EXAMINATION OF THE EFFECTIVENESS OF A CRITERIA-BASED TEAM FORMATION TOOL

BY

ALBATOOL A. ALAMRI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Professor Brian Bailey

ABSTRACT

CATME is 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 instructors configuration of the criteria, the veracity of students responses to the survey, and the soundness of the algorithm. In this thesis, 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, and potential cases of misreports. Likewise, we found some cases where the tool produced unexpected results when calculating the homogeneity of the skills of a team. Implications for instructors and tool designers to mitigate these problems are discussed.

"You're braver than you believe,
stronger than you seem,
and smarter than you think"
Cristopher Robin

ACKNOWLEDGMENTS

Praise is to Allah by Whose grace, guidance, and mercy good deeds are completed. I would like to express my deepest gratitude to my family, whose love, prayers, encouragement and support were endless. To my beloved father, my partner and counselor in this challenging journey, and to my beloved mother, the shining beacon that guides me in this stormy life. I am deeply thankful to my grandmother and aunts, whom I dearly cherish and miss. I am grateful to my friends who always believe in me, cheer me up, and bring joy to my life.

I would like to thank my dear colleagues Emily Hastings and Farnaz Jahanbakhsh for sharing their work with me and always gladly offering help. Special thanks to Christine Pisarczyk for helping with the analysis of the design projects. Words cannot express my deep gratitude to my adviser, Prof. Brian Bailey who gave me an invaluable opportunity of learning and growth. I am very thankful for his guidance, interest, encouragement, and support

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION . . . 1

CHAPTER 2 RELATED WORKS . . . 3
- 2.1 Influencing Factors on Self-Assessment . . . 3
- 2.2 Influencing Factors on Prediction . . . 4

CHAPTER 3 RESEARCH QUESTIONS . . . 5

CHAPTER 4 METHOD . . . 6
- 4.1 Instructors Selection of the Course Skills and the Projects Outcomes (RQ1) . 6
- 4.2 Students Responses to the Team Formation Survey (RQs 2,3,4) . . . 7
- 4.3 The Tools Heuristic Functions and the Composition Scores (RQ5) . . . 8

CHAPTER 5 RESULTS . . . 9
- 5.1 Correlations Between Skill Distribution and Team Performance (RQ1) . . . 10
- 5.2 Students Assessments of their Skills (RQ2) . . . 14
- 5.3 Skill Possession vs Skill Mastery (RQ3) . . . 16
- 5.4 Potential Misreports: Computer Science Students & Programming Skill (RQ4) 17
- 5.5 Composition Scores (RQ5) . . . 18

CHAPTER 6 DISCUSSION AND FUTURE WORK . . . 20
- 6.1 On the Instructors Configuration of the Criteria . . . 20
- 6.2 On the Accuracy of Students Self-Assessment . . . 21
- 6.3 On the Soundness of the Heuristic Functions of the Tool . . . 22
- 6.4 Implications for Tool Designers . . . 23

CHAPTER 7 CONCLUSION . . . 24

REFERENCES . . . 25

CHAPTER 1: 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 a 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 thesis, 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, design, 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, were significantly higher than those of the random teams: (Criteria-based:μ=9.0, Random: μ =4.9 out of 19 maximum score),(t (35) = 2.42, p = .02).

In the new analysis of the potential issues that may have led to that result, we examined the project deliverables, the data of the team formation survey and the composition heuristics used by the tool to form the teams. Several key issues were identified. First, students were largely distributed into teams based on their reported strongest skills. However,none of the technical skills analyzed showed a considerable correlation with the grade of the project deliverables (programming: r= 0.17, writing: r= -0.23, design: r= 0.1). Second, 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. Third, 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 or graduate students, mostly of whom had GPAs higher than 3.0. (on a four-point scale). Finally, the heuristic function used to score the composition goodness of the course skills 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 skills 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 discussion on areas of improvements.

CHAPTER 2: RELATED WORKS

Self-Assessments can take many forms such as evaluating skills and knowledge or predicting future behaviors and events. Interestingly, the literature on self-assessment shows that peoples assessments in comparison to objective performance measures shows relatively small to moderate correlations at best [8]. Similarly, peoples judgments and predictions of future events or behaviors tend not to prove accurate when the actual situation arrives [8].

2.1 INFLUENCING FACTORS ON SELF-ASSESSMENT

There are several factors that influence peoples assessment of knowledge and skills. Examples of such factors are: the assessors meta-cognitive abilities, the difficulty of the assessed skill, the specificity in defining skill competence, and the desirability of the skill. First, meta-cognition is a term described by Kruger and Dunning [9] as "*the ability to know how well one is performing, when one is likely to be accurate in judgment, and when one is likely to be in error*". The authors argue that incompetent people lack such meta-cognitive ability. That is, their incompetence deprives them from the ability to recognize how incompetent they are. Consequently, when those people are asked to assess their skills, they tend to grossly overestimate their competence [9].

Second, the difficulty of the assessed skill affects peoples assessment. The easier the skill is i.e. the threshold of success is low, the more confident people are in their assessments. In contrast, the more difficult the skill is, i.e. the threshold of success is high, the less confident they become [10]. This is more evident when people assess themselves in comparison to the others. Kruger [10] argues that in such cases, people anchor their assessments on their own level of competence, neglecting the competence level of the comparison group. Consequently, the easier the skill is (e.g. spoken expression), the more likely people would think they are Above-average, ignoring that it can be similarly easy to others. In contrast, the more difficult the skill is (e.g. mechanics), the more likely they would think they are Below-average, forgetting again that it is most likely as difficult to others.

