Examination of the Effectiveness of a Criteria-based Team Formation Tool

Albatool A. Alamri and Brian P. Bailey
Department of Computer Science
University of Illinois
Urbana, IL
{aaalamr2, bpbailey} @ illinois.edu

***Abstract*— In this Research Work in Progress Paper, we examine the effectiveness of CATME, a tool that implements a criteria-based team formation approach. The tool facilitates forming teams based on criteria like demographics, skills, and work styles. This information is collected from the students via an online survey. The effectiveness of this genre of tool depends on the practicality of the instructor's configuration of the criteria, the veracity of students' responses to the survey, and the soundness of the algorithm. In this work-in-progress paper, we investigate potential issues affecting these factors. Our study was conducted by performing new analysis of data collected from a prior study comparing the performance of teams formed using CATME or randomly in a user interface design course. The performance of teams was not statistically different between the two conditions. In examining the students' responses to the team formation survey, we found issues related to Self-Assessment such as inconsistencies between students' ratings of their skills and reporting of their strongest skills. Likewise, we found some cases where the tool produced unexpected results when calculating the homogeneity of the skills of a team. Implications for instructors to mitigate these problems are discussed.**

Keywords—team formation, criteria-based approach, student's self-assessment

I. INTRODUCTION

Team-based projects represent an important method of learning for students where in addition to applying learned material and developing required skills, they get to develop essential teamwork skills such as leadership, communication, and handling conflicts. For such a method of learning to be effective, teams need to be formed in a way that minimizes conflicts and maximizes learning. The literature on team composition shows that teams with balanced gender [1], balanced personality types [2], diverse academic abilities and skills [3][4] demonstrate improved performance. However, manually forming teams to satisfy multiple criteria, especially in large classes, can be extremely difficult. Consequently, instructors are increasingly leveraging tools that implement the criteria-based approach. One example of such tools is CATME [5], a representative of the criteria-based formation approach that is increasingly gaining support [6].

To form teams using CATME, the instructor can select a set of criteria such as schedule, GPA, or skills, and configure the significance of each criterion and the degree of similarity between the team members according to that criterion. A survey would then be sent to the students to gather the needed information related to the chosen criteria. Once the survey is completed, the teams can be generated by the tool, which would also provide a composition score for each team indicating how well that team matched the configured criteria.

It is important to realize that the effectiveness of this tool depends on three factors: the practicality of the instructor selection and configuration of the criteria, the accuracy of students' responses to the survey, and the soundness of the algorithm. In this paper, we investigative potential issues affecting these factors, and consequently the validity of the tools' outcomes.

Our investigation was conducted by performing a new analysis on data collected from a previous study comparing performance of teams formed randomly or with the criteria-based tool [7]. The study was conducted in an engineering User Interface Design course with team-based projects (176 students in 37 teams). The team formation approach (criteria-based vs. random) was the factor, and the project grade was the measure of team performance. The selected criteria included Gender, Leadership Preference, and Course skills (teamwork, programming, designing, writing, speaking). In teams of (4-5), students worked on a 9-week design project, delivered in stages, that comprised 40% of the final grade. The results, contrary to expectations, showed no significant difference in performance between the two conditions. These expectations were based on the fact that the composition scores given by the tool for the criteria-based teams, especially in terms of skills (μ=7.88, s=1.590), were significantly higher than those of the random teams ($t(33.18) = 4.96$, $p < .01$).

In the new analysis of the potential issues that may have led to that result, we examined the data of the team formation survey and the composition heuristics used by the tool to form the teams. Three key issues were Identified. First, there were inconsistencies in students' responses between determining the possession of a skill and the level of that skill; Students were asked to rate their writing skill level on a five-point scale (None to Expert), then in a different question to choose their strongest skills where writing was one of the choices. About 70% of them indicated a writing level of good (4) or higher, yet only ~55% of that percentage selected writing as a one of their strongest skills. Second, some of the students' responses were contrary to expectations. Of those who did not report having a programming skill (54 students), 25 of them were Computer Science students in their 3rd or 4th year of undergraduate study

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

or graduate students, mostly of whom had GPAs higher than 3.0. (on a four-point scale). Third, the heuristic function used to score the composition goodness of the skillset question, a type of "Choose Many of", operates on an assumption of commonality; the score of a skill with 1 student response only is equivalent to the score of a skill with no response, i.e. that skill's score is zero. The reason is there is no commonality between members for that skill. In these situations, the heuristic for distributing skill does not produce expected scores.

