Gender, Confidence, and Mark Prediction in CS Examinations

Brian Harrington
Dept. of Computer and Mathematical Sciences
University of Toronto Scarborough
Canada
brian.harrington@utsc.utoronto.ca

Shichong Peng
Dept. of Computer Science
University of Toronto
Canada
shichong.peng@mail.utoronto.ca

Xiaomeng Jin
Dept. of Computer Science
University of Toronto
Canada
tracy.jin@mail.utoronto.ca

Minhaz Khan
Dept. of Computer and Mathematical Sciences
University of Toronto Scarborough
Canada
minhaz.khan@mail.utoronto.ca

ABSTRACT

A common refrain heard by instructors of CS1 courses is “I’m sure I did better than that" or “I have no idea how I got that mark". Sometimes differences in a student’s expected and reported marks may be due to a mistake on the part of the grader of the work, but more often than not this is an indicator that a student is not accurately assessing their own level of achievement on a piece of work. This may be an issue of capability (some students may lack the tools to assess what they have done correctly or incorrectly?) or one of confidence (some students are certain they are making mistakes even if they have done everything correctly). Regardless of the source of the gap between predicted and reported grades, it is an important skill for students to be able to accurately assess their own capabilities and performance.

In this work, we asked students in a CS1 course to predict their own grades on each question of their final examination. Analyzing the actual and predicted grades, and the differences between them leads to several interesting results. Poorer performing students are more likely to overestimate their grades, while better performing students are more likely to underestimate their grades. Furthermore, performance on the exam is strongly correlated with ability to correctly predict marks. Perhaps most interestingly, we found that while there was no difference in performance of male and female students on the exam, female students were more likely to under predict their performance than their male counterparts.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; **Computer science education**; **CS1**;

KEYWORDS

Assessment, CS1, Confidence, Examinations, Gender

ACM Reference Format:
Brian Harrington, Shichong Peng, Xiaomeng Jin, and Minhaz Khan. 2018. Gender, Confidence, and Mark Prediction in CS Examinations. In *Proceedings of 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE’18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3197091.3197116

ITiCSE’18, July 2–4, 2018, Larnaca, Cyprus

ACM ISBN 978-1-4503-5707-4/18/07…$15.00
https://doi.org/10.1145/3197091.3197116

1 INTRODUCTION

Students in CS1 courses can vary widely in both their technical abilities and their confidence in those abilities. It would be reasonable to assume that confidence correlates highly with aptitude, but it is not uncommon to interact with students who have very inaccurate beliefs in their own achievement within a course. Moreover, students often “don’t know what they don’t know”, which can lead to students who are struggling in a course actually believing they are doing well or students who are at the higher end of the grade spectrum becoming disheartened and believing that they are performing poorly in the course.

This may be explained in part by the Dunning-Kruger Effect [9] which causes individuals with limited experience to over assess their own abilities in a field, and to only gain a more realistic perspective of their own capabilities and level of achievement as they gain a wider understanding of the scope of the field. Alternatively, this could be simply an artifact of students having natural differences in their level of self-confidence.

Two groups for whom confidence has been shown to be an issue in CS courses are students with no prior programming experience, and female students. These groups have higher dropout rates in CS programs, and confidence may play a factor in that trend.

In this study, we asked students to predict their own marks on their final examination of their CS1 course. This was done as a bonus question, so students were providing these estimates in very close temporal proximity to completing the questions. We used overestimation (predicted mark - actual mark) as a proxy for confidence to analyze how various factors affect the confidence of CS1 students.

Our data shows that in general, students overestimated their grades, with better performing students being better at predicting their marks. Students who out-performed expectations on the

exam relative to their prior performance were less likely to overestimate their marks. Students with prior programming experience performed better in the course overall, but interestingly were more overconfident than their peers with no prior programming experience. This goes against the general trend of better performing students being less overconfident. There was no difference in exam performance between male and female students, but female students had lower confidence than males.

