A Study of the Relationship Between a CS1 Student's Gender and Performance Versus Gauging Understanding and Study Tactics

Kristin Stephens-Martinez
Duke University
Durham, North Carolina
ksm@cs.duke.edu

ABSTRACT

Metacognitive monitoring is an individual's ability to assess their level of mastery. This skill is integral to learning because students decide what to study based on what they believe they do not understand. Therefore, how well can a CS1 student gauge their mastery? We had students demonstrate their metacognitive monitoring skills by predicting their scores for all three exams of a 15-week CS1 course. We collected data from two course offerings. Moreover, we had students predict both before and after each exam to understand the effect of seeing it and surveyed students on how they studied.

We found that our study's students were reasonably accurate, but low performers were worse than high performers. However, high performers did not improve between their before and after predictions. We did not have sufficient evidence that students improve their predictions over time. Prediction accuracy did not have a gender effect. Finally, we found no difference in study tactics by gender and little difference between high and low performers. Overall, we found only some similarities to related work.

CCS CONCEPTS

• **Social and professional topics → CS1**.

KEYWORDS

cs1; metacognition; metacognitive monitoring; exams; assessment; self-assessment

ACM Reference Format:
Kristin Stephens-Martinez. 2021. A Study of the Relationship Between a CS1 Student's Gender and Performance Versus Gauging Understanding and Study Tactics. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432365

1 INTRODUCTION

Metacognition is an awareness and understanding of thought processes and is a powerful predictor of learning [14, 15]. An aspect of metacognition is *metacognitive monitoring* – an individual's ability to assess the state of their cognitive activity [4]. This metacognitive monitoring skill is crucial because it informs *metacognitive control* – what guides students in how/what to study [5]. Therefore, how well can a CS1 student gauge their mastery? Especially when they have little or no prior programming experience?

We had students demonstrate their metacognitive monitoring skills by predicting their score on all three exams in a 15-week CS1 course. Students predicted before and after the exam.

For both midterms, we also surveyed students on their study tactics. We sought to understand if there is a relationship between study tactics, expected performance, and actual performance. We listed five different class material and provided options on how they used that material to study. These options mapped to (1) retrieval practice, (2) reviewing, and (3) not using the material. We split the options in this manner because retrieval practice is known to benefit students more than merely reviewing the material [10].

With our data, we sought to answer the following questions. (1) How accurate are students across the exams? (2) Is the after prediction better than the before? (3) Does accuracy improve over time? (4) Do groups of students study differently?

We found our students were reasonably accurate, with low performers making worse predictions than high/middle performers. High performers were also a little more accurate on the before prediction than middle performers and much more likely to underpredict. We found no statistical difference in accuracy based on gender. For improving the after prediction, only some of our results had strong evidence. We found little evidence for gender differences and some evidence that students improve their after prediction on earlier exams. For performance, we found no evidence high performers improve their after prediction (likely because they were already accurate), and some evidence low performers do. For improving over time (*i.e.,* between exams), we did not have sufficient evidence that accuracy increases. And for our final question, we found no differences in study tactics between genders nor performance.

Overall, we found some similarities to related work, but not entirely. We believe this work adds to our understanding of the relationship between a student's exam prediction versus gender and performance and especially for a CS1 student's metacognition.

2 RELATED WORK

The related work closest to ours is those that had students predict their exam score. We found one work within CS [9] and the rest in psychology [2, 6, 7, 12, 13]. Within CS, Harrington et al. [9] situated their work in terms of a student's confidence in themselves. They claimed that the difference between the student's actual and predicted exam score was a measurement of the student's confidence. In contrast, the psychology works see this difference as a measurement of the student's metacognitive monitoring skills. They

SIGCSE '21, March 13–20, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8062-1/21/03…$15.00
https://doi.org/10.1145/3408877.3432365

claimed that the worse a student is at predicting their exam score, the poorer their skill at assessing their mastery level.

Our work combines these related works. We use the lens of metacognition and especially metacognitive monitoring. Our interest is within the context of a CS1 class and understanding metacognitive monitoring skills relationship with gender, class performance, seeing the actual exam, and how the student studied.