The contributions of this research include the examination of the effectiveness of the criteria-based team formation approach, the identification of potential problems with that approach, and the specification of suggested improvements.

II. Research Questions

We posed the following questions to guide our examination of the data:

- **RQ1:** Given that criteria questions could be evaluative, were students able to adequately evaluate their skills?
- **RQ2:** Which is more effective, asking about the possession of a skill or the mastery of it?
- **RQ3:** Were there instances where students intentionally provided inaccurate responses?
- **RQ4:** How is the composition score of a team is computed? And is it accurate?

III. Method

To answer the research questions, and identify potential factors affecting the accuracy of the formation process, we examined two components of the process: the formation survey and the composition scores.

A. Team Formation Survey

The criteria used to form the teams included: Gender, Leadership Preference, Writing skill, and Course skills (teamwork, programming, writing, speaking, design). In contrast to gender and leadership preference, factual and preferences questions, we focused on examining the responses to the evaluative questions (Course Skills and Writing Skill) to see if students' self-assessments were accurate. The Writing Skill question was phrased as "Rate your writing skill", while the Course Skills question was phrased as: "What is your strongest skill(s) as it relates to a design project in the course?"

First, as Writing was in both questions, we compared students answers for the skill level question with the skill possession question. Furthermore, to see which question correlated more with the grade of the writing components of the project, we conducted a single-linear regression analysis for each. Second, since the course is a Computer Science course, we examined the demographics of students who did not report having a programing skill.

These two technical skills (programming and writing) surpass in necessity the rest of the skills. They are foundational to the course and are essential to complete the project which requires the development of a functional user interface in addition to report writing on a weekly basis. Therefore, if any potential inaccuracies affecting the compositions between the criteria-based teams and the randomly formed ones existed, it would be most likely related to those two skills.

B. The Composition Scores

Each formed team is assigned a score indicating the goodness of the composition based on the configured criteria. This score is a combination of the heuristic scores for the responses to each question in the survey. The tool uses a number of specialized heuristic functions for questions such as gender and race -to insure the minorities are not isolated in teams- as well as more general heuristics for questions like "Choose One of" or "Choose Many of".

The skills questions in the team formation survey are of the later types (general heuristics). Therefore, we examined the composition scores of these two questions.

IV. Results

A. Students' Evaluation of their Skills (RQ1)

The results of comparing the students' responses between the Writing Level question and the Skillset question is shown in Fig.1. We see that students report having the skill if they assess their level to be at least Average. Interestingly, of those with a writing level of Good or Expert, 45.08% of them did not report writing to be one of their strongest skills.

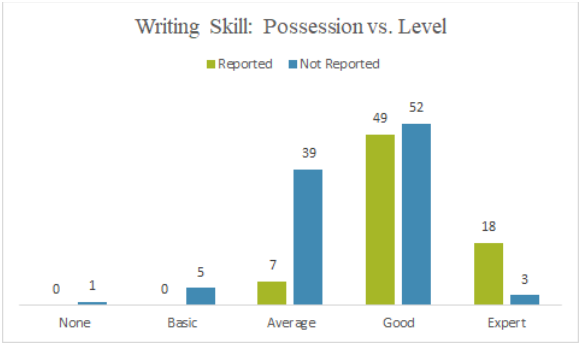

Fig. 1. Students responses about their writing skill to the Writing Level and Skillset questions

