Composing Team Compositions: An Examination of Instructors' Current Algorithmic Team Formation Practices

EMILY M. HASTINGS, University of Wisconsin-Eau Claire, USA and University of Illinois at Urbana-Champaign, USA
VIDUSHI OJHA, University of Illinois at Urbana-Champaign, USA
BENEDICT V. AUSTRIACO, University of Illinois at Urbana-Champaign, USA
KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA
BRIAN P. BAILEY, University of Illinois at Urbana-Champaign, USA

Instructors using algorithmic team formation tools must decide which criteria (e.g., skills, demographics, etc.) to use to group students into teams based on their teamwork goals, and have many possible sources from which to draw these configurations (e.g., the literature, other faculty, their students, etc.). However, tools offer considerable flexibility and selecting ineffective configurations can lead to teams that do not collaborate successfully. Due to such tools' relative novelty, there is currently little knowledge of how instructors choose which of these sources to utilize, how they relate different criteria to their goals for the planned teamwork, or how they determine if their configuration or the generated teams are successful. To close this gap, we conducted a survey (N=77) and interview (N=21) study of instructors using CATME Team-Maker and other criteria-based processes to investigate instructors' goals and decisions when using team formation tools. The results showed that instructors prioritized students learning to work with diverse teammates and performed "sanity checks" on their formation approach's output to ensure that the generated teams would support this goal, especially focusing on criteria like gender and race. However, they sometimes struggled to relate their educational goals to specific settings in the tool. In general, they also did not solicit any input from students when configuring the tool, despite acknowledging that this information might be useful. By opening the "black box" of the algorithm to students, more learner-centered approaches to forming teams could therefore be a promising way to provide more support to instructors configuring algorithmic tools while at the same time supporting student agency and learning about teamwork.

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

Additional Key Words and Phrases: Algorithms, CATME, Team formation, Team composition, Collaborative learning

ACM Reference Format:
Emily M. Hastings, Vidushi Ojha, Benedict V. Austriaco, Karrie Karahalios, and Brian P. Bailey. 2023. Composing Team Compositions: An Examination of Instructors' Current Algorithmic Team Formation Practices. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 305 (October 2023), 24 pages. https://doi.org/10.1145/3610096

Authors' addresses: Emily M. Hastings, University of Wisconsin-Eau Claire, 105 Garfield Avenue, Eau Claire, WI, 54702, USA and University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL, 61801, USA, hastinem@uwec.edu; Vidushi Ojha, vojha3@illinois.edu, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL, 61801, USA; Benedict V. Austriaco, bv2@illinois.edu, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL, 61801, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL, 61801, USA; Brian P. Bailey, bpbailey@illinois.edu, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL, 61801, USA.

2573-0142/2023/10-ART305 $15.00
https://doi.org/10.1145/3610096

1 INTRODUCTION

Algorithmic team formation tools (e.g., [2, 32, 65]) offer many benefits for instructors implementing team-based learning, and are in use at institutions around the world, serving many thousands of students [3, 4]. These tools create teams based on criteria such as skills, demographics, and other characteristics of potential team members and can reduce stress and burden for both instructors and students [37]. Criteria-based approaches have also been shown in prior work to produce teams that are higher-performing [15, 65] or have more desirable attributes [38] than those formed with random assignment or self-selection, when a small number of specific criteria are used. These benefits are especially important as courses increase in size and move to online delivery methods, where instructors cannot always know students well enough to manually create teams, or do so feasibly at scale.

Configuring algorithmic tools (i.e., selecting the criteria according to which student teams should be formed and how these traits should be distributed among teams) is a difficult problem, because these tools offer considerable flexibility, and the scope of possible configurations of criteria and weights is expansive. For example, the widely-used tool CATME Team-Maker [2, 44] offers 27 criteria by default, each with 11 possible weight values, for a total of 11^{27} possible configurations, not including any custom criteria the instructor might add. When developing a configuration, instructors have many options for sources from which they could draw inspiration. They can consult the literature on team formation [12, 13, 17, 27, 34] on which these tools are based to help make their decisions. For example, research suggests that students benefit from being in groups that have high skill diversity [34] and certain personality traits, such as agreeableness and openness to experience [12]. Instructors might also rely on their own previous experience forming teams, consult their peers for guidance, or even solicit input from their students to gain greater insight into which criteria matter most to them [29, 31]. However, not much is known about how instructors are currently approaching this process or making their decisions. Additionally, it is not clear how instructors' goals for the teamwork affect their decisions in the tool or how they assess whether the generated teams satisfy their goals. Closing this knowledge gap is important, because making poor configuration choices due to a lack of understanding of the tool or which criteria are best can generate ineffective teams and harm students' learning experiences. For example, the configurations shared in prior work suggest that instructors tend to select configurations containing many criteria at once [29, 30, 37]. These multi-criteria configurations may not lead to the same benefits in terms of team performance and satisfaction as the more focused individual criteria studied in the literature on which these tools are based, and could even be detrimental to teams.

To address these challenges, we studied instructors' current criteria selection practices for forming teams in authentic course settings. We investigate their broader goals for team formation and how these impact their process; report which and how many criteria they select and how they make these selections; and explore their understandings of their tool or process and how these beliefs impact their approaches or configurations, focusing especially on differences between instructors who do and do not use algorithmic tools. This information is collected via an online survey with two versions, the first restricted to users of representative tool CATME Team-Maker, and the second open to instructors who use any criteria-based approach to form teams (e.g., other tools, card-sorting, or spreadsheet-based methods). This survey gathered criteria configurations from more instructors and a broader range of institutions and academic areas than prior work such as [29, 30, 37]. We additionally conducted semi-structured interviews with a subset of the survey participants to gather richer information about their goals and team formation approach, including aspects such as their understanding of their tool or process and how they evaluated its output.

We received 77 survey responses and conducted 21 followup interviews, with participants from seven countries and a variety of academic departments, ranging from Computer Science to Recreation, Sport and Tourism. We found that instructors' primary goals for the teamwork in their courses were supporting peer-based learning and the development of students' teamwork skills, as well as preparing students for future careers in industry. These goals drove instructors to prioritize various aspects of diversity (e.g., gender, race, academic area, etc.) in the criteria they used to form teams, although some reported struggling to relate their goals to specific aspects of a configuration. On average, instructors using CATME selected combinations of 9-10 of the 27 available default criteria, including both practical scheduling-related concerns and those related to promoting diversity. However, the configurations themselves varied widely in terms of which criteria were selected and how many were chosen. Participants not using the tool reported creating significantly smaller configurations (2-3 criteria on average) of similar criteria, although with less of a focus on scheduling. Instructors expressed confusion about how the tool worked but offered their own interpretations such as the existence of a limit on the number of criteria that should be included in a single configuration, and described some surprising behaviors, such as intentionally obfuscating from students which criteria would ultimately be used. We also observed that instructors frequently desire to "sanity check" generated teams to ensure that their configuration goals are met, but struggle to do so when the representation of a tool's output does not match their mental models of a desired team. Our results suggest that the use of an algorithmic tool impacted the way instructors approached forming teams in their courses– for example, through the default criteria provided– and in some cases constrained instructors' processes or led them to perform additional work manually. Tool designers must therefore carefully consider any defaults in the interfaces they design, and be cognizant of how a tool and its interface can shape process and help instructors more effectively achieve their educational goals.

In this work, we contribute to the CSCW community (1) insights into the current practices of instructors using criteria-based team formation, and (2) practical implications for how designers of algorithmic team formation tools can create interfaces that are responsive to instructors' actual processes and provide more guidance in areas where they are unsure. For example, we found that instructors were sometimes uncertain about whether the configurations they selected would be effective or accepted by students, and while they valued students' opinions on what criteria to prioritize, rarely solicited this information themselves. Tools might therefore provide built-in ways to gather student input and help instructors interpret it. Including more explicit recommendations from the literature, especially in relation to the mapping between instructors' goals and specific criteria, may also help clarify the configuration process. The inclusion of this additional support may in turn produce student teams that cooperate more effectively and have better learning experiences.

2 RELATED WORK

We describe how our work contributes to the prior literature on team composition and algorithmic team formation.

2.1 Team Composition and Criteria-based Team Formation

Prior work has shown that facets of the composition of a team can impact the team's outcomes. For instance, elements such as the number of women on a team [10, 69] and certain distributions of represented personality types [12, 47] and task skills [34] can increase team performance relative to teams lacking those compositions. Team formation approaches grounded in this literature therefore select potential members for teams according to these and other criteria.

The options for which criteria to use when forming teams are numerous, with more being tested all the time. For example, in addition to criteria like demographics, skills, and work styles, prior work

has investigated "speed dating" techniques [48, 49] and using member social connections [24, 58] or transactivity [66, 67], as well as various other approaches incorporating potential team members' preferences or elements of previous interactions. This proliferation of potential criteria coupled with the fact that existing work often focuses on the effects of singular or small combinations of criteria at a time, and often in settings outside of authentic courses [30, 42, 50], will make an already difficult problem (deciding how to form teams) even more complex over time. Given the expansive set of options, it is likely that many possible configurations of criteria that instructors could select when forming teams in their courses are ineffective and might lead to poor team outcomes for their students. In addition, these configurations may not align with the values of the students in the course, who have the most at stake in the team formation process. For example, Hastings et al. [29, 31] found that students tend to prefer criteria related to logistic issues like scheduling over demographic criteria such as gender and race. Instructors may therefore benefit from additional guidance to help them make the most beneficial decisions for their specific contexts.