2.1 Differences between related works

The related work we found is noticeably heterogeneous when a prediction is collected and whether an intervention was part of the study. We collected a before and after prediction for all the course's exams and did not conduct an intervention experiment.

Studies varied on when, how, and for which exam students made predictions. Only some studies asked students to predict both before and after their exam(s) [2, 6, 7]. Others were only a before prediction [12] or an after prediction [9]. And one study collected both a per question level of confidence in being correct and an after prediction for the entire exam [13]. Studies collected predictions as the number of points [2, 6, 9], percent of the score [7, 13], and signed letter grade [12]. Studies either had students predict for every exam [6, 7, 12, 13] or only the final exam [2, 9].

Some studies only reported data [2, 7, 9, 13], while others ran an intervention experiment. Hacker et al. 2008 [6] investigated extrinsic motivation versus reflection. Miller and Geraci [12] investigated incentives versus feedback.

2.2 Accuracy

Overall, performance on the exam correlates with accuracy. The better students perform on the exam, the more accurate they are, and the worse they perform, the less accurate [2, 6, 7, 9, 12, 13]. Moreover, the better a student performs, the more likely they will underpredict, while the worse a student performs, the more they overpredict [2, 7, 9]. The central tendency distance (mean or median) between the students actual grade and prediction ranged from 2%-9% [2], 8%-10% [6, 7, 12], and over 10% [13]. Finally, two studies considered gender and they did not find any difference [2, 9].

2.3 Improving accuracy within an exam

If two predictions are collected, whether a student improves their after prediction depends on performance [2, 6, 7]. High performers are pretty accurate on both predictions. Middle or low performers usually have a less accurate before prediction but improve the after prediction. Very poor performers do not improve their accuracy between the before and after prediction. However, these results are based on the means of the before and after predictions per group of students. In contrast, our results are on the means of the difference between the predictions for each student.

2.4 Improve accuracy between exams

For whether students improve their abilities to predict their score over the course of the class, the results are mixed [6, 7, 12, 13]. Both Hacker et al. [6, 7] papers asked students to predict before and after. In the 2000 paper [7], they found overall the before prediction improved, but not the after prediction. When splitting by performance, low performers had low accuracy for both predictions, while high performers started low for both predictions and improved both over the course of the semester. However, in the 2008 study [6], there was no overall improvement. High performers started and stayed highly accurate, while low performers predicted more poorly and whether they improved their accuracy depended on their treatment group. Miller and Geraci [12] found no improvement in their first experiment. In their second, high performers showed no change in their accuracy, while low performers did show improvement over time. Nietfeld et al. [13] did not see improvement between exams.

Table 1: Statistics on the students, percentages are relative to those that volunteered that information.

	SP19	FA19
# of students	179	211
Male	64 (41.0%)	106 (50.7%)
Female	90 (57.7%)	103 (49.3%)
Prior coding experience	20 (12.8%)	44 (21.1%)
White or Asian	130 (83.3%)	186 (89.9%)

2.5 Groups vs study tactics

Only two studies considered study tactics, but not like we did. Hacker et al. 2000 [7] had no intervention and only asked students how many hours they studied, as opposed to our survey that collected how students studied. They found that the number of hours studied does contribute to a student's accuracy for both before and after predictions. However, this is significant only for the later exams, not the first one. Miller and Geraci's [12] second intervention gave students explicit feedback on how to improve their accuracy. This feedback improved the accuracy of the low performers. However, it did not improve their grade, meaning the feedback improved their predictions, not their mastery of the material.

3 METHOD

3.1 Setting and Participants

We collected data from two offerings of a 15-week introductory computer science (CS1) course in Spring 2019 (SP19) and Fall 2019 (FA19). The course uses Python 3 and is taught at a medium-sized, private, R1 institution. We collected data on all three exams in the course. For studying, students had midterm exams from prior course offerings, but not finals. The same professor taught both course offerings using the same materials with negligible differences. SP19 had 250 students, 179 (71.6%) consented to be part of our study. FA19 had 278 students, 211 (75.9%) consented.

