Self-Predicted and Actual Performance in an Introductory Programming Course

Paul Denny
Andrew Lluxton-Reilly
John Hamer
Department of Computer Science
The University of Auckland
Private Bag 92019
Auckland, New Zealand

Dana Dahlstrom
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive
La Jolla, CA, USA

Helen Purchase
Department of Computing Science
University of Glasgow
Glasgow, Scotland

ABSTRACT

Students in a large introductory programming course were asked twice to predict their scores on the final exam: once at the beginning of a six-week module, and once at the end. In between, students in only one of the two lecture streams recorded subjective confidence in their answers to individual questions on weekly quizzes. Students’ predictions were moderately correlated with their scores. Students who attended more quizzes had not only higher exam scores, but improved their predictions more than those who attended fewer quizzes. Practice recording confidence on individual quiz questions did not yield significantly more improvement in exam predictions. Several findings from previous work are confirmed, including that women were significantly more underconfident than men.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education—Computer science education, Self-assessment

General Terms
Experimentation, Human Factors, Measurement, Performance

Keywords
Learning, metacognition, confidence, gender

1. INTRODUCTION

The ability to self-assess what one does and does not know is a metacognitive skill that plays a key role in learning through, for example, allocation of study time [15, 2]. Research has shown less-skilled learners not only make more errors, but are less able to distinguish when their answers are and are not correct [11]. Earlier work has suggested a learner’s “feeling of knowing” can be made more accurate and that it would be worthwhile to develop techniques for doing so [16].

Investigators in metacognition have distinguished micropredictions, in which subjects judge how likely they are to be correct on each question, from macropredictions of overall performance [20]. The main question motivating the work we report here is whether, in the context of an introductory programming course, students who regularly recorded micropredictions during quizzes would improve their ability to make macropredictions of their final-exam performance or to answer the quiz or exam questions correctly.

Ancillary to the main thrust, we also investigated other basic questions: how informative and well-calibrated students’ micropredictions were, and how quiz attendance (likely to reflect lecture attendance in general) affected exam scores and macropredictions.

Previous work showed perceived self-efficacy at the end of a programming course was better correlated to performance than was self-efficacy at the beginning [22]; we also generally expected predictions to be better at the end. Previous work also suggests women are less confident (and less overconfident) than men in computer-science courses specifically [1, 9], and in general, particularly for tasks perceived as congruent to male sex roles [17, 13]. Accordingly we expected similar findings regarding gender.

Our research questions can be summarised as follows:

1. Quiz micropredictions: When students are more confident on an item, are they more likely to be correct? How informative and well calibrated are micropredictions, and does this vary by gender or quartile?

2. Exam macropredictions: How did students’ late exam predictions differ from their early predictions, and how well did each predict actual exam scores?

3. Attendance and performance: What is the relationship between the number of quizzes attempted and exam scores? Between quiz scores and exam scores?

4. Effects of micropredictions: Does recording item-specific confidence estimates on weekly quiz questions improve performance, or macropredictions of exam performance?

2. METHOD

The Engineering Computation and Software Development course (ENGGEN 131) is compulsory for all first-year engineering students at the University of Auckland. ENGGEN 131 consists of a
6-week MATLAB-programming module followed by a 6-week C-programming module. Students attend weekly lab sessions, work on individual projects, and sit a test (worth 5%) at the end of the MATLAB module, for which they receive marks before the start of the C module. Lectures are three days per week in two separate, nominally interchangeable streams (at 8 a.m. and 10 a.m.). Lecture attendance is not compulsory, and students may attend either stream, although most gravitate to one or the other.

We collected data during the C module in 2009. During the module, at the beginning of each Friday lecture, quizzes were held consisting of 5 multiple-choice questions on concepts covered that week. Questions were projected overhead so the time for each question was the same for all students. Students’ responses were scanned immediately after the lecture, and the same day, each student received e-mail showing each quiz question and whether their answer was correct. Correct answers were not published; students who had answered incorrectly were encouraged to attempt again to determine correct answers for themselves.

In addition to answering quiz questions, students at the 10 a.m. lectures (but not at 8 a.m.) recorded their confidence in each answer using the following scale:

(A) I’m very (90–100%) confident in my answer
(B) I’m moderately (70–90%) confident in my answer
(C) I’m somewhat (50–70%) confident in my answer
(D) I’m slightly (30–50%) confident in my answer
(E) I’m hardly (<30%) confident in my answer

Twice during the module, students in both streams were asked to predict their final-exam scores, as a percentage of questions to be answered correctly. The “early predictions” were made before the first question of the first quiz, after just two lectures of the C module; the “late predictions” were made after the last question of the last quiz, after all course material was covered and all quiz questions attempted.