This paper describes a novel, low-cost, easily replicable method for using exam mark over-prediction as a proxy for confidence, and uses this method to analyze the effect of a variety of factors on student confidence in CS1. In particular, we find evidence that students with prior experience in CS appear to be overconfident and female students appear to be less confident than their male counterparts. While this is only a single study on a single cohort of students, it provides valuable empirical confirmation of commonly held beliefs, and we hope that the ease of replication will lead to further research.

2 BACKGROUND

As a great number of research findings indicate, self-efficacy, a major factor of self-confidence, plays an important role in academic success [5, 8]. Different measurement methods have been applied to evaluate self-confidence level. A study using the third-person effect, which compares self-prediction of students from various programs at University of Newcastle to their prediction of a third person's performance, found that female students were likely to under predict their own performance, and were less confident in predicting the performance of others [1]. Using cognitive and metacognitive awareness tests has also shown that certainty of unsure events could be used as a measurement of self-confidence[7].

Several studies have specifically researched confidence and mark estimation in students. Particular focus has been put on the difference in students' predictive and actual performance to measure overestimation as a distinct factor of overconfidence[10]. The effect of gender on student confidence has also been studied. While a few studies on gender effects between members in a self-rating agreement group have found that gender was not a statistically significant factor among students in general[12], many studies have shown that female students are more likely to underestimate their performance in completing computer science courses and while completing programming tasks[2, 4, 6, 11, 13].

3 METHODOLOGY

CSCA08 is a CS1 course at the University of Toronto Scarborough aimed at students planning on pursuing a major or specialist degree in computer science. The course is taught in Python and covers fundamental computer science concepts, design, documentation and OOP principles. The course evaluation consists of exercises, assignments, quizzes, two term tests and a final exam worth 40% of the grade for the course.

The final examination in the Fall term of 2017 consisted of four normal questions and four bonus questions. The questions on the examination consisted of a tracing question where students were asked to write the output of a program, an open ended question where students were asked to comment on a given piece of code and explain why it was poorly designed, a Code Mangler [3] style question where students were required to re-arrange scrambled code, and a design question where students were asked to produce UML class diagrams for a program given a text description. The bonus questions consisted of one subquestion soliciting feedback on the course (students received marks for giving any piece of insightful feedback on how the course could be improved for future offerings), one subquestion rewarding students for following instructions (the question instructed students to read all the full question carefully before starting, asked the students to name as many of the course TAs as they could, and then said to ignore all previous instructions and simply leave the question blank for full marks), and one subquestion giving students an opportunity for freedom of expression (the question simply said to write or draw something interesting/nice/funny in the space below to receive the bonus mark).

The final[1] bonus question stated:

> Predict your mark on each question of this examination. If your estimate is within 10% of your actual grade, you get this bonus mark.

There was then a space for the student to write their predicted grades for each question, as well as a predicted total out of 55. The self-grading question was actually marked more leniently than indicated, with any student giving a reasonably plausible mark being awarded the bonus grade. 564 students wrote the exam, with 542 predicting their own marks. Two students were excluded from the study for offering implausible predictions (one student predicted a mark well above 100%, and another wrote a series of equations that could not be resolved into a number).

During the marking of the examinations, the predicted marks were also recorded. Only the final predicted grade is used in this paper, though the individual question predictions were recorded for future study.

In order to assess prior programming experience, all students completed a survey at the beginning of term asking them to choose which one of the following statements most directly applied to them:

- "I have never written a line of code in my life"
- "I have played around with coding a bit, but never written anything beyond a few lines. I don't know how to/feel confident writing loops or if statements"
- "I have written multi-line programs, using ifs and loops to build a function or method"
- "I have written many large programs combining multiple functions/methods and am familiar with object oriented programming"

Of the students who wrote the exam, 36 chose the first option, 80 chose the second option, 205 chose the third option and 59 chose the final option. For the purposes of this study, we considered anyone who chose the first two options to not have prior programming experience, and those who chose the latter two options to have prior programming experience.