The demographics between SP29 and FA19 are noticeably different, see Table 1. Hence, we include both offerings for a more representative sample. FA19 has a higher proportion of students that are white/Asian, identify as male, or have prior coding experience. We surveyed students about their prior coding experience using the same question as Harrington et al. [9]. This course assumes that students have no prior coding experience. Students with some experience are strongly encouraged to take a different course. If a student reported multiple races, we used the smallest minority status option (*e.g.,* A white and Black student, mapped to Black).

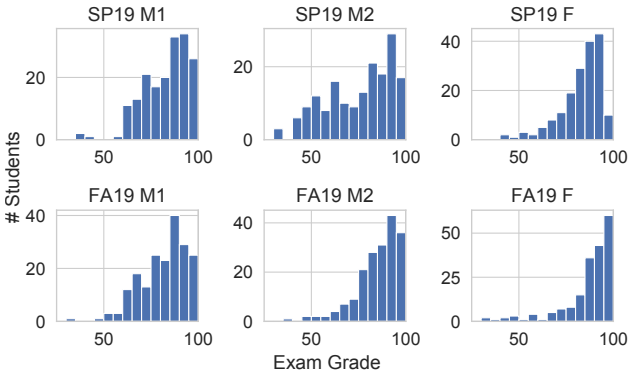

Figure 1: Histograms of exam scores. Note the x-axes are the same, but not the y-axes to help compare proportions.

We collected data on two midterms (M) and one final (F). The midterms consisted of code-tracing, multi-part debugging, short-coding, and long-coding questions. The finals were similar, except the code-tracing questions were replaced with multiple-choice comprehension questions. Midterms were written new each semester, and the finals had much overlap between both offerings. Exams were cumulative with greater emphasis on more recent material. The final had about 50% more material and points than midterms.

We define student performance based on overall exam performance and then split students into three groups. This definition is different from much of related work, which considered only per exam performance. We use this definition because if a student does well on the exams overall, this seems like a reliable indicator of good metacognitive monitoring skills than on a single exam. Using this metric, we can understand the difference between students who are likely to have good metacognitive monitoring skills versus those with poor skills. We calculated performance using only the exams based on the syllabus weighting within the 50% of the students' actual grade (15% M1 and M2, 20% F). We only used exams because they are the course's best summative assessments. We then split the students into three groups: high performers (top third), middle performers (middle third), and low performers (bottom third). Most of our analysis focuses on the high and low performers. We split by thirds to make the differences more apparent.

We incentivized predictions with two extra credit points per exam. One point could boost a midterm by 1.14% to 1.23% and SP19's final by 0.73% and FA19's by 0.79%. These amounts could shift borderline exam scores by a letter, but, we believe, they did not pressure students unduly. Exam grades could not exceed 100%, and none of our analysis includes the extra credit.

When inspecting the exam score distributions, we found SP19 M2 is noticeably different, which we believe affects our results. Figure 1 shows a histogram for each exam's scores. Notice SP19 M2's distribution is much flatter and noticeably flatter than FA19 M2. There are two possible reasons for this flattening. First, it is evidence this exam was harder than the others, which was a surprise to the students and not intended by the instructor. If so, this would skew the predictions for SP19 M2 directly. It would also skew FA19 M2 indirectly because students are strongly encouraged to use old midterms to study, and if they used SP19 M2, they would be improperly calibrated. Alternatively, the flattening is due to the differences in demographics because FA19 has a higher percentage of students with prior coding experience. We will elaborate further where appropriate how this may have affected our results.

3.2 Procedure

Students predicted their score before and after the exam by answering the same multiple-choice question. The options were 5% increments, with the last increment encompassing 6% (*i.e.,* "0%-4%", "5%-9%" ... "90%-94%", "95%-100%"). We collected the before prediction through a survey released 24-hours before the exam and due when the exam started. Students earned one extra credit point by submitting, regardless of prediction accuracy. We took the last submission if a student had multiple. We collected the after prediction on the last page of the exam. We incentivized accuracy by awarding one extra credit point if the student is correct. If the student's score landed between two options, we rounded in their favor (*e.g.,* if they earned 89.5% and predicted "85%-89%" we considered it correct)[1].