The final exam of the C module consisted of 30 multiple-choice questions. Scores on the test at the end of the MATLAB module ranged from 9 to 20.

3. RESULTS

Of 625 students enrolled, 622 sat the MATLAB test; 462 gave early macropredictions; 293 gave late macropredictions; and 613 sat the final exam at the end of the C module. Duplicate submissions and spurious macropredictions (below chance) were removed, leaving 260 students—176 men and 84 women—for which we have all 4 of these measurements.

Quiz results are shown in Table 1. Of the 559 students who submitted any quiz responses, 141 did so only in the 8 a.m. stream; 309 only in the 10 a.m. stream; and 109 at least once in each. A limitation of our data must be noted: as lecture and quiz attendance was not compulsory, and students were free to attend either stream, stream attendance may reflect an unknown self-selection bias from which the effects of making macropredictions cannot be disentangled. Certain results, as noted below, must be viewed in light of this.

In several of analyses we have organised students by quartile based on their performance in the MATLAB test conducted in the module prior to our study, providing a convenient external measure of initial ability, albeit with low resolution. Table 2 shows how students have been organised into quartiles based on the distribution of MATLAB-test marks.

3.1 Quiz micropredictions

As shown in Table 1, in the 10 a.m. stream (where micropredictions were made) mean confidence on quizzes exceeded mean correctness in the total (10.4% overconfidence), and on every quiz but the last. This is consistent with a widely reported general tendency toward overconfidence, especially on difficult tasks, and to a lesser extent toward underconfidence on easy tasks [13, 17, 4, 14].

As shown in Table 1, in the 10 a.m. stream, where micropredictions were made, mean confidence on quizzes exceeded mean correctness in the total (10.4% overconfidence), and on every quiz but the last. This is consistent with a widely reported general tendency toward overconfidence, especially on difficult tasks, and to a lesser extent toward underconfidence on easy tasks [13, 17, 4, 14].

In the 10 a.m. stream, over all item-responses from men (n = 5,790) mean confidence was 59.4%, mean correctness 47.0%, and correlation r = .279; over all item-responses from women (n = 1,710) mean confidence was 48.3%, mean correctness 44.1%, and correlation r = .221. Consistent with other findings [13, 17], men were more correct (p = .031), more confident (p < .001), and more overconfident (p < .001). The difference in correlation between confidence and correctness is also significant (p = .026).

For each student we calculated the mean confidence for all correct responses and the mean confidence for all incorrect responses. Figure 1 shows the mean confidence by quartile for all correctly and incorrectly answered quiz questions. In each quartile, students are more confident in correct answers than in incorrect ones. However, this difference is smallest for students in the lowest quartile, consistent with the unskilled-unaware hypothesis that less-skilled students are less able to determine when they are right and not correct [11]. Also, generally, the better a student performed, the more confident they were both when they were correct and when they weren’t, consistent with the noise-plus-bias account [4].

3.2 Exam macropredictions

For the 260 students yielding complete data, early and late macropredictions are strongly correlated to each other (r = .622, p < .001). Exam scores are correlated to late predictions (r = .343, p < .001) more strongly than to early predictions (r = .269, p < .001), but the difference is not significant (p = .353). The late predictions do account for some variation not explained by the early
predictions: the partial correlation between the late prediction and the exam scores, given the early predictions, is small but significant ($\hat{\rho} = 0.233, p < 0.001$). By contrast, the partial correlation between early predictions and exam scores, given late predictions, is not significant ($\hat{\rho} = 0.076, p = 0.224$). These results are broadly similar to findings relating self-efficacy and performance [22].

As with the lower quiz micropredictions, women ($n = 84$) also gave lower exam macropredictions than men ($n = 176$); in the early prediction, only slightly lower (-2.0%, $r = 0.74$), but in the late prediction, significantly lower (-4.6%, $p < 0.001$). Late predictions were lower for both men (-5.2%, $p < 0.001$) and women (-7.8%, $p \ll 0.001$); but women’s macropredictions dropped significantly more than men’s ($p = 0.044$). Unlike on the quizzes, however, women in fact achieved slightly higher exam scores (-2.6%, $r = 0.163$).