We contribute to the body of work on team composition by reporting which criteria (and combinations of criteria) instructors are using in practice to form teams in their courses, as well as the factors they consider when making these decisions and how their choices are affected by the use of a dedicated team formation tool. This greater empirical knowledge of criteria-based team formation in authentic settings can help provide designers of team formation tools greater insight into what kind of guidance would be most helpful to instructors and ultimately benefit students. For instance, it may be useful to build features for gathering student input on criteria directly into team formation interfaces in order to augment instructors' existing processes.

2.2 Algorithmic Team Formation Tools

As course enrollments grow large, it becomes more difficult for instructors to implement a criteria-based approach to team formation by hand. Online learning environments such as massive open online courses (MOOCs) can suffer from similar problems, both due to the scale of the course and the limited information available about students [66]. To address these issues, instructors can turn to algorithmic tools to facilitate and automate the team formation process. Not only do these tools decrease stress and burden for both instructors and their students [37], they have the potential to optimize team assignments better than the instructor with respect to a given set of criteria [44].

Many tools and algorithms exist for team formation. In their review, Gómez-Zará, DeChurch and Contractor describe a taxonomy to categorize these according to two axes: users' *agency* and users' *participation* [22]. In this paper, we primarily study the algorithmic team formation tool Comprehensive Assessment for Team-Member Effectiveness (CATME) Team-Maker [2, 44]. Team-Maker is a popular tool used in many universities [4] that falls into the "staffed teams" quadrant of this taxonomy (high agency, low participation) since users select the criteria that will be used by the tool to form teams, drawing from a set of 27 default criteria or personalized options created by users. A greedy randomized algorithm then creates teams based on the selected configurations. DIANA [65] and groupformation.org [32] are additional examples of algorithmic team formation tools that create staffed teams. Other tools and algorithms described in prior work include [9, 11, 35, 67], although not all of these have publicly available implementations or user interfaces to allow instructors who may not be skilled programmers or mathematicians to use them effectively in their courses [35].

The use of algorithmic tools introduces new challenges, including placing the burden on instructors creating staffed teams to configure the tool to be responsive to both their own teamwork goals and the needs of students in the course. Instructors must think carefully about exactly which and how many criteria they plan to use when forming teams, how students should be grouped relative to these criteria (i.e., should similar or different students be placed on the same team?), and how

criteria should be weighed relative to each other. Current tools provide considerable flexibility but do not offer much in the way of guidance for instructors making these decisions, which can lead them to make selections that may not offer the same benefits as simpler configurations described in the literature [30]. In addition, tools often rely on input from survey responses of uncertain quality self-reported by students [6], while at the same time giving students limited input into the team formation process, despite its importance to their learning experiences [29, 31].

Our work expands current knowledge of how instructors are using and making decisions with a representative algorithmic team formation tool in practice, including how they source criteria configurations, their understanding of the tool's algorithm, and the ways in which the tool constrains or aids their process. This greater understanding will be instrumental in helping tool designers develop features to guide and support instructor decision-making, and otherwise make team formation tools more usable. For example, tools might consider instructors' goals for the teamwork and suggest relevant criteria or weights from the literature to help support these aspects of students' collaboration.

3 RESEARCH QUESTIONS

The study addresses the following research questions:

(1) What goals do instructors have for the teamwork in their courses and how are the goals reflected in their team formation process?
(2) Which and how many criteria do instructors use to form teams in their courses, and how do they select these configurations?
(3) How do instructors understand the use of team formation tools, and how does this understanding affect their configurations?

Answering these questions will lead to increased knowledge of current instructor practices related to criteria-based team formation, and present opportunities for improving algorithmic team formation tools. These improvements may lead to more positive student teamwork experiences.

4 METHOD

To answer our research questions, we conducted an online survey of instructors who use criteria-based approaches to form student teams in their courses. In order to explore whether instructors' decisions were impacted by the use of an algorithmic team formation tool, there were two versions of the survey: one focused on instructors who use the representative algorithmic tool CATME Team-Maker [44], and another open to instructors using any criteria-based approach. The latter could include the use of other tools, spreadsheets, or more physical methods like card-sorting. For example, one participant described printing copies of students' resumes and sorting them into piles according to the relevant criteria. Others mentioned gathering student project preferences through sign-ups on physical or virtual whiteboards, gathering student demographic or academic information from prior assignments or paper surveys, or forming teams based on students' ranking in the course. We then conducted followup interviews with a subset of the participants to provide more context to their responses. The study was approved by the Institutional Review Board at our university.

4.1 The Algorithmic Team Formation Tool

For the specific version of our survey, we wanted to focus on a single algorithmic team formation tool in order to facilitate the comparison of criteria configurations and decision processes from different instructors. We selected CATME Team-Maker as a representative tool because of its widespread use and coverage in prior literature (e.g., [6, 29, 30, 37]), as well as its basis in team

formation research [44]. The CATME system (including Team-Maker and several other components) has an expansive user base, spanning more than 22,000 instructors across 88 countries [4]. As such, it is the only team formation tool we are aware of that could offer such a large pool of potential participants.

4.1.1 Team-Maker Workflow. Team-Maker allows instructors to select the criteria according to which student teams should be formed, as well as weigh the importance of criteria relative to each other. The tool provides a set of default criteria (such as Gender, Schedule, GPA, and other commonly selected items), but users can also enter their own custom criteria. For each selected criterion, the instructor assigns a weight between −5 and 5, where negative weights configure the algorithm to group students by dissimilarity on that criterion, and positive weights by similarity, with the magnitude signifying the strength of the preference. For example, an instructor might assign "Weekend Meetings" a weight of 5 to strongly prefer teams formed of students with similar weekend availability, and assign "Software Skills" a −3 to moderately prefer teams with a mixture of software experience. Once the criteria are finalized, students are asked to fill out a survey indicating their schedule, prior experience, or demographics, depending on the criteria selected. With this information, the tool uses the assigned weights and a randomized greedy algorithm to form teams [1]. Instructors can choose to rerun the algorithm until they are satisfied with the resulting teams. The tool then informs the students of their assigned teams and how to contact their teammates.

4.2 Recruitment

Recruitment was accomplished through a variety of approaches, primarily email-based. Note that we were not able to reach the entire population of CATME users, because their contact information is not publicly available. At our institution, we advertised the CATME-specific version of the survey on the weekly faculty and staff e-newsletter once per term for the duration of the study (the maximum number of posts allowed). Additionally, we worked with the university to compile a list of faculty who had used the tool over the last three years, in order to contact these users directly via email. We also reached out to instructors who had previously expressed interest in our work to encourage them to participate.

Outside our university, we reached out to colleagues at institutions publicly listed in the tool's user base [3] to encourage them to take the survey if applicable and share it with anyone they thought would be interested. We also contacted relevant offices (e.g., those concerned with classroom software or the advancement of teaching methods) at some of these institutions to share the survey. Additionally, we sent messages advertising the study to relevant email distribution lists, specifically the Association of Computing Machinery (ACM) Special Interest Groups on Computer Science Education (SIGCSE) and Computer-Human Interaction (SIGCHI), and created posts sharing the survey on the CATME Users LinkedIn group[1]. Finally, we also conducted several searches online to find academic papers and course syllabi mentioning the use of the tool, and contacted the associated authors and instructors to share the survey with them, directly reaching out to 63 people in total.

To distribute the general version of the survey, we again sent messages advertising the study to the SIGCSE and SIGCHI mailing lists, and advertised the survey on our university's weekly faculty and staff e-newsletter.

4.3 Survey Procedure

Instructors who agreed to participate in the study (via following the survey link distributed via email or online) completed the survey on the platform provided by our university's Public Affairs office. The survey had several sections:

[1] https://www.linkedin.com/groups/4188510/

(1) Consent form.
(2) Details about a course in which they form teams using a criteria-based approach (e.g., course size, topic, etc.).
(3) Details about their criteria configuration for that course. Participants who used CATME were instructed to download an .html file from the tool's website showing the criteria and weights they select. We asked participants responding to the general version of the survey to explain their process and provide any example file that could help illustrate their approach. This section also contained additional Likert items and free text questions about how and why the instructors selected these criteria.
(4) Details and demographic information about the participant (e.g., familiarity with the team formation literature, experience using their current team formation approach, etc.).
(5) Followup. Participants indicated whether they were willing to be contacted by the researchers to answer followup questions about their responses, and if so, provided an email address.

We expected the survey to take users 15-30 minutes to complete. Participants received $20 in compensation via PayPal for completing the survey.[2]

Since participants responding to the general version of the survey did not have a uniform method of sharing their criteria configurations (i.e., the .html file for the tool users), it was necessary to compile these manually. Two members of the research team read through the descriptions of the team formation processes used by these participants. They then collaboratively extracted [28, 57] the criteria mentioned, along with details about the distribution of students relative to these criteria (e.g., students might be grouped by similar or dissimilar GPA), when available.