Third, how well the competence in a skill is defined is an important factor in self-assessment. When no clear definition is provided, people rely on their own definitions or interpretations of that skill for assessment [11]. Moreover, the desirability of the assessed skills when coupled with unconstrained definition of competence has far greater influence. For instance, Dunning et al [11] show that people are more likely to rate themselves higher for ambiguous desirable traits such as sensibility or ingenuity, than they would for more constrained traits such as neatness or punctuality. The exact opposite happens with negative traits. People would rate themselves lower on ambiguous negative traits, such as naivety or submissiveness, than on more constrained traits such as sarcasm or wordiness. In other words, social desirability is inherent in peoples assessments. However, constrained definitions of competence help in producing more accurate assessments.

2.2 INFLUENCING FACTORS ON PREDICTION

In their predictions, people tend to be too optimistic and overconfident [8]. Such optimism and confidence stem from several factors. For example, people tend to not consider the unpredictability of future situational details. Also, they tend to forget about how they behaved in similar past experiences. As for the former, despite the incomplete knowledge of the future situational details, people make their prediction as if they have the perfect situational knowledge [12]. Such tendency can be reduced by asking people to describe alternative situations. That would make them realize that there are potential details that cannot be anticipated. Consequently, people would provide more realistic predictions [12].

As for the latter, when people make prediction of future behaviors such as completion of a task, they tend to ignore their past behaviors in similar experiences. That is usually attributed to what is known as the inside view; people typically focus on the nature of the task, their abilities and resources when making prediction. Conversely, in an outside view, people would make their predictions based on the objective facts of their typical behaviors in past comparable experiences [13]. Although it is in the nature of people to predict based on the inside view, incorporating the outside view in the prediction helps people provide more accurate predictions [14].

Relating these works to this tool shows that tools that solely depend on students assessments of their knowledge and abilities are highly prone to the effects of the inherent flaws of this approach. The questions in the team-formation survey that corresponds to the selected criteria could be factual (gender), preferential (leadership role), predictive (commitment level), or evaluative (skill level). It is in answering the predictive and evaluative questions that the flaws of self-assessment may manifest. Therefore, more robust methods of assessments are needed for such tools to work effectively.

CHAPTER 3: RESEARCH QUESTIONS

As the effectiveness of this tool depends on the instructors configuration of the criteria, the accuracy of students responses to the survey, and the soundness of the algorithm, we posed the following questions to guide our examination of the data:

On the Instructors Configuration of the Criteria:

- **RQ1**: Did the distribution of the skills specified in the criteria correlate with teams performance in the projects?

On Students Responses to the Team Formation Survey:

- **RQ2**: Given that criteria questions could be evaluative, were students able to adequately evaluate their skills?

- **RQ3**: Which is more effective, asking about the possession of a skill or the mastery of it?

- **RQ4**: Were there instances where students intentionally provided inaccurate responses?

On the Algorithm and Heuristic Functions of the Tool:

- **RQ5**: How is the composition score of a team computed? And is it accurate?

CHAPTER 4: METHOD

To assess the soundness of the factors influencing the effectiveness of the tool, we examined the project deliverables, the team formation survey, and the composition scores.

4.1 INSTRUCTORS SELECTION OF THE COURSE SKILLS AND THE PROJECTS OUTCOMES (RQ1)

The criteria used to form the teams included: Gender, Leadership Preference, Writing Skill, and Course Skills (teamwork, programming, writing, speaking, design). To see if the distribution of the course skills had a discernible effect on teams performance in their projects (RQ1), we examined the project deliverables in relation to each skill. The project required the design and development of a functional user interface. The deliverables of the project included: Project Proposal, User Research Report, Low-Fidelity Prototype, User Evaluations of the Low-Fidelity Prototype, Functional Prototype, and User Evaluations of the Functional Prototype.

The examination included conducting Single-Linear Regression analysis for each of the following hard-skills: programming, writing, and design. The teamwork and communication (speaking) skills were not considered as their effects are implicit and cannot be directly measured from the project deliverables. In each regression analysis, the independent variable was the number of students reporting the examined skill in each team. Those numbers were obtained from students responses to the Course Skills question. The dependent variable was the team performance, measured by grade, in the project deliverables specific to that skill. Explicitly, grade was calculated by taking the average of the unweighted original grades of the specified deliverables.

As for Programming, the project deliverable considered was the functional prototype as it was the only deliverable involving actual implementation and development. However, as this skill is not as visible as the other skills, we included a second measurement of performance in addition to the grade of that deliverable. We used a code analysis tool, Code::Stats [15], to measure the productivity in the projects, i.e. the number of lines of code excluding spaces and comments. It is important to note that there were some projects that were not included in the examination for the following reasons: There were teams that submitted their projects in APK format only, in which the original code cannot be retrieved, hence hindering the possibility of analysis. In addition, some teams used a variety of technologies, to build web pages for instance, instead of the recommended Android Studio for the development.

That made the analysis across different platforms difficult. Therefore, only 18 projects were included in the examination.

For Writing, we considered only the project components delivered in the form of reports: User Research Report, User Evaluations of the Low-Fidelity Prototype, and User Evaluation of the Functional Prototype. Project proposal was not included as it was delivered in the form of a presentation. For Design, we considered only the Low-Fidelity Prototype and the Functional Prototype. An additional analysis was conducted for each of these skills to see if there was a difference in performance between the randomly formed teams or the criteria-based teams. The rational is that performance of teams in the prior study was measured with the project grade, a cumulative grade of weighted scores of the deliverables. Since the holistic performance between conditions did not differ significantly, we wanted to examine teams performance based on skill.

4.2 STUDENTS RESPONSES TO THE TEAM FORMATION SURVEY (RQS 2,3,4)

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 Course Skills question was phrased as: What is your strongest skill(s) as it relates to a design project in the course?, and the Writing Skill question was phrased as Rate your writing skill.

