 2000.

APPENDIX A: STUDENT CRITERIA VOTE DISTRIBUTIONS

Following are the distributions of votes for criteria discussed in the four courses which were not included in the paper. Recall that 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.

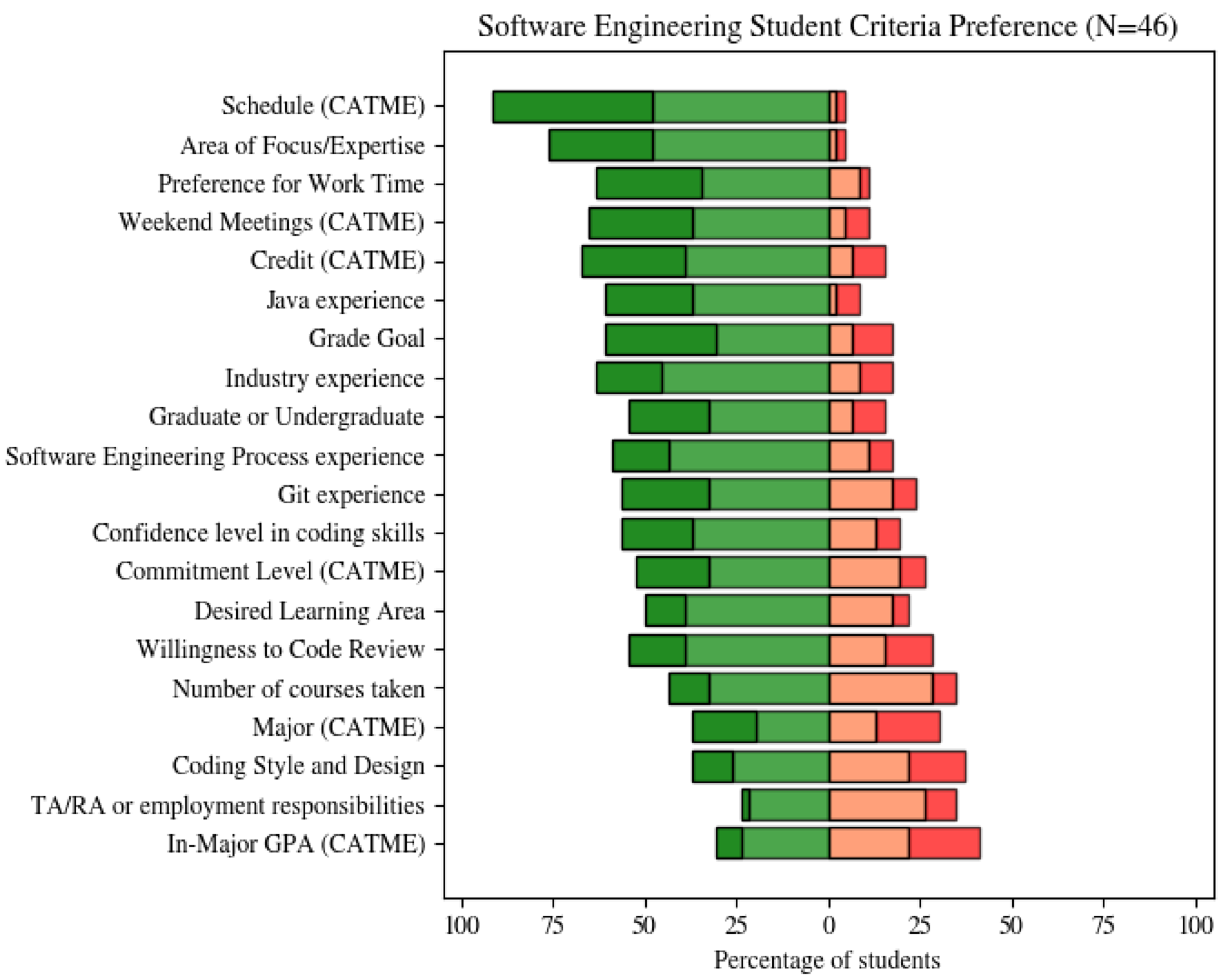

Figure A.1: Software Engineering

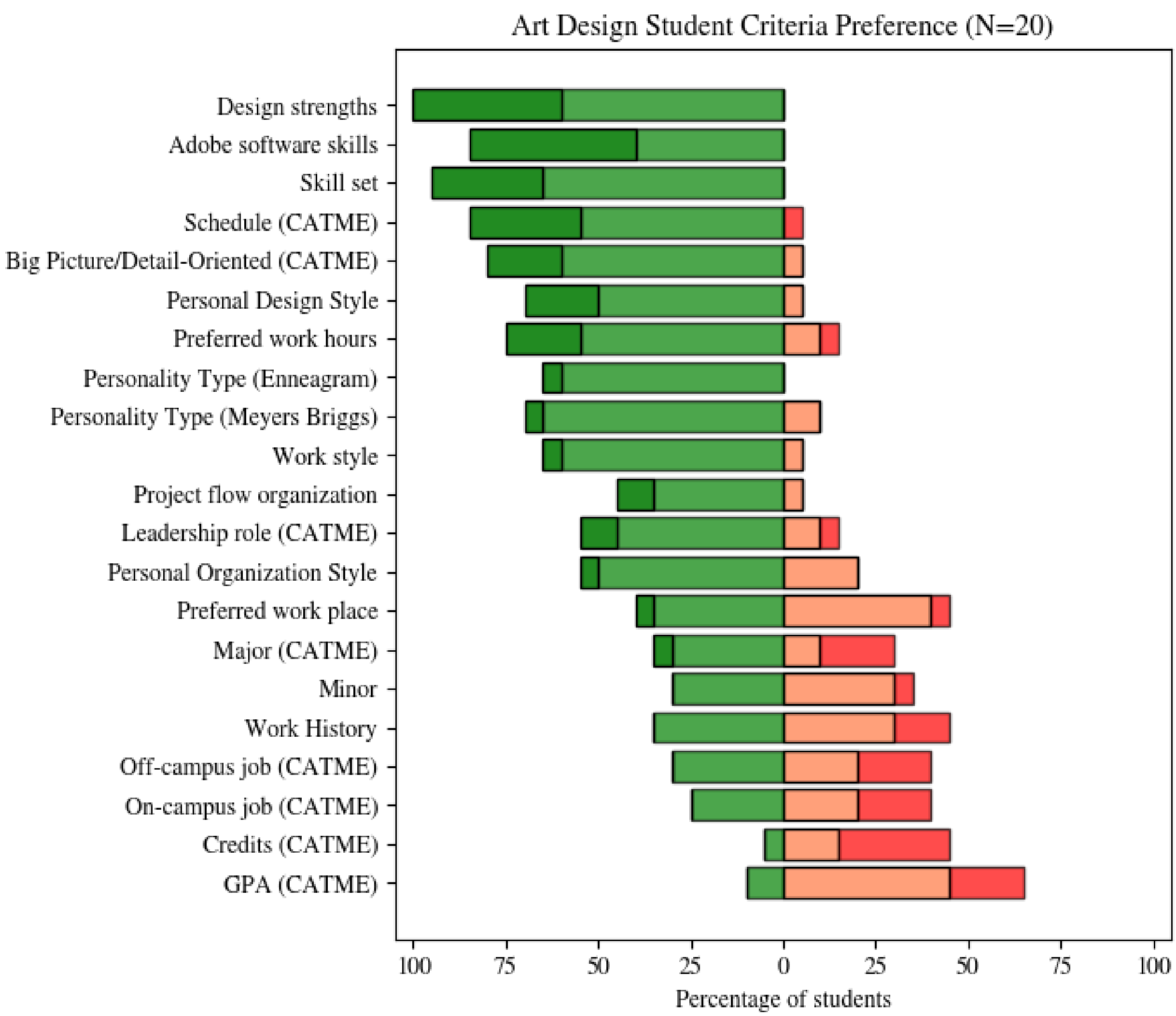

Figure A.2: Design Methods

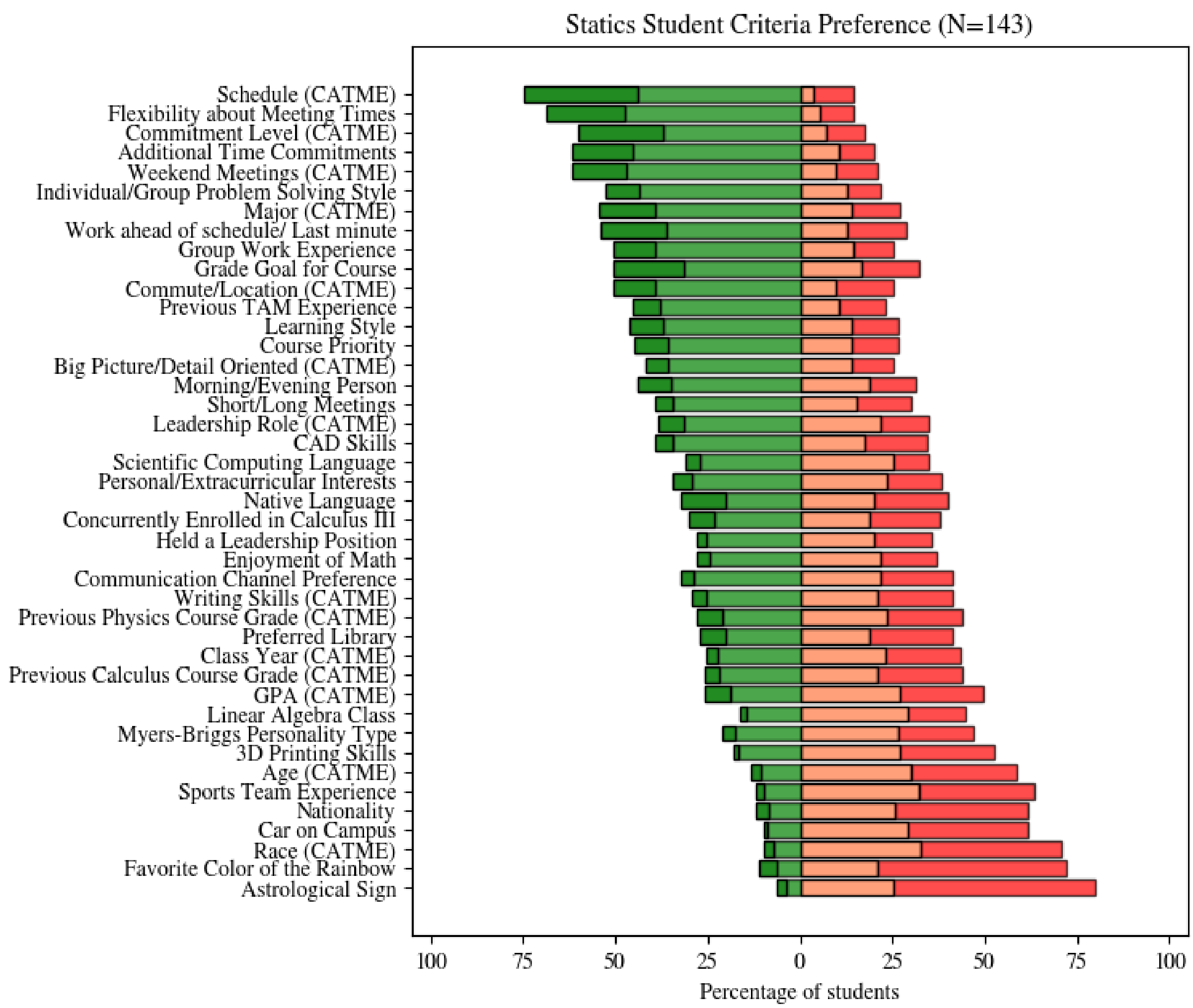

Figure A.3: Statics

Figure A.4: Design for Manufacturability

APPENDIX B: FINAL STUDENT CRITERIA CONFIGURATIONS

Following are the final criteria and weights selected by students in the four courses which were not