We chose to use ranges rather than predicting a percentage to prevent gaming and as an attempt to reduce noise. In a pilot study, we incentivized students in the after prediction by awarding the point if they predicted within 10% of their actual grade, similar to Harrington et al. [9]. However, students gamed this by "predicting" 90% of the possible points regardless of what they would likely get. Moreover, we estimated that guessing within 5% was equivalent to guessing correctly, and any value within that 5% was likely noise.

Our calibration score is similar to Miller and Geraci [12] and represents the closeness of a student's prediction. We converted the students' prediction to only the lower number in the range they predicted (*e.g.,* "95%-100%" becomes 95%). We converted a student's exam percentage score to the lower value of the range the score was in (here on called exam score). If a student's exam score fell between a range, we rounded towards their prediction (*e.g.,* if they scored 84.5% and predicted higher we converted it to 85%, if they predicted below, we converted it to 80%). We defined our calibration score by taking the absolute difference from these two values and subtracting it from 100. Therefore, high calibration scores mean that students were more accurate. When we needed to compare a single calibration score per exam, we used the mid-point between the before and after prediction. If one of the values was missing, we used the available value (usually the after prediction).

After students received each midterm score, we administered an exam wrapper that included questions on study tactics. We used a checkbox grid question with five different course materials versus options for how the student used that material. The materials included: (1) *Problem sets* - short coding questions, (2) *Reading quizzes* - quizzes due before lecture based on the assigned reading material, (3) *Peer instructions* - questions used in lecture, (4) *Assignments* - multi-hour coding assessments, and (5) *Old exams* - prior offering's midterm exams. The options included: (1) "Redid one or more by re-answering the question without looking at the answer first," (2) "Did one or more for the first time," (3) "Reviewed one or more by rereading or looking at the question or answer," and (4) "Did not use any of these to study." We categorized the first two options as

[1]We do not recommend rounding, it complicated logistics. Instead, use ≤ and <.

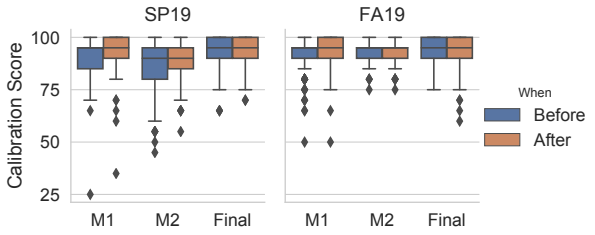

Figure 2: Boxplots for each prediction's calibration score.

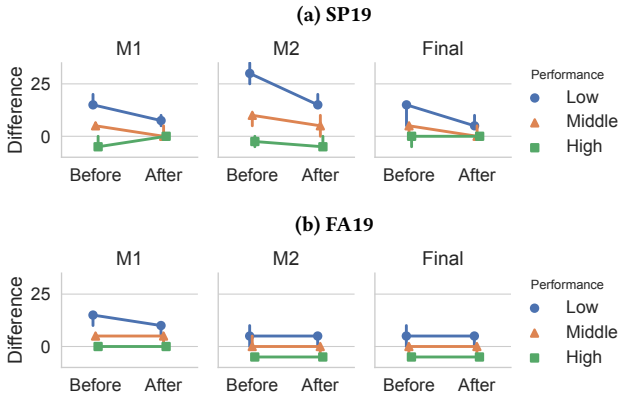

Figure 3: Median difference with confidence intervals between a student's predicted and actual exam score.

retrieval, option 3 as *review*, and option 4 as *did not use*. If a student checked more than one box for a course material, we chose the option with the lowest number (*e.g.,* if they chose option 2 and 3, we considered it option 2). This "rounding" leaned the answers towards retrieval. The exam wrapper also asked students to reflect on their study tactics and make study plans for the next exam.

We analyzed our data in Jupyter Notebooks using Python 3.7, pingouin 0.3.4, pandas 1.0.3, NumPy 1.18.1, and SciPy 1.4.1.

4 RESULTS

4.1 Student accuracy across all exams

Generally, our students are reasonably accurate within 10% of their actual grade. Figure 2 is a boxplot showing the calibration score for each prediction. All exams have a median of 90% or higher, which puts us among the related work's most common calibration range.