The survey questions were intended to answer RQ2 and were prompted by prior work such as [29, 30, 37] reporting instructor configurations that contained many criteria at once, which may not create the same benefits in terms of team outcomes as more focused selections [30]. Team formation tools make it easy to select complex configurations such as these, so gaining greater insight into how instructors more broadly are selecting criteria and weights offers a significant opportunity to improve on a potential shortcoming of algorithmic tools. The questions were refined through an iterative pilot process involving pretesting [52] with several instructors who had previously used CATME.

4.4 Instructor Interviews

To elicit additional information to provide richer context for the survey responses and answer RQ1 and RQ3, we reached out to those participants who had agreed to be contacted in the final section of the survey to ask for their participation in two further study activities: a brief followup survey, and semi-structured interviews.[3] The survey asked instructors to select their primary goals for including teamwork in their courses from a given list with an option to write in their own, and to respond to two seven-point Likert items gauging their perceptions of the success of their approach: (a) "My team formation approach (e.g., CATME or another tool, card sorting, etc.) is accomplishing these goals." and (b) "My specific set of team formation criteria (e.g., schedule, gender, GPA, etc.) is accomplishing these goals." (1=Strongly disagree, 7=Strongly agree). Instructors also had a chance to leave additional comments if desired.

The interviews focused on instructors' goals for using teamwork, the ways in which these goals affected their process, the factors affecting the development of their configuration over time, and their understanding and use of various features of the team formation tool (for instructors who used it). Interview participants completed an additional consent form and were compensated $50 for

[2]We have included the complete survey as supplemental material.
[3]We have included the followup survey and interview script as supplemental material.

	CATME			General		
	Avg.	Min.	Max.	Avg.	Min.	Max.
Times Teaching Course	5+	1-2	5+	1-2	1-2	5+
Times Using Current Formation Method	1-2	1-2	5+	5+	1-2	5+
Num. Students in Course	21-50	1-20	201-500	21-50	1-20	201-500
Literature Familiarity (out of 7)	4	1	7	4	1	6
Num. Identifying as URG	19 (out of 46)			12 (out of 31)		

Table 1. Participant information for both versions of the survey. "URG" refers to being a member of a group that is underrepresented in the participant's academic area. "Avg." column shows the mode for categorical questions where participants selected from given ranges and the median for Likert item.

their time. Interviews lasted from 30 to 60 minutes and were conducted on Zoom. Audio recordings of the interviews were transcribed by the service Rev.com.

Two members of the research team performed open coding on the transcripts to identify major themes. First, the first author developed a tentative coding scheme based on the survey and an initial reading of all the interviews, and further refined it while coding a sample of 20% of the interviews. Both coders then independently applied the scheme to a further 25% of the interviews, iterating on the scheme and revising until satisfactory agreement was reached (median Cohen's kappa over all codes was 0.66). The two coders then divided the remaining interviews (including the initial sample) and coded them using the final scheme, discussing as necessary to resolve uncertainties. [4]

5 RESULTS

We received 77 responses to our survey, 46 of which were for the CATME-specific version, and 31 for the general version. Of the 46 respondents who reported using the algorithmic tool, 30 provided uploads of their criteria configurations.

The instructors came from a variety of academic departments (including Computer Science, Economics, and Business Administration), and were primarily US-based (68), but several reported teaching at institutions in other countries: the UK (2), Malaysia (2), India (1), Australia (1), Canada (1), and Ireland (1). Most of our respondents (50) reported teaching courses in the area of Formal and Applied Sciences, with a variety of course sizes. The instructors reported multiple purposes for the teams in their courses, with the most popular being long-term projects. Most instructors (53) selected criteria and formed the teams themselves, but some (7) delegated the team formation task to teaching assistants, and in six cases, to the students in the course. See Table 1 for more information about the participants.

25 participants further completed the followup survey, and 21 agreed to be interviewed (10 CATME, 11 other). In the following sections, these results are denoted with "(N=x)", where x is the number of participants who mentioned a topic or selected an answer. Comments from individual interview participants are denoted with "P" + a numerical identifier.

See Figure 1 for an overview of the team formation process described by participants, which is discussed in more detail in the following sections.

[4]We have included the complete list of codes and their frequencies as supplemental material.

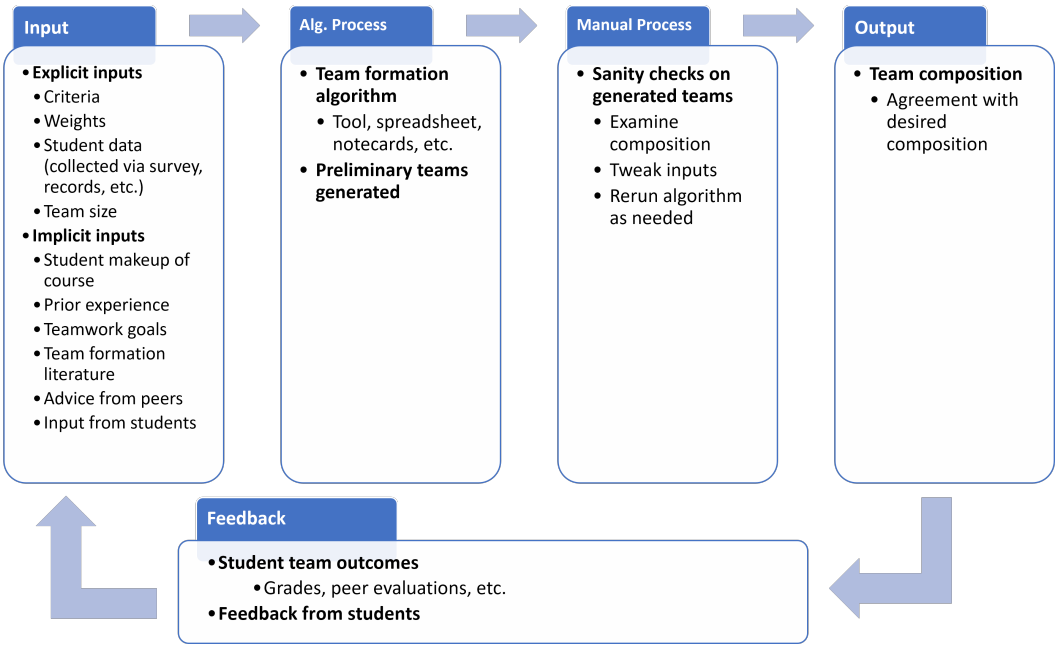

Fig. 1. An input-process-output diagram of the team formation process described by instructors in the study. Teams are formed based on both explicit and implicit inputs using a process incorporating algorithmic and manual elements, and the results of the team formation influence future iterations of the process.

5.1 Instructors Prioritize Teamwork Skills, Peer-based Learning, and Preparing Students for Industry

The goal or reason for using teamwork that was most frequently mentioned by instructors in the interviews was teaching students teamwork skills (N=15). Supporting peer-based learning (N=14) and preparing students for future careers in industry (N=14) were the next most frequently described, followed by promoting learning of the course content (N=10). Other reasons instructors cited for including teamwork in their course included fulfilling external requirements such as those posed by the Accreditation Board for Engineering and Technology (ABET [5]; N=6), reducing grading (N=5), promoting higher quality projects (N=4), building social bonds among students (N=4), and continuing processes that had been in place when they took on the course (N=2). We did not observe any patterns in the goals cited by instructors using the algorithmic tool as compared to those using other approaches.

These results from the interviews largely agreed with those of the followup survey, where instructors cited promoting the learning of course content and supporting peer-based learning most frequently (N=21 for both), followed by promoting teamwork skills (N=19) and preparing for industry (N=16). These were followed by promoting higher quality outcomes (N=9) and reducing grading (N=8). It is important to note that these goals are not mutually exclusive. Instructors described having multiple goals in both the interviews and the followup survey, and might have, for example, chosen to prioritize teaching teamwork skills in order to prepare students for industry [64]. A chi-square analysis did not reveal a statistically significant difference in the pattern of goals between instructors using the tool or other methods ($\chi^2(6)=4.94$, $p=0.55$).

When asked how these goals for the teamwork affected their approach to team formation in the course, the most common answer was that instructors prioritized placing diverse students together (e.g., different genders, academic areas, etc.; N=13). For example, two instructors who had prioritized preparing students for industry explained:

> *"The nice thing is I'm able to say to them, 'Look, you have to do this in the real world... You're going to get a job. You're going to go work, and you have no say in what group that you're in, so you better be able to do that.' There's an implicit skill that we're sort of teaching through this."* (P3)

> *"So I'm trying to give them the diversity of working with others. At the same time, hold them accountable so that when they're out in the industry, they're used to the dynamics of having to work with others."* (P19)

A few instructors also described placing similar students together (N=2) or not directly considering the goals when forming teams (N=2).