included in the paper. Recall that Weights range from -5 to 5, with the sign indicating similarity (+) or dissimilarity (-) of team members.

Table B.1: Software Engineering

Criterion	Weight
Schedule	3
Preferred Work Time (e.g., morning, evening, etc.)	3
Weekend meetings	3
Taking course for 3 or 4 credits	2
Focus or experience area	-3

Table B.2: Design Methods

Criterion	Weight
Schedule	3
Morning vs. evening person	3
Leadership role	-2
Big picture vs. detail-oriented	-2
Work habits (i.e., last minute or work ahead)	3
Adobe software skill	-3
Design skills	-3
Design style	3
MBTI personality type	2
Enneagram personality type	1
Preferred project management software	1
Personal organizational style	2

Table B.3: Statics

Criterion	Weight
Schedule	3
Weekend meetings	3
Extracurricular activities	2
Flexibility about meeting times	3
Commitment level	3

Table B.4: Design for Manufacturability

Criterion	Weight
Schedule	4
Big picture vs. detail-oriented	-2
CAD skills	-3
Weekend meetings	3
Course priority	3
Scheduling in advance	3
Flexibility about meeting times	2
Commute	1
Expected grade	1
Experience with teams	1
Leadership role	-2
ColorCode personality type	-2
Hands-on skills	-3
Shop skills	-3
Presentation comfort	-2
Major	2
Personal interests	2
Engineering interests	-2
Number of extracurriculars	2