4.2.1 Adequacy of Students Self-Assessments

To see if students were able to evaluate their skills adequately (RQ2), we examined their responses to two skills: writing and teamwork. Those two skill, unlike the rest of skills, were assessed more than once. As for the writing skill, since it was in both evaluative questions, we compared students answers for the skill level question with the skill possession question.

As for the teamwork skill, we used CATMEs peer-assessment that was collected after the completion of projects. In this assessment, students rate themselves and their peers in five categories (Contribution, Interaction, Keep on Track, Expect Quality, Knowledge/skills). Each rating is on a 5-pt scale, where 1 is Poor, 3 is Satisfactory and 5 is Outstanding [16]. Using this assessment, we evaluated the performance of those who reported teamwork skill. Performance was measured by aggregating the averaged peer ratings in those five categories; lowest score is 5 (Poor), middle score is 15 (Satisfactory), and highest score is 25 (Outstanding). Additionally, we examined the agreement level between students rating of themselves and their peers.

4.2.2 Skill Possession vs Skill Mastery

To see which of the two evaluative question is more effective (skill possession or skill mastery) (RQ3), we conducted another Single-Linear Regression analysis for the Writing Skill question. We then compared its result with the previous regression analysis for the writing skill that was based on the Course Skills question.

4.2.3 Potential Misreports

Since the course is a Computer Science course, we examined the demographics of the students who did not report having a programing skill (RQ4).

4.3 THE TOOLS HEURISTIC FUNCTIONS AND THE COMPOSITION SCORES (RQ5)

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. Also, it uses 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 (RQ5).

CHAPTER 5: RESULTS

Tables 5.1 and 5.2, and Figure 5.1 provide an overview analysis of students responses to the Course Skills question: "What is your strongest skill(s) as it relates to a design project in the course?". Teamwork was the most reported skill (72.16%), and in most selections of two or more skills. Programming, second most selected skill (69.32%), was the most reported among those reporting only one skill. Following teamwork skill and programming, there is an interesting drop in numbers (∼28%) for the rest of reported skills. Writing, speaking and design are close in numbers. Yet, between those three, writing and speaking were more closely related.

Table 5.1: Skills ordered by the number of students reporting them

Teamwork	**Programming**	**Writing**	**Speaking**	**Design**
127	122	74	66	63
72.16%	69.32%	42.04%	37.5%	35.8%

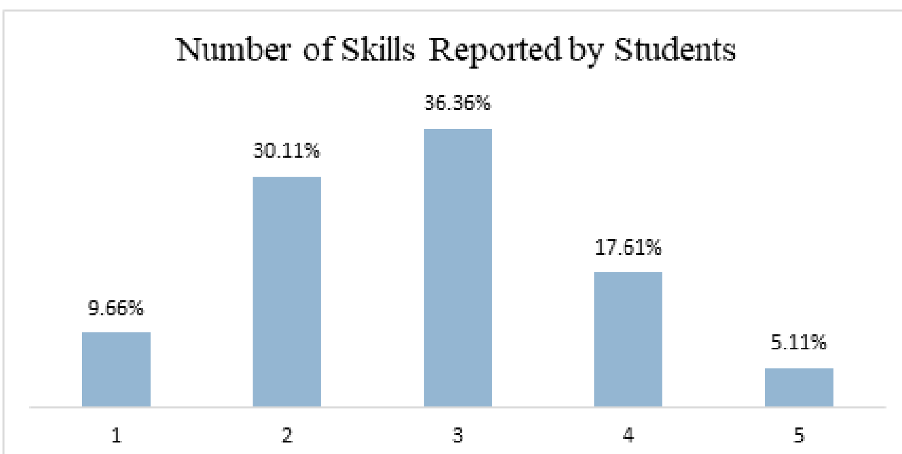

Figure 5.1: Number of skills reported by students

Table 5.2: Most reported skill choices

Skills Reported	**Top Three Skill Choices**	**No. of Students**
1	Programming	9/17
	Team Skills	3/17
	Speaking	2/17

Table 5.2: Most reported skill choices

Skills Reported	Top Three Skill Choices	No. of Students
2	(Teamwork, Programming)	16/53
	(Teamwork, Writing)	9/53
	(Programming, Design)	8/53
3	(Teamwork, Programming, Writing)	13/64
	(Teamwork, Programming, Design)	11/64
	(Teamwork, Writing, Speaking)	9/64
4	(Teamwork, Programming, Writing, Speaking)	12/31
	(Teamwork, Programming, Design, Speaking)	4/31
	(Teamwork, Programming, Writing, Design)	4/31

5.1 CORRELATIONS BETWEEN SKILL DISTRIBUTION AND TEAM PERFORMANCE (RQ1)

The results of the regression analyses between the number of students with a certain skill per team and their performance in project deliverables related to that skill showed the following:

5.1.1 Programming Skill

Since programming is a skill not as visible as the other skills, we had two measurements of performance: the grade of the functional prototype, and the number of lines of the code (productivity). Figure 5.2 shows the regression analysis between the number or programmers per team and the grade of the functional prototype. The number of programmers per team was determined by students reporting programming in the Course Skill question.

The result showed small correlation ($r = +0.17$), showing an increase in grade by (0.26) for each addition of a programmer to the team. However, no significant regression equation was found (F (1,16) =0.47, p=.0.51).