B. Skill Possession vs Skill Level (RQ2)

To know which is more effective to ask about, skill possession or skill level, two single-linear regression analyses were calculated to predict the grade in the written components of the project based on the number of writers in a team. In the first analysis, the number of writers was determined by those reporting writing as one of their strongest skills. In the second analysis, the number of students was determined by those indicating a writing level of Good (4) or Expert (5) on a scale of 5 in the writing level question. The reason for counting only the good and expert writers is to see the actual effect of their distribution in teams since only about half of them considered writing as one of their strongest skill. The first analysis had a

small correlation (-0.23), that was not significant ((F (1,35) =1.98, p=0.168). On the other hand, the second analysis showed a stronger correlation, with a coefficient of (-0.34) and a significant regression equation (F (1, 35) =4.56, p=0.04), R^2 of 0.115. See Fig.2 and Fig.3.

Noteworthy that both correlations were negative as the Grade deceased -1.13, for each additional writer in the team in the first analysis, and -1.62 in the second analysis.

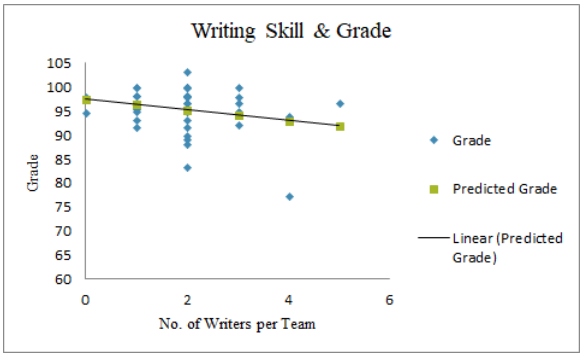

Fig. 2. Regression Analysis between grade and the number of writers per team as determined from the skillset question

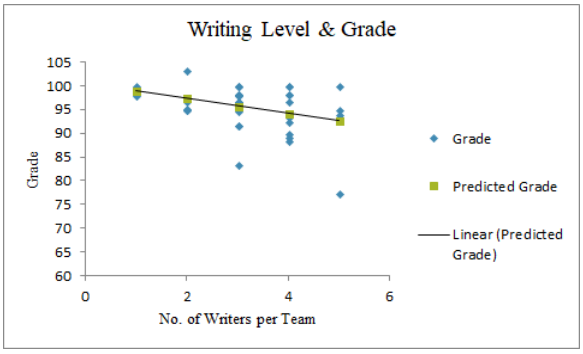

Fig. 3. Regression Analysis between grade and the number of writers per team as determined from the writing skill level question

C. Computer Science Students & Programming Skill (RQ3)

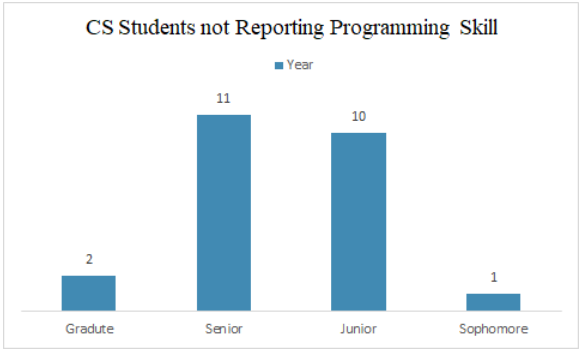

Fig. 4. Computer Science students who did not report programming skill

About 70% of students reported Programming as one of their strongest skills. In examining the demographics of those who did not choose that skill (54 students), we found that 25 of them were Computer Science students in their 3rd or 4th year of undergraduate study or graduate students, whom are expected to master that skill, see Fig. *4*. Moreover, 10 of them had GPAs above 3.5 (on a four-point scale) indicating high academic performance, Table *1*.

TABLE I. CS STUDENTS NOT REPORTING PROGRAMMING SKILL

	Graduate	Senior	Junior	Sophomore
GPA ≥ 3.5	2	5	3	-
3.5> GPA ≥ 3.0	-	4	6	-
GPA < 3	-	2	1	1

D. Composition scores (RQ4)

While analyzing the teams' compositions, we noticed cases where teams having all skills reported were scored lower than teams with similar compositions yet lacking a skill or two. Fig.5 shows the compositions of two teams in the course. Despite that Team-22 had students with all the required skills, its composition was scored (-1.88), which is lower than Team-34 that lacked two skills (1.25).