5.1.1 Instructors Use Algorithmic Tools to Streamline Team Formation. Beyond their general goals for the teamwork, we also examined the other reasons instructors reported for why they use the team formation process that they do. One survey question asked instructors to rank their reasons for using their current method to form teams from 1 (most important) to 6 (least important), using possibilities drawn from Jahanbakhsh et al. [37]. We summed these scores across participants to generate the following ranking, shown from most important to least important (score shown after each in parentheses for the CATME-specific version of the survey):

(1) Streamline team formation process for yourself as the instructor (108)
(2) Create teams with an increased chance to produce successful outcomes (151)
(3) Create a fair and consistent experience for students (155)
(4) Tool has a good theoretical background and is based on rational criteria (168)
(5) Students learn to work with unfamiliar people (170)
(6) Reduce stress and burden for students (206)

For instructors responding to the general version of the survey, we saw a similar ranking, but the most important reason reported by instructors who use the tool (streamlining the process) was now at the bottom of the ranking. This finding makes sense, because instructors forming teams by hand frequently must put in more effort to create the teams than they would using a tool.

The interviews revealed similar reasons for using particular approaches, especially for those instructors using the algorithmic tool. Many mentioned that they valued specific features of the tool (N=9) such as the ability to easily group by schedule, as well as the more general convenience it offers (N=8). The ability to blame the tool [37] and have "plausible deniability" (P20) if students took issue with the team formation process was also commonly mentioned (N=5), as were issues of how well the process did (or did not) reflect common team formation practices in industry (N=5). Several instructors mentioned other reasons for using their approach, including recommendations from other faculty (N=3), the literature (N=2), and requests from students (N=2).

5.2 Little Consensus on Criteria and Weights

Across both versions of the survey, instructors reported using a total of 87 distinct criteria, including both default criteria from the tool, as well as custom criteria. Using the configurations provided by respondents to the CATME-specific version (N=30), we extracted the top 20 most popular criteria, along with their associated weights (see Figure 2). Although criteria like Gender, Schedule, and Race appeared in most instructors' configurations, there were few obvious patterns in the weights for any criteria (with the exception of Schedule, which the tool limited to positive values only).

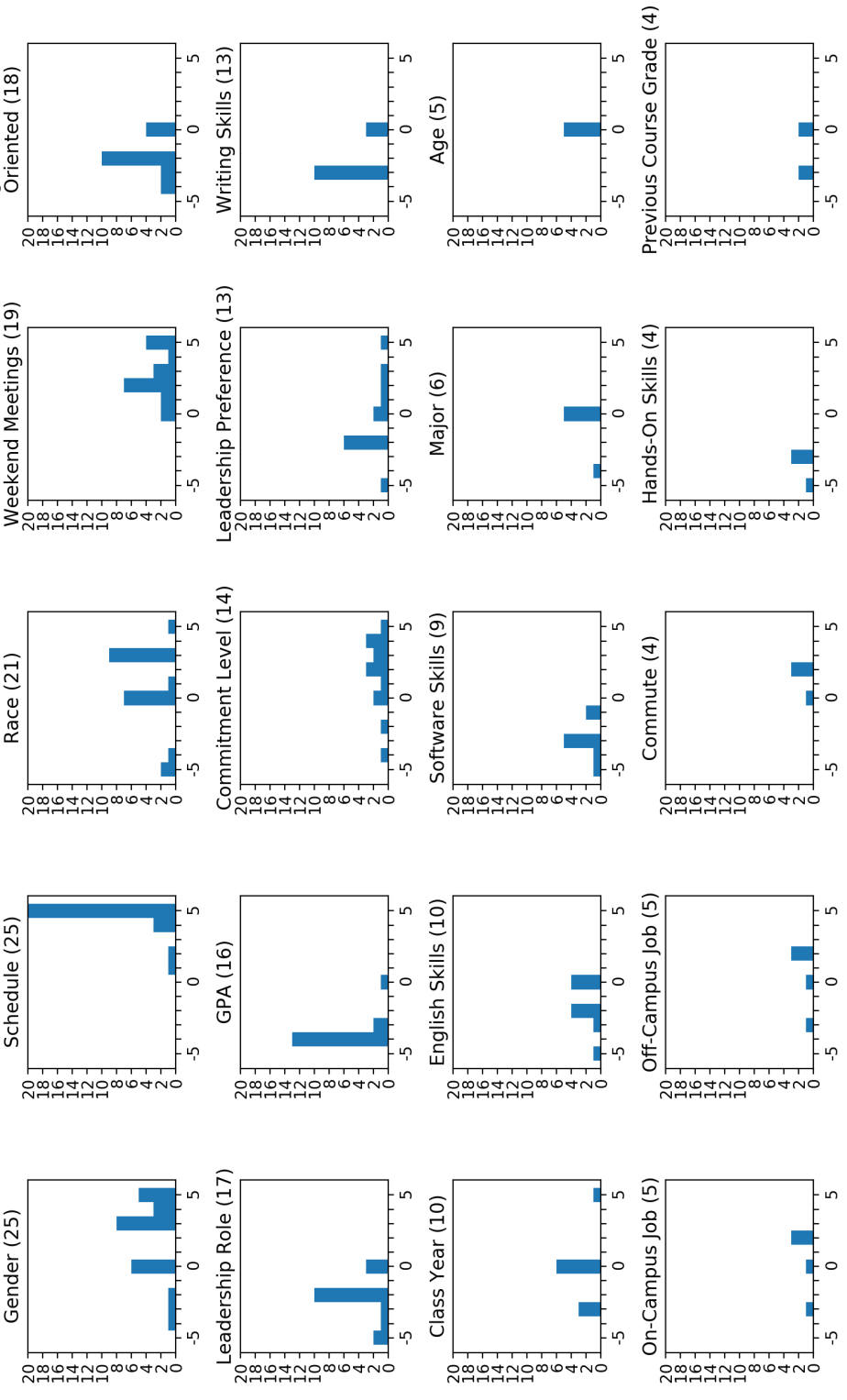

Fig. 2. Top 20 most popular criteria used by instructors in the CATME-specific survey, and their associated weights. The parenthetical number represents the number of times that criterion appeared in instructor configurations. The x-axis in each histogram shows the weights assigned to the criterion, while the y-axis shows the count of instructors who assigned that weight. Note that all the top 20 most popular criteria are default selections in the tool. There is little consensus for most criteria on the weights that should be assigned.

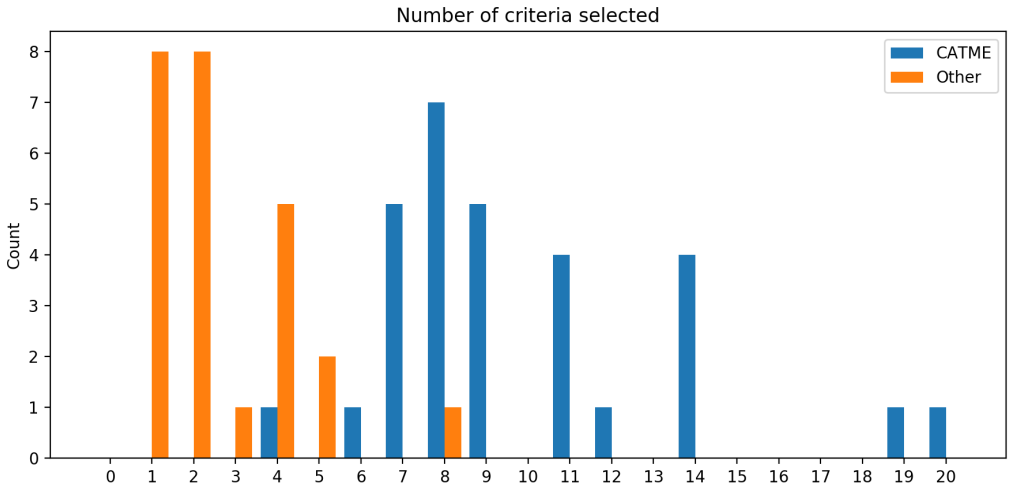

Fig. 3. The number of criteria selected by instructors in their team formation process.

For those instructors who chose criteria not included by default in the tool, they seemed to use these custom criteria either to get specific information about their students' experience (e.g. GitHub, grade in a specific preceding course) or their preferences for projects. Eight instructors using CATME noted including custom criteria for these purposes. On average, participants in the CATME-specific version selected 9.9 criteria at a time in their configurations. See Figure 3 for a distribution of configuration sizes.

We found that respondents to the general version of the survey chose similar criteria to those using the tool, including Gender, Race, and GPA. Student preferences for project topics were also commonly used as a criterion, similar to the instructors using the tool, who often included these preferences as custom criteria. In contrast, however, we found that only one instructor mentioned scheduling constraints, which was one of the most common criteria among tool users. This difference makes sense, as managing logistic issues such as schedule can be very difficult to do by hand. Similarly, we noted that these respondents reported using significantly fewer criteria in their configurations, an average of 2.6 (Mann-Whitney-Wilcoxon W=765.5, p=0.00). This result is in line with comments made in the survey by a number of instructors that their approaches may not scale as easily as tool-based team formation for larger or online courses.