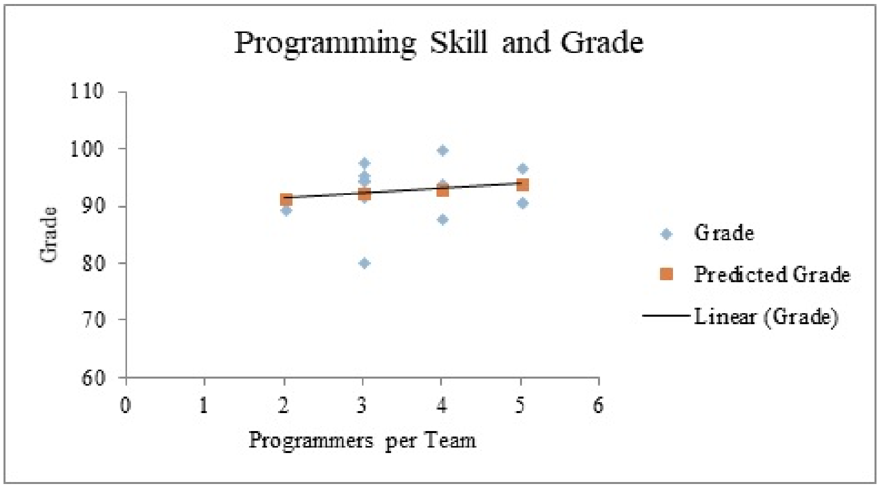

Figure 5.2: Regression analysis between grade and the number of programmers per team as determined from the course skill question

Figure 5.3 shows the regression analysis between the number or programmers per team and the code productivity, i.e. the number of lines of code excluding spaces and comments. These numbers were obtained from Code::Stats [15], a code analysis tool. The result showed a stronger correlation (r =+0.31), with an increase in productivity by 250 lines for each addition of a programmer to the team. Yet, no significant regression equation was found (F (1,16) =1.64, p=.22).

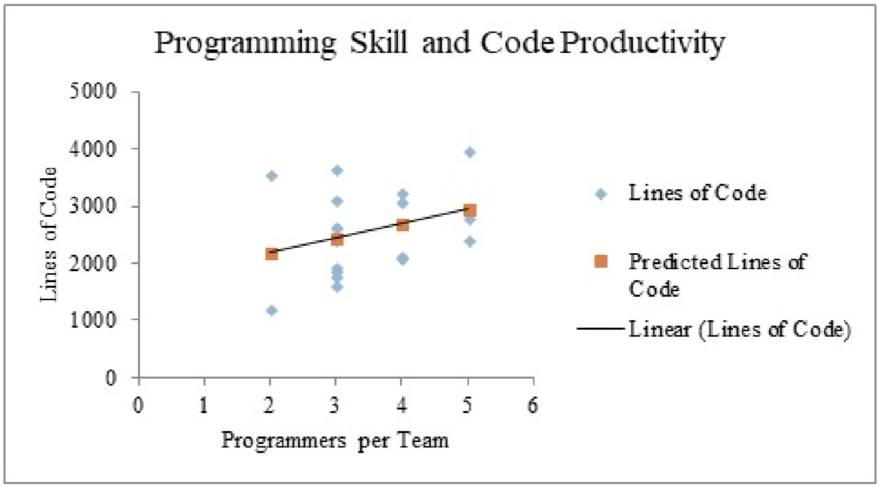

Figure 5.3: Regression analysis between the lines of code and the number of programmers per team as determined from the course skill question

We conducted an additional regression analysis to see how the number of code lines correlate with grade, see Figure 5.4. Interestingly, the correlation was strong (r = +0.51), and significant regression equation was found (F (1,16) =5.72, p=.0.029). However, for each additional 1000 lines, grade increased by 1 only.

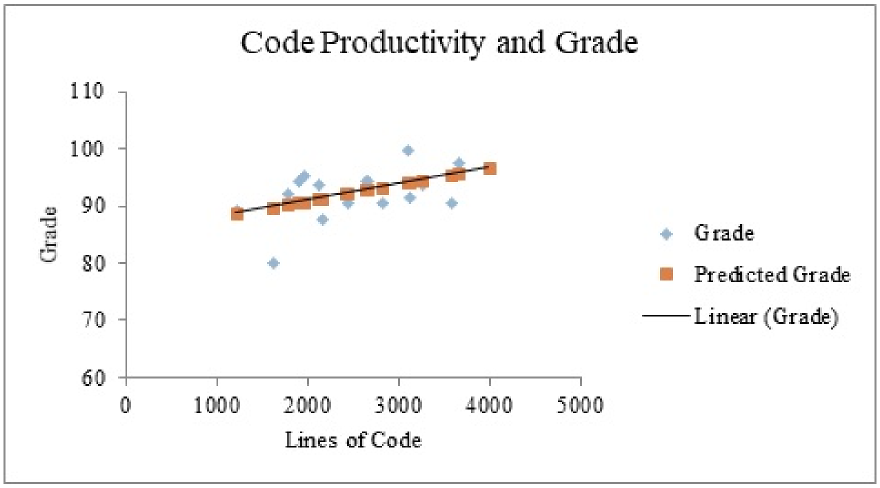

Figure 5.4: Regression analysis between grade and the lines of code in each teams project

5.1.2 Writing Skill

Figure 5.5 shows the regression analysis between the number of writers per team and the grade of the written deliverables. The result showed negative small correlation ($r = -0.23$). With each addition of a writer, the grade decreases by (-1.3). No significant regression equation was found ((F (1,35) =1.98, p=0.17).