When splitting by performance, our results are similar to related work but not completely. Low performers are generally worse at predicting than middle and high performers. However, our high and middle performers are more similar than different, with high performers being slightly better at the before prediction. Figure 3 shows the median[2] difference between the student's prediction and their exam score and a confidence interval. A positive difference means an overprediction. Across all 12 predictions, an equal number of predictions have high performers closer to 0, the same, or farther

[2]We are reporting the median because the distributions were skewed.

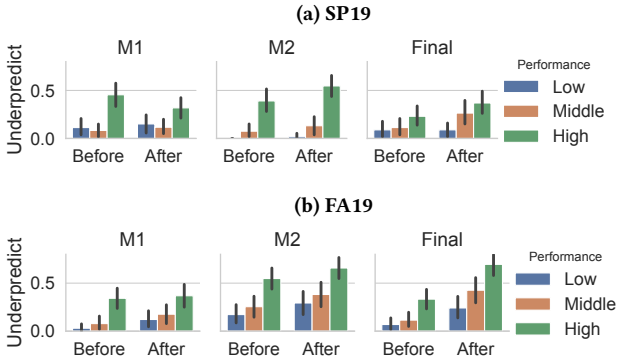

Figure 4: Fraction of students that underpredicted on each exam split by performance.

Table 2: 95% bootstrap confidence intervals to comparing high/low performer's mid-point calibration scores. A significant result (*) is if the interval does not include 0.

	M1	M2	F
SP19	*[5.00, 12.50]	*[10.00, 17.50]	[0.00, 7.50]
FA19	*[2.47, 10.00]	[0.00, 2.50]	[0.00, 5.00]

away than middle performers. However, when considering which predictions are better, it is more of the before predictions (3 out of the 4), which is similar to related work [2, 7]. Also of note FA19's medians stabilized and are closer to 0 than SP19's, which could be because of the previously discussed effects of SP19 M2.

In terms of underpredicting students, we are closer to related work [2, 7, 9], where high performers are more likely to underpredict. Figure 4 shows the fraction of students that underpredicted per group. The high performers always have a higher fraction and often much higher. We can also see the effect of SP19 M2 in the fewer middle/low performers underpredicting in SP19 M2 and larger numbers across all groups in FA19 M2. Moreover, both offering's high performing groups either have negative or 0 medians in Figure 3, meaning at least half underpredicted or made a perfect prediction.

To check for a statistical difference between gender and performance, we used the mid-point calibration scores split by course offering and exam, resulting in 6 distributions. We then split by gender or used high/low performers. However, few of our distributions were normal. Therefore, we could not use t-tests to compare the groups and instead used other statistical techniques.

We found no evidence gender influences how well students can predict their exam score. To find this, we considered only students that self-identified as male or female. We confirmed with a Levene's test that the two groups have equal variance. Then, we used a Mann Whitney U test, a recommended test when the distribution is not normal but does have equal variance. Its H_0 is that the distributions are equal. None of our tests came out statistically significant.

For performance, we have only some evidence that a high performer will usually have a better calibration score than a low performer. Our high and low performing groups did not have equal

variance, so we could not use the Mann Whitney U test. Instead, we used bootstrap [1, Ch13.2], a recommended technique when most common statistical tests will not work. It allows us to estimate the confidence interval of a test statistic. The difference between the group medians is our test statistic. First, from our data set, we randomly sampled with replacement the same number of students as in our current data set. Next, we split the sample by performance, like in the methods section. Then we subtracted the low performers' median from the high performers', so positive values mean high performers have a higher median. We used medians because of skewed distributions. Table 2 shows the 95% confidence intervals for each exam after running this process 1,000 times. Of our six exams, only three intervals did not include 0 (a.k.a. no group difference). This result means that only some of the time are high performers better calibrated than low performers.

4.2 After prediction better than the before?

While Figure 3 implies students as a group improve, we investigate this question at the student level. We found only some results have enough evidence for a strong conclusion. When considering students overall, students seem to improve on earlier exams but not later ones. For gender, there is little evidence of a difference. For performance, we have no evidence that high performing students improve their after prediction. However, we have some evidence that low performers do improve it and, when they do, it is statistically significantly different than high performers.