5.2.1 Instructors Consider Students When Selecting Criteria But Do Not Solicit Their Input. For instructors who responded to the CATME-specific version of the survey, the factor they considered most frequently when determining their configuration was the student makeup of the course (i.e., the specific population of students and their characteristics), possibly due to this information's prevalence in the tool's interface. For example, with regards to Gender, one participant explained that they chose to distribute female students across teams in a course with a significant female population, whereas when there were three or fewer women in the course, they were assigned to the same team. Somewhat contradictorily, despite this interest in students' characteristics, student input was the least frequently considered factor. This tendency to not consider student input also seems to contrast with how important instructors rated student preferences to be (median=5 on a scale from 1 (Strongly disagree) to 7 (Strongly agree) for survey item "I would find it useful when selecting the configuration for this course to be able to see what criteria students in my

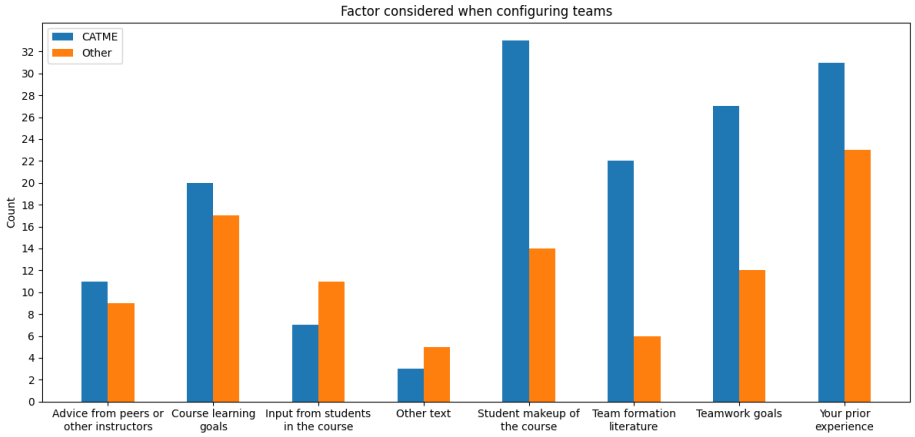

Fig. 4. Factors surveyed instructors reported taking into consideration when selecting criteria.

course prefer to be used."). For instructors not using the tool, their own prior experience was the most frequently selected factor that contributed to determining the criteria they selected. Student makeup of the course, which was the most popular factor in the CATME-specific version of the survey, was ranked third by these instructors. However, a chi-square analysis did not reveal a statistically significant difference in the pattern of factors selected across the two survey variants ($\chi^2(7)=13.36$, $p=0.06$). See Figure 4 for the complete distributions of responses.

In the interviews, instructors described considering similar factors when selecting criteria and weights, with student makeup of the course and course goals tying for the most frequently cited factors (N=14 for both). Student input was again the least frequently considered factor, with only one instructor mentioning taking it into account.

5.2.2 Transparency Concerns. Several instructors in the interviews (N=4) mentioned feeling unsure during the configuration process that they had selected the best criteria for their course. One common concern, especially among instructors who had selected criteria such as Gender and Race/Ethnicity, was about how transparent they should be with students during this process, and whether configuration decisions like these could be perceived negatively by students:

> *"I just have an uneasiness about it. That's just related to, 'Wait, are we [being] profiled?'... I don't know how that's going to be. Whereas with women being at least 25-ish percent [of the students in the course], it's like, that's conceivable that I got paired with another woman. That's not unreasonable, versus there's four Black students in... a class of 400. And I got paired with one. I could see that not going as well. I just don't know."* (P8)

Twelve of the interviewed instructors brought up matters of transparency and related ethics when describing their team formation process, suggesting that these are important considerations when forming teams. One instructor explained:

> *"One of the things that I've found over the years with teaching is that if you explain to students why you're doing things, they tend to be more receptive. And I always tell them if you have a question about why I'm including these criteria in the survey, let me know. And most years, I do have a discussion with them about why am I incorporating race and why am I incorporating gender and why am I incorporating course grade or GPA. So I think [including] the multiple characteristics... in addition to helping me, I hope, get good*

teams, I think it also shows the students that you're trying to get good teams and that you're trying to build their teamwork skills." (P10)

5.2.3 Instructors Learned About Forming Teams over Time. Many instructors described a learning process involving trial and error over multiple semesters when developing their criteria configuration or formation process more generally. Some reported experiences that had led them to form teams later in the term when they knew more about the students (N=3), form teams of a different size (N=5), or change the direction or magnitude of the weight assigned to a criterion (N=3), and many reported valuing different criteria altogether than they did originally (N=12). Six instructors mentioned that they better understood the tool or their other process after repeated use, and many now felt that their experiences had confirmed that their team formation process was working as intended (N=12). This sentiment is echoed in the results from the followup survey, which found that instructors felt both their team formation approach (median=6 on a scale from 1 to 7) and their specific criteria configurations (median=6) were accomplishing their goals.

When asked to explain how they gauge the success of their team formation process, most instructors mentioned using feedback from students in the form of surveys or peer evaluations (N=13), while a few also described observing the quality of teams' collaboration processes and outcomes (e.g., projects, reports; N=3). Interestingly, although distinguished in the interview protocol, most instructors did not seem to differentiate between determining the success of their team formation process and the success of the teams the process generates. For this latter question, instructors again mentioned peer evaluations (N=20), team outcomes (N=15), team process (N=13), and occasionally students' ratings of other teams' work (N=4). This apparent conflation of the tool or process' success with that of the teams is in line with comments from several instructors (N=6) about how the team formation process is not always the most important aspect of the course or students' experiences. Instead, other factors such as team-building may be equally important:

"For me personally, I'm not so interested in the team formation, but I am in the improving of teamwork. And so I would definitely be curious to have better tools to, 'Okay, this is the problem we're having in the team, here's some tools that can help you solve that problem.' I would love that, that would be really helpful for me, because the team formation... it is what it is. I like it, it's easy, and I have pretty good success with it. But when it comes to the problems that they're having within the team, I feel like there aren't great resources in tackling individual problems, short of my own life experience." (P9)

5.3 Instructors Develop Mental Models Beyond the Tool

Multiple participants (7) in the CATME-specific survey variant and all instructors interviewed who had used CATME (N=10) mentioned they were unsure about how the tool's algorithm actually worked. As a specific example, when asked about one feature of the tool (numeric scores representing how well the generated teams fulfill the desired configuration), four instructors reported not using the scores due to not understanding what they represented, and a further two had not realized the scores existed. However, many instructors did offer their own interpretations about how the tool functioned (N=8). For instance, when asked to explain their thoughts on including multiple criteria in the team formation algorithm (e.g., do the benefits of individual criteria stack?), many instructors (N=6) agreed that using additional criteria was helpful (e.g., to capture different aspects of diversity), but there was a limit at which there were no further benefits or the constraints became too difficult to satisfy:

"Of course, the more extreme you go, then the more defeating the purpose of what CATME does, because it's just going to end up, more or less, trying to satisfy everything and nothing gets satisfied." (P12)

> *"And so, I do believe if you put too many variables in, you may start to variable yourself right out of teams that function at all."* (P17)

This understanding was shared by instructors not using the tool (N=8).

We also found that instructors tended not to accept the teams generated by the tool without question. Nine of the 10 instructors who used CATME described performing some kind of "sanity check" on the results of the team formation. Most frequently, this took the form of going over the composition of each team to check for potential problems (N=9), although some instructors also reported rerunning the algorithm multiple times and comparing the results (N=6). Instructors frequently described using the tool's feature allowing them to separate or group specific students during this process (N=9), often in cases where they knew relevant information about students the tool couldn't capture (e.g., personalities, friend groups, students who were dating or had previously dated, etc.). Again, similar behaviors were described by the instructors using other criteria-based formation methods, nine of whom described performing some kind of sanity check after their initial pass at forming teams. Interestingly, of the nine instructors who initially reported not performing a sanity check and simply accepting the results generated by the tool or their other process, six went on to describe performing one of the checks mentioned above. Only three instructors who reported not performing a sanity check actually seemed not to perform one.

Instructors' intervention in the team formation process through sanity checks and the manual manipulation of team assignments seems to indicate their understanding that algorithmic tools may not necessarily produce "perfect teams", even though survey results showed that creating teams with an increased chance of success was a major reason for using the tool at all. One instructor explained:

> *"Overall... this does do a nice job of putting together teams in a methodical way that I would not have thought of otherwise. So, I call it a partnership. CATME does take a cut at it. And some teams, they're all CATME put together teams. Other times, I will move people around based on the things we talked about earlier."* (P20)

This interpretation is consistent with comments by instructors describing aspects of the students or the projects they will be completing which are distinctly human or messy, and therefore cannot be easily captured by the tool or team formation process (N=9). For example, one instructor described an experience from her own time as a student when her teammates stepped up to adjust work distribution on a project long after the team had been formed:

> *"Now, the backstory was I found out I was pregnant the day I started grad school. My kid was supposed to be born after the end of spring quarter, but he was early. I was literally trying to get my operating systems project done while nursing a newborn... I will still to this day talk in glowing terms about my partners who... did pick up the slack for me. Because when you have these situations, one of the beauties of group work is that our lives are messy, and not everything works all the time. And occasionally, you'll have a student who falls down for whatever reason, and it is nice when you have groups that are willing to say, 'Okay, I know this isn't like this person. I'm going to help this person out,' and sees it as an altruistic thing."* (P3)

Controlling during team formation for this kind of unexpected situation or how team members would respond would likely be impossible. However, instructors did describe situations where they provided additional scaffolding (N=10) in assignments or lectures to offset some of these aspects that are difficult to account for during team formation (e.g., asking students in a discipline-homogeneous class to approach a problem as if they were from a different academic area) or reached out to intervene in struggling teams (N=18), again indicating their understanding of the importance of their involvement, even after teams have been formed.