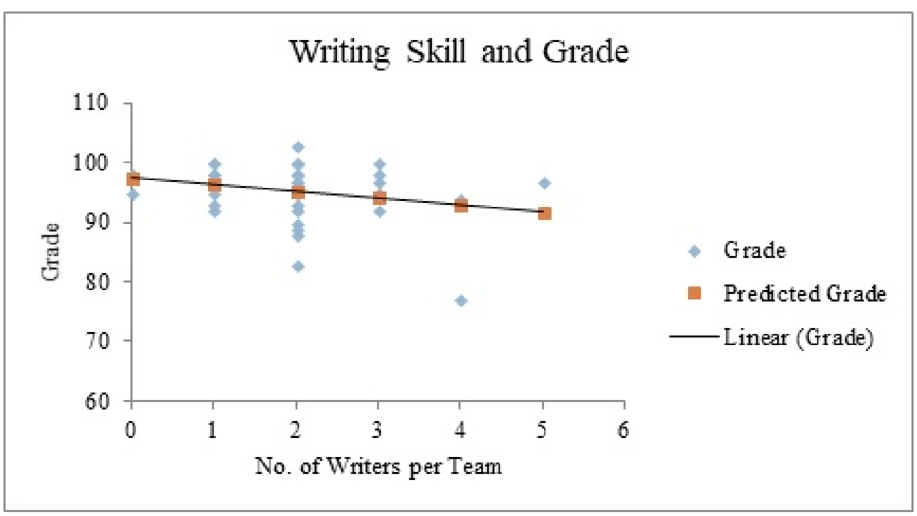

Figure 5.5: Regression analysis between grade and the number of writers per team as determined from the course skill question

5.1.3 Design Skill

The number of designers per team and grade barely correlated ($r = +0.1$), see Figure 5.6. Grade increases by 0.4 for each addition of a designer. No significant regression equation was found too, (F (1,35) =0.36, p=0.56).

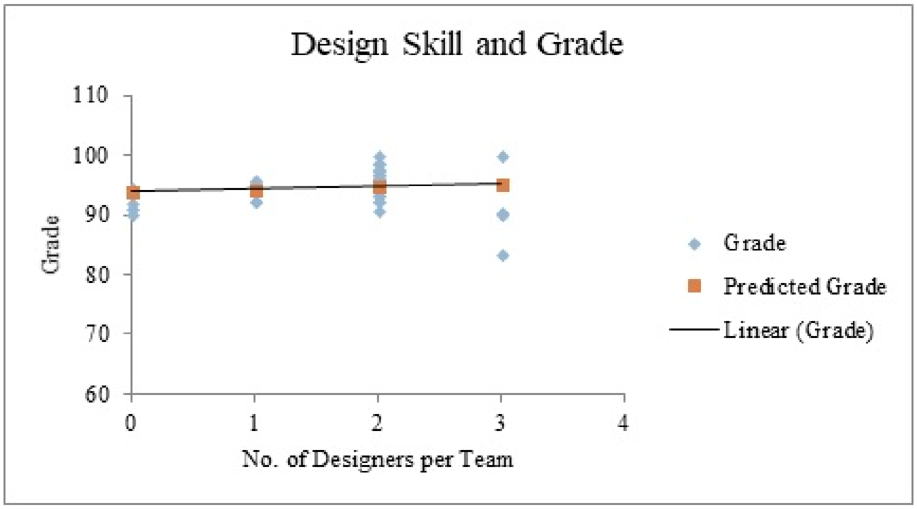

Figure 5.6: Regression analysis between grade and the number of designers per team as determined from the course skills question

5.1.4 Performance of Random Teams and Criteria-based Teams in terms of Skills

We analyzed teams performance between the conditions based on each skill. The results, seen in Table 5.3, were similar to the previous study; No significant difference in performance between the conditions for any of the skills was found.

Table 5.3: Team performance between conditions based on skills as measured from the project deliverables

Skill	Performance Measure	Test	Value	P
Programming	Lines of Code	U(7,11)	41	0.860
	Grade	U(18,19)	183.5	0.707
Design	Grade	U(18,19)	184.5	0.685
Writing	Grade	U(18,19)	167.5	0.916

5.2 STUDENTS ASSESSMENTS OF THEIR SKILLS (RQ2)

5.2.1 Teamwork Skill Students Assessments and Peers Assessment

Figure 5.7 shows students reporting of the teamwork skill and their performance level in their teams as assessed by their peers via CATME. For those who reported teamwork skill, Figure 5.8 shows the agreement level between their assessments of their teamwork performance and their peers assessment of them.

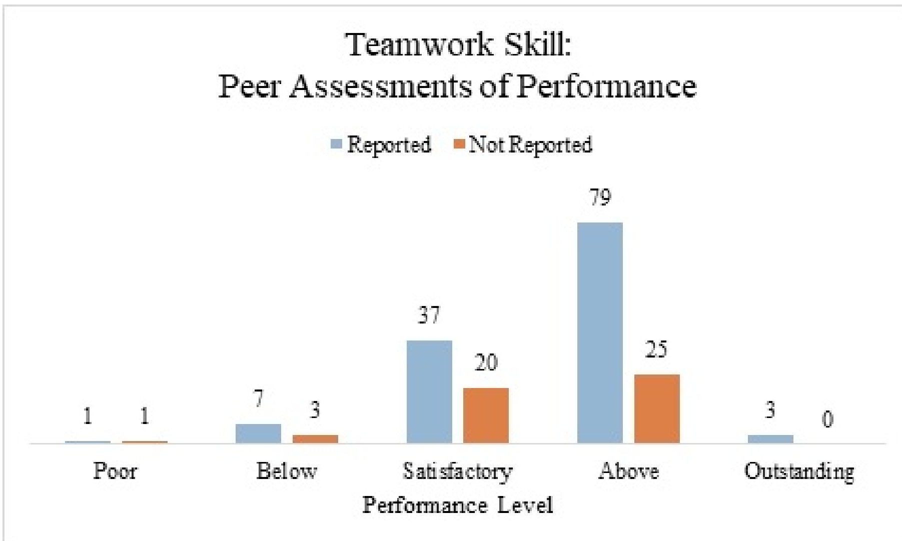

Figure 5.7: Teamwork skill reports in the course skill question and Peer-Assessment of teamwork performance

From Figures 5.7, we see that (∼14%) of students whose performance level is above satisfactory did not consider teamwork skill to be one of their strongest skills. Nevertheless, the majority of those who reported teamwork skill have a satisfactory performance or better. Moreover, for those who reported teamwork skill, Figure 5.8 shows that almost all students assessments of their performance matched their peers assessment of them. Those with mismatching assessments are shown in Table 5.4. The cause of disagreement, provided in CATME, is that those with performance level below satisfactory overrated themselves. In contrast, those with a performance level above satisfactory underrated themselves.