To check if students improved their calibration score between an exam's predictions, we subtracted the after score from the before. So a positive value means improvement.

We ran different t-tests on the differences depending on what we were investigating. To understand if students improved within an exam, we ran 1-sample t-tests with $H_0 : \mu = 0$, which means no change in calibration. To test if there is a difference when splitting by gender and performance, we ran 2-sample t-tests. Table 3 shows the means for each 1-sample t-test and the p-value for the 2-sample t-tests. A * means the test was statistically significant. Since we are running 7 statistical tests per subset of data (split by exam), we used a Bonferroni correction [8] of 0.05/7 as our significance threshold.

Our evidence that students overall improve on earlier exams is in column one of Table 3, where half the tests were statistically significant. Students improved on SP19 M1, SP19 M2, and FA19 M1, all earlier exams. SP19 M2 was the noticeably different exam, and we thankfully see improvement in the after prediction. In turn, FA19 M2's 1-sample t-test is not statistically significant, which may be due to the previously discussed effects of SP19 M2.

When considering Table 3's gender columns, the most robust result is that none of the 2-sample t-tests were statistically significant. This lack of evidence implies no difference between the genders. From the 1-sample t-test results, we conclude gender does not strongly influence whether a student improves. Only two exams per group had a significant result. These are the same exams that were significant for the overall results. Moreover, almost all means are close to the corresponding exam's overall mean.

The performance columns in Table 3 are noticeably different. In the high performers' column, none of the 1-sample t-tests were significant, so there is no evidence they improved. However, there are three pieces of evidence that low performers do sometimes improve. First, half of the low performers 1-sample t-tests are significant. Second, those same exams are also significant for the 2-sample t-tests. Third, only two of the three were from the set that were significant for the overall results. Therefore, there may be an actual difference and not due to overall trends. Besides, when comparing the high and low performers' means, they are noticeably different.

A potential reason for the weaker results is the decreasing number of students in the data due to fewer before predictions.

4.3 Does accuracy improve over time?

We do not see sufficient evidence that students increase their accuracy over time. We used a one-way repeated measures ANOVA on each course offering with the mid-point calibration score as the dependent variable and the exam as the independent variable. The H_0 was that all exam means are the same. We chose an ANOVA because it is robust against non-normal data.

We do not have sufficient evidence SP19 nor FA19 improved. SP19's means were significantly different, $F(2, 342) = 33.063$, MSE = 43.726, $p < 0.001$, $\eta^2 = 0.162$. However, the exam means are not consistently increasing, 92.57%, 88.53%, and 94.07% respectively. FA19's means were not significantly different, $F(2, 368) = 1.266$, MSE = 26.583, $p = 0.28$, $\eta^2 = 0.007$. Sphericity was violated, so we used the Greenhouse-Geisser corrected p-value. The means also were not consistently increasing, 92.77%, 94.01%, and 93.57%. This lack of an increasing improvement is possibly due to SP19 M2.

4.4 Are there differences in how different groups of students study?

We do not see strong evidence for a difference in how students study when split by gender nor when split by performance. The latter is surprising because we expected high and low performers to fall into their respective groups due to different study tactics.

To measure the difference between groups, we used a two-step process that involved condensing our data into first a five-element student vector and then a frequency vector per group. We first created a *study vector* for each student. Each element represented one of the study materials. We marked an element as a 0 if they did not use that material, 1 for review, and 2 for retrieval. We defined a *group vector* as a $3^5 = 243$ long vector where each element represents the proportion of students with that study vector.

To see if there is a difference between the two groups, we used a permutation test [1, Ch12.1]. Our statistic is the distance between the two group vectors (*i.e.,* $\sqrt{(x_1 - y_1)^2 + \ldots + (x_{243} - y_{243})^2}$). To calculate our confidence interval, we permuted a student's group label and calculated our statistic 1,000 times. Since we are doing this twice on the same data set, we use a Bonferroni correction [8] of $p = 0.05/2$, for a confidence interval of 97.5%. If the statistic for the actual labels falls outside the confidence interval, we know with $p = 0.05/2$ that there is a difference between the groups.