To determine the gender of the students, in the pre-course survey, each student was asked to self identify as one of: *Male, Female,*

[1] this was actually the 2nd bonus question chronologically in the test

Transgender, or *other* (with the option to fill in a text box if *other* was chosen. There was also an option to choose *prefer not to say*. Of the students who wrote the exam, 285 identified as Male, and 87 identified as Female. No students chose any other option.

In order to determine exam performance relative to expectations, we calculated a linear best fit expectation for students given their marks on all term work (Assignments, Exercises, Quizzes and Term Tests) relative to their examination scores. Then, any student whose exam score was above the expected value was labeled as having *overperformed* and any student whose exam score was below the expected value was labeled as having *under performed*. Similar processes were completed to find expected exam values given only non-test-like coursework (Assignments and Exercises) and given only prior test-like coursework (Quizzes and Term Tests).

4 RESULTS

After collecting the data, the first obvious question was whether students are able to predict their own marks with any degree of accuracy. Plotting the students' predicted grades vs the actual grade received, as seen in Figure 1, shows a relatively strong correlation ($r = 0.69$). Spearman's rank-order correlation also shows a high coefficient ($\rho = 0.68$, $p < 0.001$), which indicates that in general, a student who does better than his/her colleague will be likely to give a higher predicted mark. In general, students overpredicted their marks, with an actual exam average of 53.3%, and a predicted average of 71.9%.

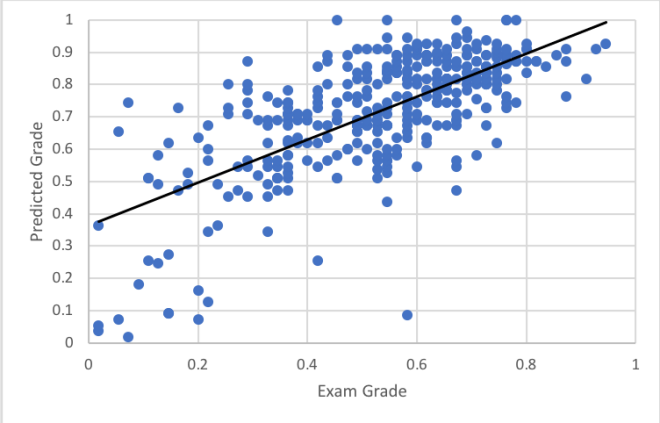

Figure 1: Predicted vs. Actual Grades

Once we established that students have some reasonable ability to predict their own grades, the next question was whether that ability was in any way correlated with their performance in the examination itself. Figure 2 shows the correlation of prediction error (the percentage by which a student overpredicted their exam mark) with the grade achieved by that student in the exam. A moderate negative correlation is evident ($r = -0.43$), and this is confirmed by a Spearman's rank-order coefficient ($\rho = -0.48$, $p < 0.001$).

This is consistent with our initial hypothesis, and that of the Dunning-Kruger Effect. Students who are performing poorly in the course are much more likely to overpredict their marks than those who perform well. This would likely indicate that the poor performing students are not only failing to grasp the core material, but also failing to accurately understand their own level of achievement on the exam. In other words, they don't know what they don't know.