### 3.3 Attendance and performance

Since the following analysis does not rely on micro- or macropredictions, we can use data for all the students who took the exam (at least one quiz ($n = 562$)). Not surprisingly, students who did well on the weekly quizzes tended to do well on the exam. Total mean quiz correctness is strongly correlated with exam score ($r = 0.495, p \ll 0.01$). Also not surprisingly, quiz attendance is moderately correlated to exam performance ($r = 0.36, p \ll 0.001$).

Each effect remains even when controlling for the other: the partial correlation between quiz scores and exam scores, given quiz attendance, is large ($\hat{\rho} = 0.471, p \ll 0.01$) and the partial correlation between quiz attendance and exam scores, given quiz scores, is moderate ($\hat{\rho} = 0.321, p \ll 0.001$).

Figure 2 shows the relationship between the number of quizzes submitted, and the mean exam mark achieved.

A number of studies have been published showing a strong positive relationship between lecture attendance and exam performance in disciplines such as clinical science, economics, civil engineering and physiology [19, 21, 7, 18], but we are not aware of similar studies in introductory computer programming.

### 3.4 Effects of micropredictions

As noted above, the nature of our data limits our ability to reach firm conclusions about the effects of recording micropredictions, since these may be confounded with self-selection effects on attendance in each stream. With that caveat, Table 1 shows quiz scores in the 10 a.m. stream, where students made micropredictions, were higher than in the 8 a.m. stream ($p < 0.002$).

There is a small correlation between quiz submissions in the 10 a.m. stream and exam scores ($r = 0.141, p = 0.023$), but after controlling for total submissions and for quiz scores, the effect entirely disappears. It could be that making micropredictions improved exam scores, but that the effect was wholly mediated by improvement in quiz scores. Our present data do not permit us to distinguish among these possibilities.

We also looked at effects of attending the 10 a.m. stream on the calibration and statistical informativeness of final-exam macropredictions. Calibration is how close a prediction is to the actual score; miscalibration is the magnitude of the difference between the two.

Of the 260 students for whom we have complete data, students who submitted all 6 quizzes in the 10 a.m. stream ($n = 75$) had a mean miscalibration of 12.3% on the early prediction and 14.2% on the late prediction, while students who submitted no quizzes in the 10 a.m. stream ($n = 60$) had a mean miscalibration of 12.9% on the early prediction and 16.6% on the late prediction. The difference between miscalibrations on the late prediction is not significant ($p = 0.231$).

To measure the effect of making micropredictions on improvement in macropredictions, we first computed two linear regressions: $L_1$, modeling exam scores based on early predictions, and $L_2$, modeling exam scores based on late predictions. The residuals of those regressions are $R_1$ and $R_2$, and the unexplained variability in the exam scores is the squared residuals, $R_1^2$ and $R_2^2$, respectively. The improvement in prediction, then, is $R_1^2 - R_2^2$.

There is a small but significant correlation between total number of quizzes attended and prediction improvement ($r = 0.14, p < 0.05$) but the number of quizzes taken in the 10 a.m. stream (where the confidence estimates were being made) is not significantly correlated to prediction improvement ($r = 0.07, p = 0.3$). So while it does
appear that practice answering questions and receiving the associated feedback, and perhaps simply attending lectures, improved students’ abilities to predict their exam performance, practice estimating their confidence for each question did not seem to improve that skill significantly more.

4. DISCUSSION

4.1 Self-assessment skills

Effective feedback is a critical aspect of learning that occurs when students reflect on their own performance in the light of new information. This information can be obtained from an outside source, as is the case with marks obtained from a quiz, or from awareness gained through self-reflection. We asked students to estimate confidence on individual questions with the expectation that the act of committing to the confidence of an answer would improve self-awareness of strengths and weaknesses. These self-assessment skills are important for life-long learning [2].

Performing the confidence estimates had no noticeable effect on their ability to estimate what they would get in the final exam. We speculate that students may already be aware of their confidence in an answer whether they explicitly state it or not, or alternatively, the degree of self-reflection required by making a simple judgment of confidence might not be substantial enough to improve self-assessment ability. The nature of the self-assessment involved a very specific judgment on a localised task which may not scale well to the complexity and richness of assessment that students would expect to see in an exam. In any case, we found no evidence to suggest that there are benefits for students who explicitly state their confidence in an answer.

Both the early and late predictions of exam score are moderately correlated with exam performance. Given this moderate correlation, we believe that simply asking students to estimate the grade they will get in the final exam would be a low-cost, easy to administer way to identify students who are “at risk.” Such students could be targeted for additional help at an early stage in the course.