We have little evidence of differences in how the groups studied. Only the test for SP19 M1 split by performance came out significant. Table 4 has the confidence intervals and actual statistic for each midterm exam and groups. We have data only for the midterms because we did not ask students how they studied for the final.

Table 3: Columns 1-5 are the means of the student's before minus the after calibration score. Standard deviation is in (). N is in []. * is rejecting $H_0 : \mu = 0$. Columns 6-7 are the 2-sample t-tests' p-values comparing the groups. * is statistical significance.

	Overall		Male		Female		High		Low		Male vs Female	High vs Low
SP19 M1	2.26 (7.01) *	[157]	1.82 (8.05)	[55]	2.62 (5.91) *	[82]	1.72 (5.13)	[58]	3.40 (9.94)	[53]	0.51	0.27
SP19 M2	4.28 (9.01) *	[139]	4.33 (8.55) *	[52]	4.01 (8.98) *	[71]	-1.11 (4.78)	[54]	11.82 (9.60) *	[44]	0.85	0.00 *
SP19 F	1.02 (6.48)	[83]	1.48 (6.50)	[27]	1.06 (5.55)	[47]	-1.94 (7.15)	[31]	4.14 (5.10) *	[29]	0.77	0.00 *
FA19 M1	1.42 (5.90) *	[165]	1.65 (5.36) *	[82]	1.28 (6.38)	[82]	-0.17 (3.94)	[58]	4.39 (6.60) *	[49]	0.69	0.00 *
FA19 M2	0.76 (4.91)	[144]	1.31 (4.58)	[65]	0.32 (5.12)	[79]	0.74 (3.39)	[54]	1.74 (6.64)	[43]	0.23	0.34
FA19 F	-0.35 (5.42)	[72]	-0.42 (4.47)	[36]	-0.29 (6.32)	[35]	0.60 (3.26)	[25]	-0.80 (7.44)	[25]	0.92	0.40

Table 4: 97.5% confidence intervals and actual statistic using a randomization test when comparing how groups studied. A $*$ means the actual statistic is outside the interval.

	Male vs Female		High vs Low	
SP19 M1	[0.12, 0.22]	0.20	[0.14, 0.24]	0.27$*$
SP19 M2	[0.12, 0.23]	0.18	[0.13, 0.27]	0.21
FA19 M1	[0.11, 0.21]	0.15	[0.13, 0.23]	0.19
FA19 M2	[0.16, 0.31]	0.26	[0.17, 0.33]	0.21

5 DISCUSSION

We started this work with the idea that students with good metacognitive monitoring skills can make better predictions than those with weaker skills. Assuming high performers are better at this skill than low performers, our results align with this idea. Our evidence that low performers sometimes improve their after prediction suggests that the low performers can only directly recognize their mastery level (by seeing the exam) as opposed to their general mastery. Finally, it is gratifying to see no evidence of a gender effect.

For study tactics, the evidence we collected found no difference based on gender and performance. We believe that finding no difference in performance is likely due to the evidence we collected than there is no difference. We collected cognitive strategies [3]. However, study tactics involve more than these. There is the metacognitive control skill of choosing what topics to study and for how long. It is part of self-regulated learning (SRL) that plays a role in learning to program [3, 11]. The low performers use the same cognitive strategies as high performers but are "going through the motions" without necessarily learning as much as the high performers.

This work adds to the evidence that having some metacognitive monitoring skills does not necessarily lead to better exam performance. We believe this is due to poor metacognitive control skills. Further study is needed to understand the relationship between metacognitive skills and performance.

6 THREATS TO VALIDITY

A primary threat is the "space" around a student's exam score to over/under predict, especially for a high score. A student with a high score cannot predict a score greater than 100%. Therefore, there is a ceiling to how much they can overpredict. This ceiling is probably why high performing students are more likely to underpredict. Students with low exam scores have the opposite circumstances. They have more "room" to improve their accuracy. For example, a low performer's before prediction may be inaccurate, but once they see the exam, they can easily improve their after prediction.