Finally, eight of the instructors using CATME mentioned being constrained by the tool in some way (i.e., some element of their desired team formation process was not supported by the tool):

> *"The first thing I always check is I look at the teams where students have said they don't want to do weekend meetings. So they either say no weekend meetings or avoid weekend meetings. And I make sure that they have schedule blocks that align with their teammates that are not on the weekend. Because that's a big thing, CATME doesn't do that. It doesn't link the questions as far as I can tell."* (P10)

This phenomenon was not unique to CATME users (three instructors using other methods also mentioned being constrained somehow by their process). However, it is clear that instructors using algorithmic tools do face situations where they must perform extra work outside of the tool in order to carry out their team formation process as desired, or else abandon aspects of their ideal process to adhere to the affordances of the tool.

5.4 Tool Users Displayed Unique Behaviors

The interviews revealed several interesting behaviors of instructors using the studied tool that were not exhibited by any of the instructors using other methods of team formation. First, several participants (N=4) described including more criteria in the "About you" survey the tool distributes to students than they ultimately intended to use in their configuration, either because they were not sure initially whether they would use this extra information, or because they wanted to intentionally obfuscate which criteria would be used:

> *"I figured if I just ask about Race and Gender and Grade... I think they would guess why they're being grouped the way they do. That's why I put like Year, and Age and Major and Big Picture/Detail Oriented... I ignored the rest, but I asked about them not just to broadcast, 'Hey, here's how we're going to make groups.' Because again, I'm fearful of that piece of students feeling like, he's putting us together because we're minorities."* (P8)

As P8 mentions, this latter behavior is linked to the issues of transparency that many instructors raised, and may also be related to concerns about students attempting to game or manipulate the team formation process if they know too much about how it works (N=2). None of the instructors using other methods exhibited this behavior.

Additionally, seven instructors mentioned the default criteria and weights set by the tool, and how these impacted their own selections:

> *"The first thing I noticed is that the default CATME [criteria]... they have default weights. And so I thought they know something that caused them to choose these weights. So I was hesitant to change those. And then I noticed that the ones that other people create typically have no weights. And I was like, maybe they know something I don't know. You know what I mean? I wasn't sure... I really hesitated to change those weights for quite a while. But then eventually I started being willing to experiment."* (P14)

> *"If I remember correctly... Gender, Race, and GPA are at the top of my list of parameters for these teams. That would suggest to me that those are important, more important than some of those other features... It struck me as a signal. The fact that I have to check these check boxes at the top does signal to me that those are the higher priority items." (P8)*

6 DISCUSSION

In this section, we discuss the answers to the research questions of this study, relate our findings to the existing literature, and provide implications for how tool designers might better address the needs of instructors using algorithmic team formation tools. While the instructors studied

here have developed strategies to leverage the tool or approach they currently use based on their own mental models and the literature, clarifying certain aspects of tools or including additional guidance could improve user experiences and facilitate adoption by others. Note also that while some of the findings described here are specific to CATME Team-Maker, we believe that many may generalize to other tools and team formation approaches. Future research can continue to explore these opportunities.

6.1 RQ1: Instructors' Goals for Teamwork

Instructors in this study most frequently described using teamwork in their courses to promote the learning of teamwork skills, support peer-based learning, and prepare students for future careers in industry. These goals align with existing literature, which has shown both benefits in using teamwork to achieve them and areas of opportunity where they are not currently being met [14, 26, 39, 46, 53, 56, 61, 68]. Similar prior work also exists for other goals mentioned by instructors such as promoting social bonds among students [25, 40, 43], reducing grading [20], and supporting larger or better quality work outcomes [7]. Our study provides insight into the relative importance instructors assign to these goals in the context of criteria-based team formation. We also found that while many instructors did consider these goals when choosing which criteria to use in their configurations, most often by prioritizing various aspects of diversity on teams (e.g., students of different genders, races, or academic areas being placed together), the goals were not necessarily reflected clearly in instructors' processes and some struggled to relate goals to specific settings in the tool. Several instructors mentioned not considering their goals at all when deciding their team formation process, and many cited convenience or efficiency as the primary reason they had chosen to use the studied algorithmic tool rather than these other aims. Current team formation interfaces tend not to focus on instructors' broader goals for the teamwork in their courses, but rather the low-level details of how student characteristics should be distributed across teams, often on a criterion-by-criterion basis. Tool designers could therefore elicit information from instructors about their goals at the beginning of the configuration process, not only to help them solidify their goals and think more explicitly about how they relate to their process, but also to suggest an appropriate default set of criteria and weights for them to refer to and adapt based on the relevant literature for each goal. Tools might even suggest theory-grounded goals themselves for instructors who are unsure of their priorities, for example encouraging the formation of diverse teams to enhance learning outcomes. Machine learning (ML) techniques and aggregated user data could also be used to tailor these suggestions for individual instructors based on their own prior configurations or characteristics of their course. However, in any case, it is important to choose defaults carefully because, as we observed and as prior work shows, users tend to rely heavily on defaults and are reluctant to change them [59].

6.2 RQ2: Instructors' Criteria Configurations

Instructors in this study used a variety of team formation criteria aligning with prior literature on team composition, such as Gender [69], Schedule [55], Race/Ethnicity [60], and curricular or project interests [15]. The weights or relative priorities they assign to these criteria vary widely, and instructors who used the studied algorithmic tool typically selected a large number of criteria (9-10) in one configuration, even though most acknowledged that there is likely a limit after which adding additional criteria is not useful or may actively harm teams. These configurations are consistent with those reported in prior work by Jahanbakhsh, Hastings, et al. examining the same tool [29, 30, 37], suggesting that these choices generalize beyond the context of the single university studied in this existing work. On the other hand, instructors in our study who did not use the tool selected significantly fewer criteria, on average 2-3 in one configuration. One possible explanation for the

difference in configuration sizes is the flexibility and lack of feedback offered by team formation tools, which allow a large number of criteria to be selected at once without understanding their potential impacts on team outcomes. There may also be a nudging effect [36, 62] present, where the inclusion of certain criteria in a tool's list of defaults prompts instructors to select them, although they may not have thought of using them before. Tool designers may wish to notify instructors somewhere in the interfaces they create that such complex configurations could be difficult to satisfy and may not have the same benefits as more focused selections of criteria [30], or provide literature-based templates of appropriate sizes on which instructors could base their own choices.

6.2.1 Factors Considered When Selecting Criteria. Instructors did not often explicitly consider the literature when selecting criteria or weights (the factor appearing roughly at the midpoint of the ranking of possible factors considered for both the survey and the interviews). Instead, they favored their own prior experience and goals for the teamwork, as well as the student makeup of the course in order to select the configurations most meaningful to their own contexts and facilitate the completion of the teamwork for their specific students. It is possible that algorithmic tools– and the associated trial and error of actually forming teams– offer a more engaging opportunity for instructors to learn about team formation than studying the literature, a form of learning by doing rather than by reading. It may also be that tools grounded in the research on team composition (e.g., [44]) lend a certain degree of credibility to instructors' configurations, regardless of their own actual familiarity with this literature. As described previously and in work by Jahanbakhsh et al. [37], instructors mentioned that one desirable feature of using a tool was that it allowed them to be perceived as somewhat removed from the team formation process: although they may have selected the configuration, it was ultimately the tool that formed the teams, and the tool could be blamed if there were any problems. This removal also offered a sense of credibility:

> *"Even though I'm behind the scenes like in the Wizard of Oz pulling the strings, something about it being CATME helps them accept the team when it's put forward."* (P14)

It is possible that this perceived credibility applies not just to students in the course, but also extends to instructors' own beliefs about tools.

Designers may wish to more explicitly incorporate knowledge or advice from the team composition literature into tools' interfaces, since instructors do not always consult this body of work when developing their configurations. Relying more on the literature could make instructors more confident in their choices and the credibility of the tool, and ease the burden of selecting a configuration from an expansive variety of potential options. This guidance could come in the form of suggesting specific criteria or complete default configurations for courses in different disciplines, with references to the related existing research.