Team22	Student	Team	Prog.	Writing	Speaking	Design	Score	
	T22-S1	x		x			-1.88	
	T22-S2		x		x			
	T22-S3	x	x		x			
	T22-S4	x	x		x	x		
	T22-S5							
	Value	0.563	0.5625	0	0.5625	0		
	Norm	0.126	0.125	0	0.125	0	0.376	Sum
							-1.88	Weighted

Team34	Student	Team	Prog.	Writing	Speaking	Design	Score	
	T34-S1		x				1.25	
	T34-S2	x	x					
	T34-S3	x	x		x			
	T34-S4							
	T34-S5	x			x			
	Value	0.5625	0.5625	0	0.25	0		
	Norm	0.125	0.125	0	-0.5	0	-0.25	Sum
							1.25	Weighted

Fig. 5. Composition scores for two teams in the course. Team-34, lacking two skills is scored higher than Team-22 that had all the skills.

Returning to CATME's Help Page to understand how the compositions were scored, we found that the heuristic function for that question operates on an assumption of "Commonality". To elaborate, the skillset question was of type "Choose Many of". Its heuristic function measures the homogeneity of one option at a time where 1 means all students selected that option and 0 means none selected it. The individual values are normalized, then summed, and finally multiplied by the weight. In in our case, the weight was set to (-5), indicating the heaviest weight in the formation process for dissimilar distribution of skills. The issue here is a key assumption in the function that an option having 1 response only is equivalent to 0 responses as there is no commonality between members for that option. In CATME's help page, this assumption of commonality is justified and explained with an example of choosing mutual sports interests as a formation criterion; cases showing no commonalities, i.e. only one member selecting a certain sport, need not to be considered.

V. Discussion

The factors influencing the effectiveness of the tool are the instructor's specification of criteria, the students' accuracy in self-reports, and the operation of the algorithm. The results of our analysis give insights into potential issues affecting these factors.

The results of RQ1 shows that the phrasing of evaluative questions affect students self-assessments; in response to questions asking about skill level and others asking about selecting their strongest skills, students may report a skill with a level above average, yet they may not necessarily select it as one of their strongest skills.

Furthermore, the investigation of which of these questions correlates more with grades, the results of RQ2 shows that skill level questions show more correlation with performance than skill possession questions. The correlation between the number of writers per team, determined by the skill level question, and the grade of the written components of the project was higher in magnitude and more significant ($r=0.34$, $p<0.05$) compared to the number of writers per team as determined by the skillset question ($r=0.23$, $p>0.05$), which was not significant.

These findings suggest that when the instructors select the formation criteria and design the formation survey, it is better, when asking about skills, to use skill level questions such as "Rate your level of Skill X". It is also recommended to provide a clear description to each level of the skill to reduce the effect of subjective assessment [8].

The findings of RQ4 shows that as the heuristic function of the "Choose Many of" question works with the assumption of commonality between team members, it gives unexpected scores when the goal of the question is to have diverse skills within a team. Fig.5 shows how the heuristic function can favor a team lacking a skill over a team having a similar composition yet have only one student reporting possession of that skill. As a result, the heuristic function of the tool generates teams with less preferred compositions.

This finding strengthens the previous suggestion that instructors should avoid evaluative questions of the type of "Choose Many of" as it is ineffective on both the students' assessment level and the heuristic computation level.

For the outcomes of RQ3, we provide a couple of explanations as to why students did not report having skills they are expected to have. First, the phrasing of the skillset question as choosing their "strongest" skills may have affected their assessment of their skills. Consequently, they may have been discouraged from reporting programming as one of their skills. A second explanation is that some student may have intentionally misreported their expected skills. For example, when students were asked in a final survey about their experiences with CATME and its weaknesses, student 'S7' said:

> "As a strong performer, I think I was much more likely to have a [bad] team in which I had to do a lot of work. I wish I would have undersold my strengths to be matched with qualified partners. When there is a skill mismatch, tasks are less likely to be shared evenly, I'd think."