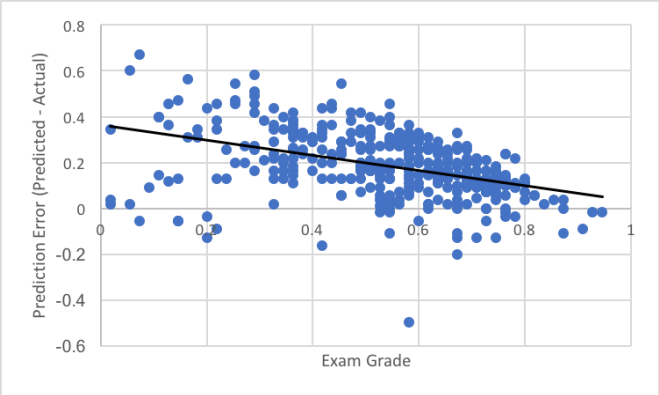

Figure 2: Prediction Error (Predicted Grade - Actual Grade) vs. Actual Grades

One could argue that the correlation between prediction error and exam grade is merely an effect of the grading system. Since students at the low end of the spectrum must necessarily be overpredicting and students at the higher end of the grade scale do not have much room to overpredict. However, similar correlations ($r = -0.47$, $\rho = -0.48$, $p < 0.001$) are found when comparing the magnitude of the error (the absolute value of the predicted error) and the exam grade as seen in Figure 3.

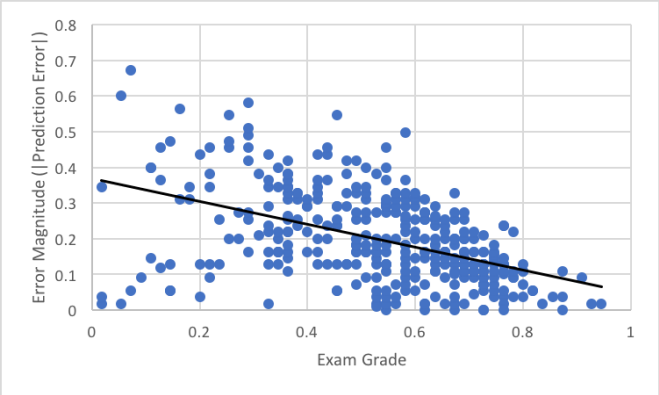

Figure 3: Error Magnitude vs. Actual Grades

This indicates that students who are performing better on the exam (and by proxy have a better understanding of the course material) are less likely to overestimate their marks and are better at assessing their own level of achievement.

4.1 Effect of Prior Programming Experience

Students who had programming experience prior to starting the course performed better in the final examination than those who