6.2.2 Transparency and Student Input. Instructors described taking matters of transparency into account during their team formation process. On one hand, many instructors mentioned the value of involving students more closely in the process and making sure they were aware of the decisions that were made in the tool. This stance aligns with the literature on algorithmic transparency (e.g., [19, 41, 45, 63]), as well as prior research showing that increased agency and ownership can improve student learning outcomes (e.g., [16, 18, 21, 33, 51]). Although the instructors in our study did not often consider direct input from students on the criteria and weights they used, they very frequently considered the student makeup of the course while making their selections, and often included such factors as student project preferences or preferences for specific classmates students did (not) want to work with. As described, they also often took into account their own knowledge about the skills, backgrounds, and personalities of individual students, usually accommodating these adjustments by hand. Although it would be difficult to fully automate all of these aspects in the tool, designers

could begin by introducing collective student opinions into the algorithm (e.g., [29, 31]), including a channel through which students could share opinions or concerns about their prospective teams before they are finalized, or including a dedicated project preference criterion. Using an approach like the LIFT workflow proposed in [29] also aligns with instructors' stated goals of supporting student learning about teamwork and peer-based learning, as it provides opportunities for students to engage with the question of what makes a good team and learn from their peer's viewpoints and experiences.

However, some instructors also expressed concerns about students knowing too much about the criteria configurations or team formation process used, and in some cases intentionally obfuscated which inputs would ultimately be used by the algorithm by including more criteria in the tool's student-facing "About you" survey than they intended to use. This approach is more in line with prior work showing potential drawbacks or complexities to algorithmic transparency (e.g., [8, 41]), as well as concerns by instructors about students potentially manipulating the team formation algorithm, as described previously and in prior work [6, 29]. Instructors concerned with this issue may not think to practice this obfuscating behavior without being prompted by a tool; therefore, designers might also consider building in a suggestion to do so or otherwise consider issues of transparency more explicitly. For example, as suggested in [29], reducing tools' reliance on data that is self-reported by students with each team formation is a promising approach. We suggest that tools might rather retain students' responses from previous courses or help them build up "teamwork profiles" consisting of prior team outcomes or peer evaluations that might be incorporated into the team formation algorithm, shared with future employers, or used to help establish their role on a current team, perhaps during a team-building activity like those described in [30]. Since these profiles would persist and have implications beyond the context of a single team project and potentially be seen by others, students might therefore have more incentive to maintain their accuracy.

6.3 RQ3: Instructors' Understanding of the Team Formation Tool

In this study, we have used CATME Team-Maker [44] as a representative algorithmic team formation tool. While it is not the only such tool, identifying areas where instructors experienced difficulties can help to shed light on how the design of similar tools might be improved. In particular, one aspect of the studied tool that our results suggest could be problematic is that its configuration interface is *algorithm-centered*. As previously described, many instructors found the numeric scores representing how well each generated team fulfills the desired configuration of criteria to be confusing or not intuitive, or did not realize they existed at all. Instructors also expressed doubt about how other aspects of the tool worked, including the numeric weights used to set the relative importance of the criteria. It is possible then that while numeric representations of the tool's inputs and outputs may map easily to the underlying algorithm, they do not match adequately with users' understandings of how forming teams works. In addition, there is no clear relationship between these inputs and outputs visible in the tool's interface, making it more difficult to relate instructors' educational goals with settings in the tool. This problem is also present in other non-tool-based methods like spreadsheets and card-sorting, where there is no obvious way for instructors to identify how well their current set of teams is fulfilling their desired configurations without frequently pausing to manually check or maintaining an ongoing sense of the teams' compositions in their mind. Metrics proposed in prior work on evaluating team formation (e.g., [54]) are also largely numeric and may be difficult to calculate for instructors who are less technologically savvy or are not using a tool. Performing the kind of "sanity checks" on teams that most instructors in the study mentioned therefore requires substantial effort, whether using a tool or another criteria-based process.

Designers might then consider alternative ways of representing the inputs and outputs of team formation algorithms in order to better facilitate sanity checking and align more with instructors' mental models and goals. For example, during the configuration of a tool, instructors might specify what an "ideal" or "gold standard" team looks like for their context (e.g., a team where women are not outnumbered and where there are at least two programmers of medium skill or higher). The interface could then visualize the distance between each generated team and this ideal, either holistically or criterion-by-criterion. Text explanations could also be generated by tools, either for this purpose or to check that the specified configuration is correct (e.g., "You have described a team where..."). These features may be especially useful to new faculty, those using the tool for the first time, or those who have had unsuccessful experiences in prior terms and therefore cannot rely on their prior experience when configuring the tool. Future research could compare the impacts of such alternative representations on instructors' perceptions of the configuration process and the ultimate success of the team formation.

7 LIMITATIONS AND FUTURE WORK

One limitation of this work is that many of the participants in the study (31%) were Computer Science instructors, probably due to the mailing lists used during recruitment (i.e., SIGCHI, SIGCSE, and the ACM overall are computing associations). The configurations and responses reported by these instructors may be discipline-specific and not generalize to other areas. The sample for the tool-specific version of our survey is also limited to CATME Team-Maker users, and may not be representative of the choices made by users of other tools. Future work could investigate whether the trends observed here generalize to a broader sample of instructors in different areas and using different methods of algorithmic team formation, or to different contexts such as organizational or professional teams.

In addition, it is not clear how the increasing amounts of remote instruction brought on by the COVID pandemic may impact team formation behavior. For example, many instructors in our sample mentioned incorporating more elements of self-selection into their process since early 2020, to allow students to interact more with their friends and otherwise reduce burden. Future work could investigate whether such behaviors continue in the long term, or what other changes instructors have made to their processes.

Additional research is also needed to determine the effectiveness of the criteria configurations instructors select in practice. Although the instructors in our study reported being satisfied with their processes and believed they were accomplishing their goals, metrics such as student team grades or other outcomes could be examined to give a fuller picture of the configurations' quality, as well as the effects of other tool settings on student learning processes.

The results of this study suggest possibilities for future research directions. One possible extension of this work would be to gather additional forms of data to complement the survey and interviews described here. For example, future work could examine log data from instructors' use of a team formation tool, and apply ML techniques to extract patterns of interest. Future work could also compare instructors' experiences with algorithmic interfaces featuring differing levels or sources of support (e.g., suggestions for criteria or weights based on the literature, student preferences, other instructors' choices, etc.) during the configuration process. In both cases, these studies could provide more insight into whether and how instructors' choices in the tool translate into more effective student teams. Finally, future research might also further explore how algorithmic and manual team formation approaches might be combined. While we here discuss a process where users manually review and potentially modify teams after they are generated by the tool, this interaction might also occur earlier in the process, such as tool-assisted self-selection of teams (e.g., [23]). Future work can continue to investigate these possibilities.

8 CONCLUSION

Current algorithmic team formation tools provide little in the way of support for instructors deciding how to configure the tool in order to select the most appropriate criteria and weights given their goals for the teamwork in the course, and many options exist for sourcing potential configurations (e.g., the literature, other instructors, students, etc.). To determine how instructors are making these decisions and identify opportunities where tools might provide more guidance, we conducted a survey and interviews examining instructors' current practices when forming student teams in their courses with a criteria-based approach. We found that instructors most frequently prioritized the development of teamwork skills and the support of peer-based learning, and selected configurations in support of these goals. However, instructors also expressed doubts about how the algorithm actually worked and some questioned whether the criteria and weights they had selected would accomplish their desired outcomes and educational goals. Most also conducted "sanity checks" on the output of the tool or their other process to determine whether the generated teams matched their desired specifications or required further intervention. In general, instructors also did not solicit any input from students when selecting criteria, despite acknowledging that this information might be useful and considering issues of transparency with students. The results of this study further highlight the need for learner-centered workflows and suggest that opening the "black box" of the team formation algorithm more to students could be a promising way to aid instructors in configuring algorithmic tools while at the same time giving their students increased agency and an opportunity to learn more about teamwork. Through this work, we hope to give tool designers and instructors the knowledge they need to create and use algorithmic team formation tools effectively. These improved decisions, in turn, may help students learn more from each other, produce more effective project solutions, and have more favorable attitudes toward the course topic.

ACKNOWLEDGMENTS

This work was partially funded by the Strategic Instructional Innovations Program (http://ae3.engineering.illinois.edu/siip-grants/) at the University of Illinois and by NSF awards CCF-1439957 and IIS-2016908. We thank Prof. Emma Mercier, Prof. Darko Marinov, Wendy Shi, Tiffany Li, and Sofia Meyers for their invaluable feedback.

REFERENCES

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

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

[3] 2020. *CATME User Institutions (Alphabetical by Country).* Retrieved 19 August 2020 from https://info.catme.org/instructor/history-research/our-user-base/catme-user-institutions-alphabetical-by-country/

[4] 2020. *Usage of CATME System.* Retrieved 14 September 2020 from https://info.catme.org/instructor/history-research/

[5] ABET. 2019. 2020-2021 Criteria for Accrediting Computing Programs. https://www.abet.org/wp-content/uploads/2019/12/C001-20-21-CAC-Criteria-MARK-UP-11-30-19-Updated-2.pdf.

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

[7] Nancy Allen, Dianne Atkinson, Meg Morgan, Teresa Moore, and Craig Snow. 1987. What Experienced Collaborators Say About Collaborative Writing. *Iowa State Journal of Business and Technical Communication* 1, 2 (1987), 70–90. https://doi.org/10.1177/105065198700100206 arXiv:https://doi.org/10.1177/105065198700100206

[8] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645 arXiv:https://doi.org/10.1177/1461444816676645

[9] Donald R Bacon, Kim A Stewart, and William S Silver. 1999. Lessons from the best and worst student team experiences: How a teacher can make the difference. *Journal of Management Education* 23, 5 (1999), 467–488.