While this is important to consider, we believe that it reduces the strength of the evidence about the high performers more than the low performers. So for our results about performance, the evidence that low performers are not very accurate still stands. In contrast, the evidence that high performers are reasonably accurate might partially be measuring noise. As for the lack of evidence that high performers improve their after prediction, we may have seen some evidence if high performers had more "room" to predict.

Other threats to validity include our different methodology from related work, incomplete data, and SP19 M2. Whether our methodology differences matter is not clear, but direct comparison is more difficult. For data, we had a high rate of after predictions but many missing before predictions. See Section 3.1 for details on SP19 M2.

Finally, our study tactics data likely suffers from self-reported recall bias. Students may remember being more productive than they were. This bias could be why there was no difference in how high and low performers studied.

7 CONCLUSION

We asked students to predict their exam score before and after the exam to understand their metacognitive monitoring skills. We also surveyed students on their study tactics. We found that students' predictions were within 10% of their actual grade, and low performers are worse at predicting than high and middle performers.

When comparing a student's before versus after prediction, students overall do improve on earlier exams compared to later ones. Implying some of their metacognitive monitoring skills are developing such that knowing the exam content does not help the after prediction. There is little evidence that gender has an effect. There is some evidence that performance does have an effect, where high performers do not improve, and low performers do improve.

However, we found little evidence that students generally improve their accuracy on subsequent exams. This result may be due to one unusually difficult exam or demographic differences.

Finally, we split students by gender and performance and compared how these groups studied. We found no difference with gender. For performance, we did not have sufficient evidence that there is a difference in how high versus low performers study.

Overall, our results partially align with related work. However, we believe, this work adds to our knowledge of how a student's gender and performance influences their prediction accuracy and study tactics. It also provides a methodology for others to follow.

REFERENCES

[1] Ani Adhikari and John DeNero. 2017. *Computational and Inferential Thinking: The Foundations of Data Science.*

[2] William R. Balch. 1992. Effect of Class Standing on Students' Predictions of Their Final Exam Scores. *Teaching of Psychology* 19, 3 (1992), 136–141. https://doi.org/10.1207/s15328023top1903_1

[3] Susan Bergin, Ronan Reilly, and Desmond Traynor. 2005. Examining the Role of Self-Regulated Learning on Introductory Programming Performance. In *Proceedings of the First International Workshop on Computing Education Research (ICER '05)*. 81–86. https://doi.org/10.1145/1089786.1089794

[4] John Dunlosky and Janet Metcalfe. 2008. *Metacognition.* Sage Publications.

[5] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.

[6] Douglas J Hacker, Linda Bol, and Kamilla Bahbahani. 2008. Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning* 3, 2 (2008), 101–121.

[7] Douglas J Hacker, Linda Bol, Dianne D Horgan, and Ernest A Rakow. 2000. Test prediction and performance in a classroom context. *Journal of Educational Psychology* 92, 1 (2000), 160.

[8] Patricia Haden. 2019. Inferential Statistics. In *The Cambridge Handbook of Computing Education Research*, Sally A. Fincher and Anthony V. Robins (Eds.). Cambridge University Press, Chapter 6, 133–172.

[9] Brian Harrington, Shichong Peng, Xiaomeng Jin, and Minhaz Khan. 2018. Gender, Confidence, and Mark Prediction in CS Examinations. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018)*. 230–235. https://doi.org/10.1145/3197091.3197116

[10] James M Lang. 2016. *Small teaching: Everyday lessons from the science of learning.* John Wiley & Sons.

[11] Dastyni Loksa and Amy J. Ko. 2016. The Role of Self-Regulation in Programming Problem Solving Process and Success. In *Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER '16)*. 83–91. https://doi.org/10.1145/2960310.2960334

[12] Tyler M Miller and Lisa Geraci. 2011. Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning* 6, 3 (2011), 303–314.

[13] John L Nietfeld, Li Cao, and Jason W Osborne. 2005. Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Educational* (2005), 7–28.

[14] Marcel VJ Veenman, Bernadette HAM Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and learning* 1, 1 (2006), 3–14.

[15] Margaret C Wang, Geneva D Haertel, and Herbert J Walberg. 1990. What influences learning? A content analysis of review literature. *The Journal of Educational Research* 84, 1 (1990), 30–43.