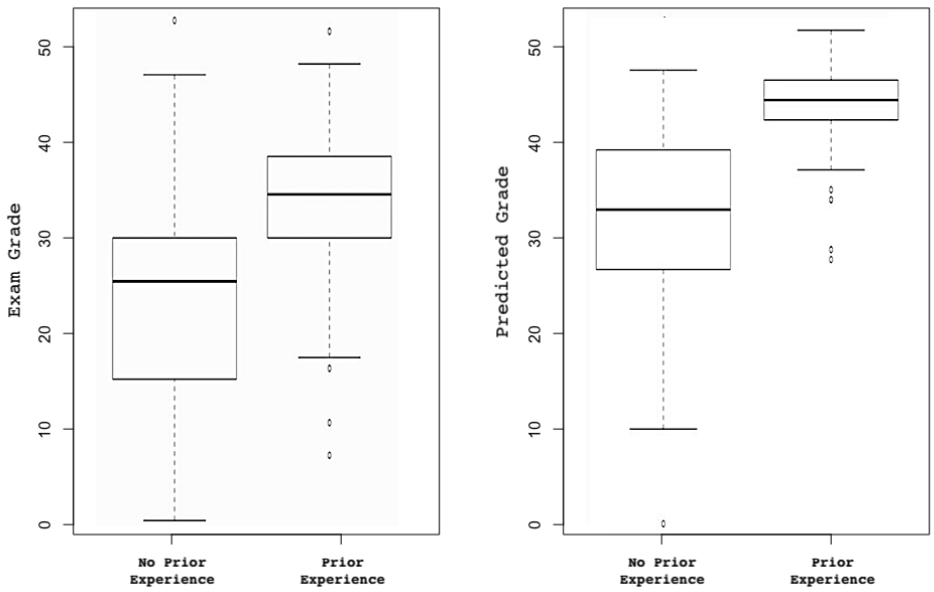

Figure 4: Actual and Predicted Grades by Prior Programming Experience

had no prior programming experience (average of 63.8% vs 48.7%). The data analysis performed to this point indicates that we should therefore assume they would be less likely to overestimate their marks. However, the reverse trend is true. As can be seen in Figure 4, students with prior programming experience were actually more likely to over-predict their marks, with students having no prior experience over-predicting by an average of 16.8%, and those with prior experience over-predicting by 21.8%. However, these results are not statistically significant according to a two-sample t-test ($t = -1.8555$, $p = 0.67$).

While the results are not statistically significant, the fact that the general trend is the opposite of what we would assume given our previous conclusion that better performing students tend to be less likely to over estimate their marks is interesting. The reversal of the trend implies that some other factor is affecting the level of estimation error. This could be due to the fact that students who enter the course with prior programming experience perceive the course as easier, and have higher confidence in their own abilities. This could explain the data for this group showing the reverse trend of the general population. Over confidence in students with prior experience leads to overestimation of marks, resulting in a higher absolute error. Conversely, students who enter the course with no prior experience may perceive the course as more difficult, and be less confident in their abilities. This effect is not large enough to be statistically significant, but it is worth noting that this data was collected at the end of a 12 week intensive CS1 course, so much of the initial effects of prior experience may have worn off over the course of the term.

4.2 Exam Performance Relative to Term Performance

In order to set a mark that we would have expected for a student given their previous mark, we used a simple linear regression to produce a direct mapping between term mark and exam average. This allowed us to have a predicted exam mark for any student given their grades in the previous assignments and tests. We then classified students as either *outperform* (having achieved a higher mark on their exam than their previous marks would predict) or *under perform* (having achieved a lower mark on their exam than their previous marks would predict).

As seen in Figure 5, students who under performed also over-predicted ($t = -3.9261$, $p < 0.001$). Interestingly, the predicted marks were almost identical in the two groups. The *outperform* group had an average predicted mark of 75.2% while the *under perform* group had an average predicted mark of 73.3%. This same calculation was repeated with a number of variations:

- leaving a "buffer" around the expected value, so students who only received a mark 2%, 5% or 10% above/below expectations were classified
- only using previous "exam-like" work (quizzes and term tests) as a predictor
- only using previous "non-exam-like" work (assignments and exercises) as a predictor

The pattern was identical in all cases. The predicted marks remained almost identical among the two groups.

The fact that the predicted marks are no different for overperformers than for underperformers implies that the confidence of

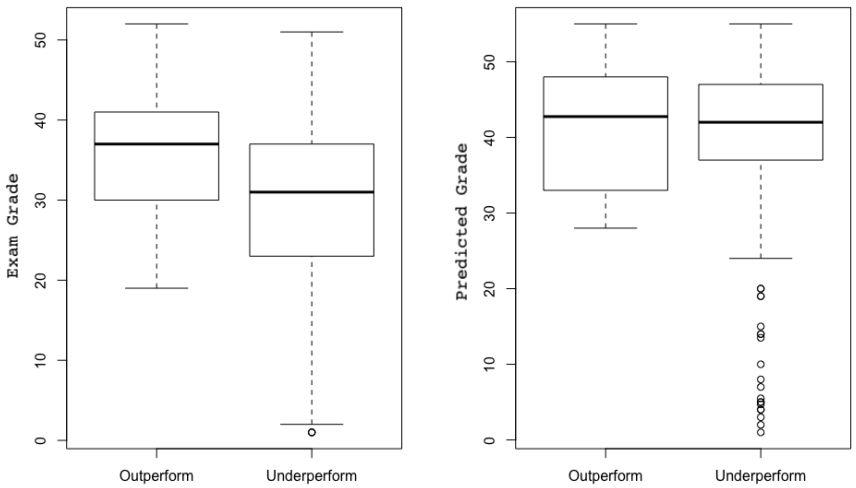

Figure 5: Actual and Predicted Grades by Exam Performance Relative To Expectations

students is strongly anchored to their previous experience in the course. This implies that actual performance during the test is having little effect on confidence. Students who did well in the past will assume they are doing well on the test, and those who performed poorly during the term will assume they are still performing poorly, regardless of how they are actually doing during the test itself.