In this study, students who made the early prediction had already completed the MATLAB section of the course. However, we note that the MATLAB section of the course was taught by different teaching staff to the second section of the course, using a different language (MATLAB vs C). Although the experience of students in the first part of the course is likely to have impacted on their early prediction, we believe that this situation is analogous to students who experience a course such as CS1 and continue to a related course such as CS2. We speculate that student predictions of success in an early part of a CS2 course may be grounded in their experience from CS1, but may still show moderate correlation with exam performance in their current course.

The feedback provided through the weekly quizzes in the form of telling the students which questions they got correct during the course did appear to have minor, but positive impact on the students’ self-knowledge, in terms of their ability to accurately predict their exam score. We believe that although students have some ability to predict their own performance, the regular use of quizzes as a learning tool may improve the ability of students to make accurate self-assessments.

4.2 Confidence and gender

In this study, students’ estimates of confidence are related very specifically to assessments of specific course content. However, the difference in confidence ratings with respect to gender are consistent with the more holistic ratings of confidence typically reported in the Computer Science Education literature [1, 9]. Not only are males more confident than females in the general sense, they exhibit more confidence than females in specific MCQ questions. One implication of this result is that MCQs that use confidence-based marking schemes [6] or negative marking schemes that penalise guessing may result in differences between the scores obtained by men and women that are not directly attributable to their difference in knowledge/ability/understanding.

4.3 Exam performance

Of the different factors investigated, quiz correctness was found to have the strongest correlation with exam score, even when controlled for attendance. It is, perhaps, not surprising that student who scored highly on the quizzes also scored highly on the exams.

More interestingly, we found that quiz attendance was positively correlated to exam scores, even when controlled for quiz correctness. This suggests that participation in learning activities conducted during lectures is a factor in exam performance, regardless of the correctness score. In other words, both quiz correctness and quiz attendance predict exam performance.

An important consideration here is that participation in the quizzes was voluntary. Assigning a portion of the final grade for participation in learning activities can have adverse effects. Published work suggests extrinsic rewards are one of the least effective forms of feedback [8] and extrinsic rewards may undermine intrinsic motivation for learning [5].

The positive correlation between the voluntary quiz attendance and exam score (when controlled for correctness) suggests that improving participation in classroom activities may yield improvements in learning for students of all abilities. We should be looking to implement pedagogies that involve students in voluntarily participating in their own learning, without the imposition of formal assessment.

5. CONCLUSIONS AND FUTURE WORK

We set out to investigate the effects of making micropredictions during weekly quizzes, particularly on macropredictions of final exam score. Although the results do not suggest making micropredictions improved the macropredictions or the exam scores, quiz attendance is positively related to improvement in the macropredictions, and is positively related to exam scores, even when controlling for quiz scores.

Students’ macropredictions at both the beginning and the end of the module were moderately correlated to their exam scores, the later predictions being only somewhat better. Though quiz attendance did mediate improvement in the macropredictions, both this effect, and the improvement itself, were surprisingly small.

That the early macropredictions are nearly as good as the late ones suggests students at the outset already had a reasonably good sense of their likely performance in the C module, probably in part from their experience in the MATLAB module just preceding. In a whole course, one might expect macropredictions to be less good at the beginning, and to improve more by the end. Future work could investigate this hypothesis.

As mentioned in Section 3, the structure of the course and our design for collecting data did not permit a controlled study of the effects of practice making micropredictions. A design with a clear, ideally randomized, intervention model may lead to conclusions we are unable to draw from our data.

Another suggestion for future studies of this kind is to ask students to predict not only their own individual exam scores, but also the class average (or median) exam score. This would facilitate a clearer separation of information about students’ intrinsic self-assessment versus their judgments of extrinsic factors such as the
difficulty of the material. It would also permit analysis of, and controlling for, the well-known “better-than-average” (BTA) effect, in which self-appraisals are more favorable than appraisals of others [3]. Future work could also investigate what factors affect student confidence at the beginning of an introductory computer-programming course, and as it progresses.

Although our results do not suggest practice making micropredictions in itself improves metacognitive ability, the impetus remains to develop techniques that do. Although prior work has suggested that improving skill itself increases metacognitive competence [11], others have called this claim into question and proposed other explanations, generally predicated on task difficulty [12, 10, 4, 14]. Further work could investigate which of the explanations proposed best accounts for results in computer-programming courses specifically, and could evaluate other possible ways to increase the accuracy of self-assessment, perhaps independent of skill.

In the meantime, our results suggest asking students to predict their own final exam scores may be a simple way to identify potentially at-risk students, even at the beginning of a course before students have been exposed to enough material that their learning could be assessed directly.

6. REFERENCES