The figures and the table give one more information; of those with a performance level below satisfactory, five of them reported teamwork skill, and their assessments of themselves matched their peers. To elaborate, those students in the beginning of the semester reported teamwork skill to be one of their strongest skills. Yet, when they completed CATMEs peer assessment near the end of the semester, they assessed their performance as below satisfactory.

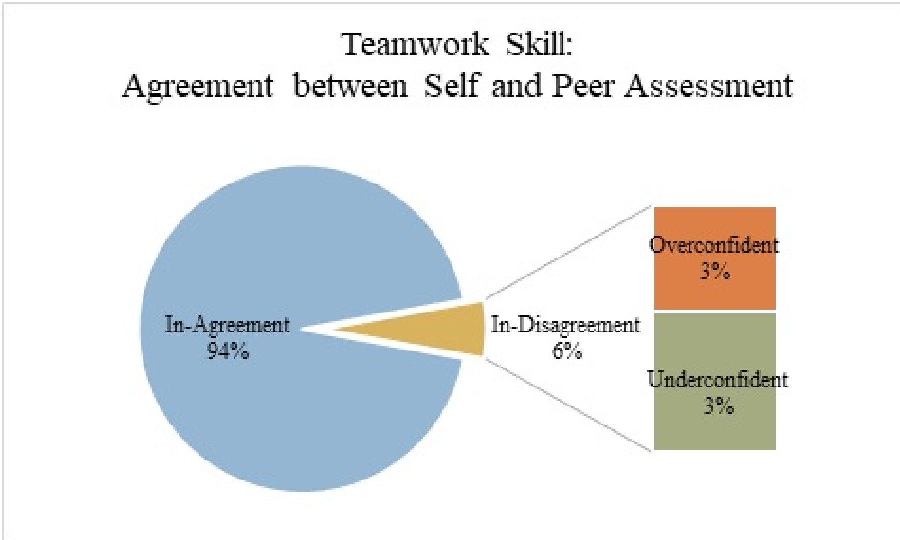

Figure 5.8: Agreement level between self-assessment and peer-assessment of teamwork performance for those who reported teamwork skill

Table 5.4: Performance level of the students whose self-assessment differed from their peers'

Disagreement Cause	Performance Level	No. of Student
Over-confidence	Poor	1
	Below	2
Under-confidence	Above	3
	Outstanding	1

5.2.2 Writing Skill Students Assessments of Skill Possession and Skill Mastery

The results of comparing the students responses between the Writing Level question and the Course Skills question is shown in Figure 5.9. 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.

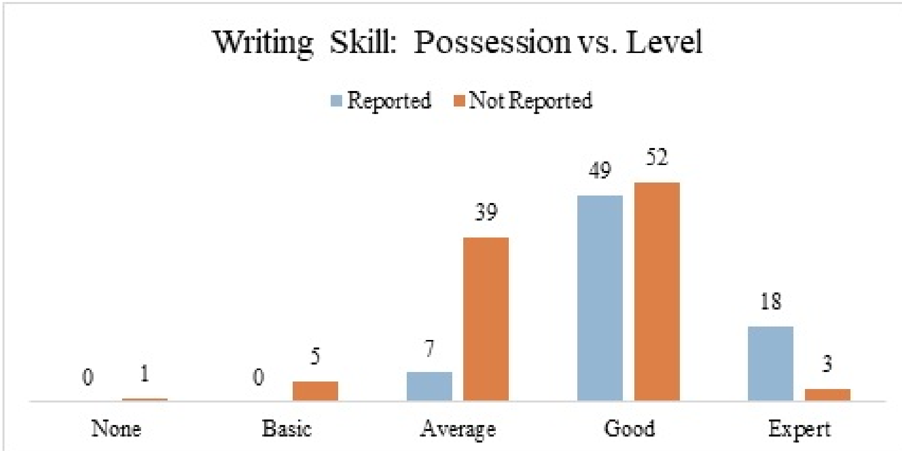

Figure 5.9: Students responses about their writing skill to the Writing Level and Course Skills questions

5.3 SKILL POSSESSION VS SKILL MASTERY (RQ3)

Motivated by the results of RQ2, we conducted an additional single-linear regression analysis for the Writing Level question to know which is more effective to ask about, skill possession or skill level. We then compared its results with the previous regression analysis where writing skill was reported in the Course Skill question.

In the previous analysis (skill possession), the number of writers was determined by those reporting writing as one of their strongest skills. In this analysis (skill mastery), the number of writers was determined by those indicating a writing level of Good (4) or Expert (5) on a scale of 5 in the writing level question, Figure 5.10. 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 results for the mastery question showed stronger correlation ($r = -0.34$) than the results of the possession question ($r = -0.23$). Furthermore, unlike with the possession question analysis, a significant regression equation was found (F (1, 35) =4.56, p=0.04), R^2 of 0.12. That is, the number of good and expert writers in a team can predict their grades in the written components of their projects. Noteworthy the grade deceased in both analyses for each addition of a writer to the team; decrease by (− 1.13) in the skill possession analysis, and by (− 1.62) in the skill mastery analysis.