[10] 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.
[11] Mehdi Beheshtian-Ardekani and Mo A Mahmood. 1986. Education development and validation of a tool for assigning students to groups for class projects. *Decision Sciences* 17, 1 (1986), 92–113.
[12] Suzanne T Bell. 2007. Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of Applied Psychology* 92, 3 (2007), 595–615.
[13] Suzanne T Bell, Shanique G Brown, Anthony Colaneri, and Neal Outland. 2018. Team composition and the ABCs of teamwork. *American Psychologist* 73, 4 (2018), 349.
[14] David Boud, Ruth Cohen, et al. 2014. *Peer learning in higher education: Learning from and with each other*. Routledge.
[15] 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.
[16] 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.
[17] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 873–888. https://doi.org/10.1145/2998181.2998250
[18] 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.
[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] Curt J. Dommeyer. 1986. A Comparison of the Individual Proposal and the Team Project in the Marketing Research Course. *Journal of Marketing Education* 8, 1 (1986), 30–38. https://doi.org/10.1177/027347538600800104 arXiv:https://doi.org/10.1177/027347538600800104
[21] Yohanan Eshel and Revital Kohavi. 2003. Perceived classroom control, self-regulated learning strategies, and academic achievement. *Educational psychology* 23, 3 (2003), 249–260.
[22] Diego Gómez-Zará, Leslie A DeChurch, and Noshir S Contractor. 2020. A taxonomy of team-assembly systems: Understanding how people use technologies to form teams. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36.
[23] Diego Gómez-Zará, Mengzi Guo, Leslie A DeChurch, and Noshir Contractor. 2020. The impact of displaying diversity information on the formation of self-assembling teams. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[24] Diego Gómez-Zará, Matthew Paras, Marlon Twyman, Jacqueline N Lane, Leslie A DeChurch, and Noshir S Contractor. 2019. Who would you like to work with?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
[25] Hee Young Han. 2009. *Relationship between students' emotional intelligence, social bond, and interactions in online learning*. University of Illinois at Urbana-Champaign.
[26] Randall S. Hansen. 2006. Benefits and Problems With Student Teams: Suggestions for Improving Team Projects. *Journal of Education for Business* 82, 1 (2006), 11–19. https://doi.org/10.3200/JOEB.82.1.11-19 arXiv:https://doi.org/10.3200/JOEB.82.1.11-19
[27] Alexa M Harris, Diego Gómez-Zará, Leslie A DeChurch, and Noshir S Contractor. 2019. Joining together online: the trajectory of CSCW scholarship on group formation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
[28] 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.
[29] Emily M. Hastings, Albatool Alamri, Andrew Kuznetsov, Christine Pisarczyk, Karrie Karahalios, Darko Marinov, and Brian P. Bailey. 2020. LIFT: Integrating Stakeholder Voices into Algorithmic Team Formation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(Chi '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376797
[30] 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.
[31] Emily M. Hastings, Sneha R. Krishna Kumaran, Karrie Karahalios, and Brian P. Bailey. 2022. A Learner-Centered Technique for Collectively Configuring Inputs for an Algorithmic Team Formation Tool. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for

Computing Machinery, New York, NY, USA, 969–975. https://doi.org/10.1145/3478431.3499331
[32] 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.
[33] Susan Horwitz, Susan H Rodger, Maureen Biggers, David Binkley, C Kolin Frantz, Dawn Gundermann, Susanne Hambrusch, Steven Huss-Lederman, Ethan Munson, Barbara Ryder, et al. 2009. Using peer-led team learning to increase participation and success of under-represented groups in introductory computer science. *ACM SIGCSE Bulletin* 41, 1 (2009), 163–167.
[34] 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.
[35] Roland Hubscher. 2010. Assigning students to groups using general and context-specific criteria. *IEEE transactions on learning technologies* 3, 3 (2010), 178–189.
[36] Dennis Hummel and Alexander Maedche. 2019. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics* 80 (2019), 47–58. https://doi.org/10.1016/j.socec.2019.03.005
[37] 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.
[38] David S Jalajas and Robert I Sutton. 1984. Feuds in student groups: Coping with whiners, martyrs, saboteurs, bullies, and deadbeats. *Organizational Behavior Teaching Review* 9, 4 (1984), 94–102.
[39] Jéssica Mendes JORGE[a], Alexandre Crepory Abbott de OLIVEIRA, and Andrea Cristina dos SANTOS. 2020. Analyzing how university is preparing engineering students for Industry 4.0. In *Transdisciplinary Engineering for Complex Socio-technical Systems–Real-life Applications: Proceedings of the 27th ISTE International Conference on Transdisciplinary Engineering, July 1–July 10, 2020*, Vol. 12. IOS Press, 82.
[40] Janna Juvonen. 2006. Sense of Belonging, Social Bonds, and School Functioning. (2006).
[41] 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.
[42] Cameron Klein, Deborah DiazGranados, Eduardo Salas, Huy Le, C Shawn Burke, Rebecca Lyons, and Gerald F Goodwin. 2009. Does team building work? *Small Group Research* 40, 2 (2009), 181–222.
[43] Sneha R. Krishna Kumaran, Deana C. McDonagh, and Brian P. Bailey. 2017. Increasing Quality and Involvement in Online Peer Feedback Exchange. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 63 (dec 2017), 18 pages. https://doi.org/10.1145/3134698
[44] 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.
[45] 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.
[46] Susan Lerner, Diane Magrane, and Erica Friedman. 2009. Teaching teamwork in medical education. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine* 76, 4 (2009), 318–329.
[47] 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.
[48] Ioanna Lykourentzou, Robert E Kraut, and Steven P Dow. 2017. Team Dating Leads to Better Online Ad Hoc Collaborations. In *CSCW*. 2330–2343.
[49] 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.
[50] Joseph E McGrath, Holly Arrow, and Jennifer L Berdahl. 2000. The study of groups: Past, present, and future. *Personality and social psychology review* 4, 1 (2000), 95–105.
[51] Jeffrey A Mello. 1993. Improving individual member accountability in small work group settings. *Journal of Management Education* 17, 2 (1993), 253–259.
[52] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. *Survey Research in HCI*. Springer New York, New York, NY, 229–266. https://doi.org/10.1007/978-1-4939-0378-8_10
[53] Humberto Oraison, Loretta Konjarski, and Samuel Howe. 2019. *Journal of Teaching and Learning for Graduate Employability* 10, 1 (2019), 173–194. https://search.informit.org/doi/10.3316/informit.580981748647262

[54] Asma Ounnas, David E Millard, and Hugh C Davis. 2007. A metrics framework for evaluating group formation. In *Proceedings of the 2007 international ACM conference on Supporting group work*. 221–224.
[55] Michael A Redmond. 2001. A computer program to aid assignment of student project groups. *ACM SIGCSE Bulletin* 33, 1 (2001), 134–138.
[56] Marcel M Robles. 2012. Executive perceptions of the top 10 soft skills needed in today's workplace. *Business Communication Quarterly* 75, 4 (2012), 453–465.
[57] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.
[58] 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.
[59] Rajiv C. Shah and Christian Sandvig. 2008. SOFTWARE DEFAULTS AS DE FACTO REGULATION The case of the wireless internet. *Information, Communication & Society* 11, 1 (2008), 25–46. https://doi.org/10.1080/13691180701858836 arXiv:https://doi.org/10.1080/13691180701858836
[60] 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.
[61] Debra Smarkusky, Richard Dempsey, J Ludka, and Frouke de Quillettes. 2005. Enhancing team knowledge: instruction vs. experience. In *ACM SIGCSE Bulletin*, Vol. 37. ACM, 460–464.
[62] Cass R Sunstein. 2014. Nudging: a very short guide. *Journal of Consumer Policy* 37, 4 (2014), 583–588.
[63] 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.
[64] Sander Valstar, Caroline Sih, Sophia Krause-Levy, Leo Porter, and William G. Griswold. 2020. A Quantitative Study of Faculty Views on the Goals of an Undergraduate CS Program and Preparing Students for Industry. In *Proceedings of the 2020 ACM Conference on International Computing Education Research* (Virtual Event, New Zealand) *(ICER '20)*. Association for Computing Machinery, New York, NY, USA, 113–123. https://doi.org/10.1145/3372782.3406277
[65] 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.
[66] Miaomiao Wen, Keith Maki, Steven Dow, James D. Herbsleb, and Carolyn Rose. 2017. Supporting Virtual Team Formation through Community-Wide Deliberation. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 109 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134744
[67] 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.
[68] David L. Williams, John D. Beard, and Jone Rymer. 1991. Team Projects: Achieving their Full Potential. *Journal of Marketing Education* 13, 2 (1991), 45–53. https://doi.org/10.1177/027347539101300208 arXiv:https://doi.org/10.1177/027347539101300208
[69] 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.

Received January 2023; revised April 2023; accepted July 2023