4.3 Gender

There was no statistically significant difference in the exam marks of the male and female students in the class. Female students averaged 52.1% and male students averaged 53.7%, a difference of 1.6%. However, as seen in Figure 6, female students predicted an average mark of 67.8% while male students' average predicted mark was 73.2%, a difference of 5.4%. A two-sample t-test shows that this difference is statistically significant ($t = 2.442$, $p = 0.015$).

When analyzing the error magnitude of the two groups, female students' predicted scores deviated from their actual scores by an average of 17.8% while males' predicted scores deviated from their actual scores by an average of 20.5%. A two-sample t-test did not find this difference to be statistically significant ($t = 1.756$, $p = 0.079$).

This result provides clear empirical evidence of something that has been known anecdotally in computer science programs (and STEM programs in general) for a long time. Female students are less confident than their male counterparts. The female students in the course performed almost exactly equally to the male students, but they were clearly less confident (or in this case less overconfident) than their male classmates. Furthermore, the fact that the error was statistically significant, while the magnitude of the error was not implies that this was not simply a case of females being somehow better at predicting their own grades, but that the effect was clearly a downward shift in predicted marks.

5 DISCUSSION & FUTURE WORK

The experiment detailed in this paper has produced several interesting results, most notably providing clear evidence of the difference in confidence levels of male and female students. As with all education research there are some threats to the validity of the study, but its simple nature and low cost of implementation should make it a good candidate for replication and validation. That same simplicity also means that this experiment can be adapted in a variety of ways to study new aspects of confidence and to be applied to new domains.

5.1 Threats to Validity

While 542 students is a fairly large size for a CS1 course, it is still a relatively small sample size. Furthermore, the fact that this experiment was only performed on a single cohort of a single class writing a single exam means that all conclusions must be taken with caution. It is possible that some artifact of the class, or the cohort of students, or the test itself may be responsible for the effects seen in this study.

The course started the year with 800 students, which means that 258 students who were initially enrolled in the course either dropped the course or failed to write the final exam. Particularly when assessing aspects such as gender and prior experience that we know have an impact on drop rates, this must be taken into account.

We have made several assumptions in the analysis of this paper. We have assumed that students were motivated to make the most accurate predictions possible on their marks to get within the 10% threshold for a bonus mark. We have assumed that the exams were marked accurately and fairly, and in a manner that was not unexpected to the students writing the test, so that students had a reasonable chance of assuming the marking scheme. We have only