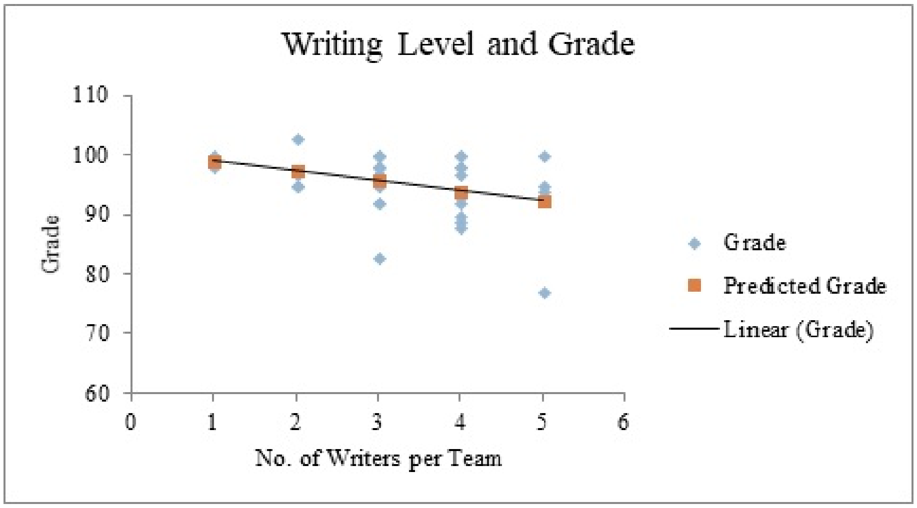

Figure 5.10: Regression analysis between grade and the number of good and expert writers per team as determined from the skill level question

5.4 POTENTIAL MISREPORTS: COMPUTER SCIENCE STUDENTS & PROGRAMMING SKILL (RQ4)

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 Figure 5.11. Moreover, 10 of them had GPAs above 3.5 (on a four-point scale) indicating high academic performance, Table 5.5.

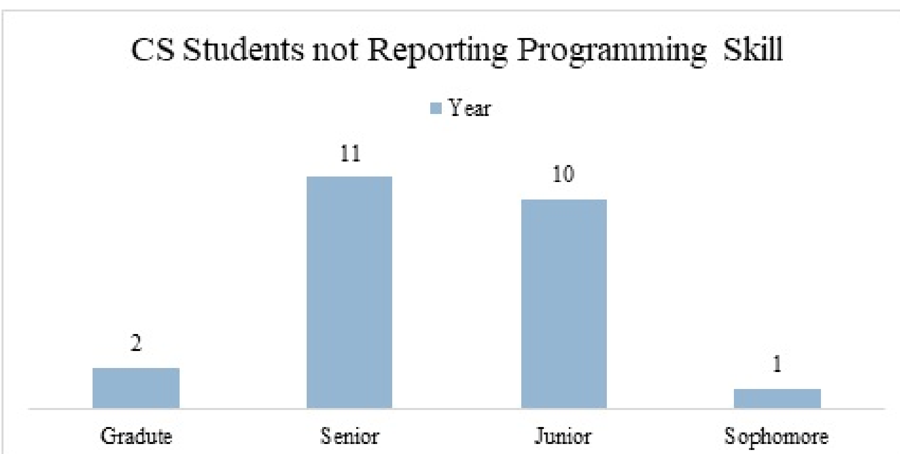

Figure 5.11: Computer Science students who did not report programming skill

Table 5.5: Year and GPAs of the CS students who did not report programming skill

	Graduate	**Senior**	**Junior**	**Sophomore**
GPA $\geq$ 3.5	2	5	3	-
3.5> GPA $\geq$ 3.0	-	4	6	-
GPA <3	-	2	1	1

5.5 COMPOSITION SCORES (RQ5)

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. Figure 5.12 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

Figure 5.12: 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 CATMEs 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 Course Skills 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 CATMEs 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.

CHAPTER 6: DISCUSSION AND FUTURE WORK

The factors influencing the effectiveness of the criteria-based team formation tool are: the instructors selection and configuration of the criteria, the accuracy of students responses to the survey, and the soundness of the algorithm. The results of our analysis give insights into potential issues with these factors.

6.1 ON THE INSTRUCTORS CONFIGURATION OF THE CRITERIA

When the team formation criteria are configured, the team formation survey is constructed, and each criterion is associated with a suitable question to elicit the needed information from students. The criteria could be either selected from the criteria list supported by CATME or added manually by the instructor. CATMEs criteria comes with pre-defined questions, whereas new criteria need its questions to be defined by the instructor. In the examined team formation survey, the instructor used both types of criteria (existing and new). For demographics and work styles, existing criteria in CATME were used. For the course skills, the instructor defined a new question that was phrased as "What is your strongest skill(s) as it relates to a design project in the course?". The answers reflected the following skills: teamwork, programming, writing, speaking, and design. To have skill diversity in teams, students were grouped dissimilarly based on that criterion.

The results of RQ1 gives insights into how skill distribution correlated with teams performance in their projects. First, none of the analyzed skills (programming, writing, and design) showed a considerable or significant correlation with the grade of the project deliverables (programming: $r = +0.17$, writing: $r = -0.23$, design: $r = +0.1$). This suggest that skill distribution based on students assessment of their strongest skills is not a significant predictor of their project grades.

Second, besides grade, programming skill was assessed by a second performance measure (code productivity). This measure showed a moderate correlation ($r = +0.31$) with the number of programmers per team. Interestingly, the correlation between the two performance measures showed a significantly stronger correlation ($r = +0.51$, $p < 0.05$). This indicates that higher coding productivity is not necessarily due to the increase of programmers in a team. Rather, it is the productivity of a few strong contributors in the team.