Another student 'S176' stated that:

> "Obviously, students can enter false or misleading information to try and game the system to end up on a team that might not fit in with the instructor's goals".

These statements suggest that some students may misreport their skills for the fear of handling or performing larger shares of workload in their projects.

Overall, the main concern regarding the effectiveness of this tool is its sole dependency on students' self-reports. The questions in the team formation survey could be factual (gender), preferential (leadership role), predictive (commitment level), or evaluative (skill level). In answering the predictive and evaluative questions, the flaws of self-assessment manifest. Research shows that people's assessments of their knowledge and skills in correlation to objective performance measures tend to be relatively small, moderate at best [9][10]. In addition, people tend to be overconfident in their judgements and predictions of future events or behaviors, which do not always prove to be accurate when the actual situation arrives [8]. Also, there may be cases of inaccurate reports regardless of the reason behind such behaviors.

The insights of this paper motivate finding better means of skill assessment in the tool. For instance, peer-assessments could be used instead of self-assessment. Peers-evaluations generally show more reliability and correlation to instructors' evaluations than self-reports [11] [12]. Furthermore, the tools need to be more resistant to any attempts of gaming the system or manipulating the outcomes of the team formation process.

VI. Future Work

We will examine the data to measure the effect of skill distributions on project outcomes and compare students' assessments with their peers. Also, we are interested in testing the generalizability of our findings by analyzing similar dataset from different courses. Finally, we intend to revise the tool in accordance with the insights form this study and test its effectiveness for team formation.

VII. Conclusion

In this work in progress paper, we report the results of an examination of the effectiveness of CATME, a criteria-based Team formation tool. Looking for potential factors affecting the validity of the tool, we identified several issues. There were inconsistencies between students' ratings of their skills and reporting of their strongest skills. Also, there are potential cases of students misreporting their skills. Likewise, we found some cases where the tool produced unexpected results when calculating the homogeneity of the skills of a team. We hope this work leads to the design of team formation tools that are more effective for the instructors and the students.

Acknowledgment

Thanks are due to Emily Hastings and Farnaz Jahanbakhsh as the analysis of this paper was based on the data collected from their study on team formation approaches.

REFERENCES

[1] 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.

[2] I. Lykourentzou, A. Antoniou, Y. Naudet and S. P. Dow, "Personality matters: Balancing for personality types leads to better outcomes for crowd teams.," in 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016.

[3] L. C. Brickell, L. C. Porter, L. C. Reynolds and C. R. 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.

[4] 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.

[5] 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.

[6] 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 CHI Conference on Human Factors in Computing Systems, ACM., 2017 .

[7] E. Hastings, F. Jahanbakhsh, K. Karahalios, D. Marinov and B. Bailey, "Structure or Nurture? The Effects of Team-Building Activities and Team Composition on Team Outcomes," unpublished.

[8] D. Dunning, C. Heath and J. M. Suls, "Flawed self-assessment: Implications for health, education, and the workplace.," *Psychological science in the public interest ,* vol. 5, no. 3, pp. 69-106, 2004.

[9] B. C. Hansford and J. A. Hattie, "The relationship between self and achievement/performance measures.," Review of Educational Research , vol. 52, no. 1, pp. 123-142, 1982.

[10] N. Falchikov and D. Boud, ""Student self-assessment in higher education: A meta-analysis."," Review of Educational Research , vol. 59, no. 4, pp. 395-430, 1989

[11] N. Falchikov and J. Goldfinch, "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks.," Review of educational research, vol. 70, no. 3, pp. 287-322, 2000

[12] K. Topping, "Peer assessment between students in colleges and universities.," Review of educational Research , vol. 68, no. 3, pp. 249-276, 1998.