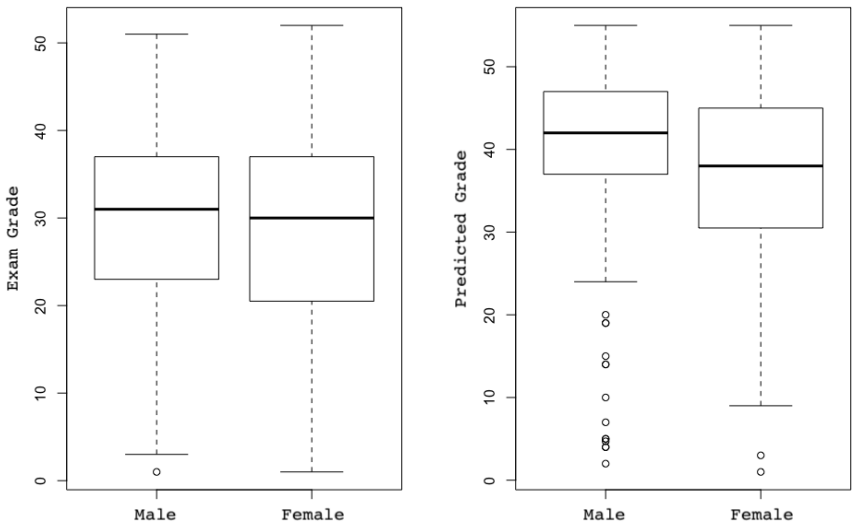

Figure 6: Actual and Predicted Grades by Gender

analyzed the final marks and the predicted final marks, assuming that asking them to predict the mark for each question and then sum those numbers is the same as simply asking them to predict their final mark. This may be biased by things such as primacy or recency effects or the relative difficulty of individual questions.

This work is largely predicated on the assumption that the difference between predicted and actual grades is a good proxy for confidence. While it is reasonable to assume that a confident student will predict a higher grade than an unconfident student, it is possible that a higher prediction may not necessarily imply confidence, and other psychological or pedagogical factors could be affecting the relationship.

5.2 Future Work

This study is easily replicable, and we intend to take advantage of that fact. In future, we intend to replicate this study across multiple tests, not just the final exam, which would provide interesting information on students who drop the course, as well as providing a longitudinal view of students' ability to predict scores as they progress through the course. We also intend to replicate the study across both CS1 and CS2 to extend that longitudinal view.

Many other factors could be affecting confidence, such as the race or immigrant status of the students or other demographic data. Elements of the test itself could impact confidence reporting, such as question ordering, relative difficulty of questions, novelty of questions, or length of the test. All of these are easily varied and tested over time.

This experiment was designed in large part to study the effects of gender on confidence in CS courses, but there is nothing about the methodology that restricts it to any particular course. In future we hope to perform the same experiment on other types of courses as a control.

REFERENCES

[1] Mirella Atherton. 2015. Measuring confidence levels of male and female students in open access enabling courses. *Issues in Educational Research* 25, 2 (2015), 81 – 98.
[2] Tor Busch. 1995. Gender differences in self-efficacy and attitudes toward computers. *Journal of educational computing research* 12, 2 (1995), 147–158.
[3] Nick Cheng and Brian Harrington. 2017. The Code Mangler: Evaluating Coding Ability Without Writing Any Code. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE '17)*. ACM, New York, NY, USA, 123–128.
[4] Sapna Cheryan, Victoria C Plaut, Paul G Davies, and Claude M Steele. 2009. Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology* 97, 6 (2009), 1045.
[5] Hashmia Hamsa, Simi Indiradevi, and Jubilant J. Kizhakkethottam. 2016. Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Procedia Technology* 25 (Jan 2016), 326–332.
[6] Eszter Hargittai and Steven Shafer. 2006. Differences in Actual and Perceived Online Skills: The Role of Gender*. *Social Science Quarterly* 87, 2 (2006), 432–448.
[7] Sabina Kleitman and Lazar Stankov. 2007. Self-confidence and metacognitive processes. *Learning and Individual Differences* 17, 2 (2007), 161 – 173.
[8] Meera Komarraju and Dustin Nadler. 2013. Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences* 25, Supplement C (2013), 67 – 72.
[9] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.
[10] Don A. Moore and Paul J. Healy. 2008. The Trouble with Overconfidence. *Psychological Review* 115, 2 (2008), 502–517.
[11] Monika Sieverding and Sabine C. Koch. 2009. (Self-)Evaluation of computer competence: How gender matters. *Computers & Education* 52, 3 (2009), 696 – 701.
[12] Ellen Van Velsor, Sylvester Taylor, and Jean B. Leslie. 1993. An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management* 32, 2-3 (1993), 249–263.
[13] Jennifer Wang, Hai Hong, Jason Ravitz, and Marielena Ivory. 2015. Gender differences in factors influencing pursuit of computer science and related fields. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 117–122.