Third, some of the correlations of the skills were positive (programming and design) while others were negative (writing). At the surface level, this might cause confusion as to how to distribute skills. However, considering the regression equations and the beta values of all the analyzed skills gives a clearer picture. In the positive correlations, each addition of a student with a specific skill had a negligible effect on grade; it increased by (0.26) for each additional programmer and by (0.4) for each additional designer. In the negative correlation, however, we see a larger decrease in grade (-1.3) for each additional writer in the team. This suggests that balancing the number of students with a skill in team is better, and the desired increase in grade can be achieved by improving the performance of those within the teams already.

6.2 ON THE ACCURACY OF STUDENTS SELF-ASSESSMENT

The results of RQ2 gives insights about students assessment of their skills. In the examination of the writing and teamwork skills, we noticed a pattern; in each skill, there was a number of high performers who did not consider that skill to be one of their strongest skills. In the writing skill, about 70% of students indicated a writing level of good (4) or higher on a 5-pt scale. However, about 45% of that percentage did not report writing as a one of their strongest skills. Similarly, in the teamwork skill, ∼59% of students had a teamwork performance level above satisfactory (4 on a 5-pt scale). Yet, about 24% of that percentage did not report teamwork as one of their strongest skills. These results show that students self-assessments are influenced by the phrasing of evaluative questions; they may be competent in a skill, but they may not necessarily view that skill to be one of their strongest skills.

In addition, the literature on Self-Assessment shows that the difficulty of the assessed skill influences peoples assessment; the more difficult the skill is, the more likely people think of themselves as below-average [10]. This explains the contrasting difference in the numbers of reporters and non-reporters in each of the levels of "Average" and above between the writing skill and the teamwork skill, see Figures 7 and 9. Students with a skill level of average and above showed greater confidence in reporting the teamwork skill compared to reporting the writing skill.

The results of RQ3 shows which evaluative question, skill possession or skill mastery, correlates more with grades. The results showed 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 course skill 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 strongly recommended to provide a clear description to each level of the skill to reduce the effect of subjective assessment [11].

As for the results of RQ4, which investigated if students intentionally misreported their skills, we provide a couple of explanations as to why students did not report having skills they are expected to have. Again, the phrasing of the course skills 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. However, it is possible that some student may have intentionally misreported their expected skills. To clarify, 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 out of the fear of handling or performing larger shares of workload in their projects. It also highlights that some students may try to game the system to satisfy personal goals.

6.3 ON THE SOUNDNESS OF THE HEURISTIC FUNCTIONS OF THE TOOL

The findings of RQ5 shows that the heuristic function of the Choose Many of, which the course skills question was based on, gives unexpected scores when the goal of the question is to have diverse skills within a team. The reason for such unexpected scores is the assumption of commonality between team members that the heuristic function works with. Figure 12 shows how the 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. Consequently, as this heuristic function cannot satisfy the goal of diversity, it would lead the tool to generating 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.

6.4 IMPLICATIONS FOR TOOL DESIGNERS

The main concern regarding the effectiveness of this tool is its sole dependency on students self-reports. In answering the predictive and evaluative questions such as commitment level or skill level., the flaws of self-assessment manifest. Research shows that peoples assessments of their knowledge and skills in correlation to objective performance measures tend to be relatively small, moderate at best [17] [18]. In addition, people tend to be overconfident in their judgments and predictions of future events or behaviors, which do not always prove to be accurate when the actual situation arrives [8]. Furthermore, there are potential cases of inaccurate reports regardless of the reason behind such behaviors.

It would be of immense value to test the generalizability of our findings by analyzing similar dataset from different courses. Nevertheless, this thesis sufficiently motivates revising the tool in accordance with the insights obtained from analyzing the data. Specifically, more effective means of skill assessment in the tool are need. 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 [19] [20]. Equally important, the tool needs to be more resistant to any attempts of gaming the system or manipulating the outcomes of the team formation process.

CHAPTER 7: CONCLUSION

In this thesis, 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. First, the technical skills analyzed showed no considerable correlation with the grade of the project deliverables. Second, 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. Finally, we found some cases where the tool produced unexpected results when calculating the homogeneity of the skills of a team. We hope this thesis leads to the design of team formation tools that are more effective for the instructors and the students.

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] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[10] J. Kruger, "Lake wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments," *Journal of personality and social psychology*, vol. 77, no. 2, p. 221, 1999.

[11] D. Dunning, J. A. Meyerowitz, and A. D. Holzberg, "Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability," *Journal of personality and social psychology*, vol. 57, no. 6, p. 1082, 1989.

[12] D. W. Griffin, D. Dunning, and L. Ross, “The role of construal processes in overconfident predictions about the self and others,” *Journal of Personality and Social Psychology*, vol. 59, no. 6, p. 1128, 1990.

[13] R. Buehler, D. Griffin, and M. Ross, “Inside the planning fallacy: The causes and consequences of optimistic time predictions,” in *Heuristics and biases: The psychology of intuitive judgment*, T. Gilovich, D. Griffin, and D. Kahneman, Eds. New York, NY, US: Cambridge University Press, 2002, pp. 250–270.

[14] R. Buehler, D. Griffin, and M. Ross, “Exploring the “planning fallacy”: Why people underestimate their task completion times,” *Journal of Personality and Social Psychology*, vol. 67, no. 3, p. 366, 1994.

[15] M. Ahlroth, “Welcome to code::stats,” 2018. [Online]. Available: https://codestats.net/

[16] M. W. Ohland, M. L. Loughry, D. J. Woehr, L. G. Bullard, R. M. Felder, C. Finelli, R. A. Layton, H. R. Pomeranz, and D. G. Schmucker, “The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self-and peer evaluation,” *Academy of Management Learning & Education*, vol. 11, no. 4, pp. 609–630, 2012.

[17] 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.

[18] 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.

[19] 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.

[20] K. Topping, “Peer assessment between students in colleges and universities,” *Review of educational Research*, vol. 68, no. 3, pp. 249–276, 1998.