Peer Instruction in Computing: The Value of Instructor Intervention

Daniel Zingaro[a], Leo Porter[b]

[a]Department of Curriculum, Teaching and Learning, Ontario Institute for Studies in Education, University of Toronto
[b]Skidmore College, Saratoga Springs, NY

Abstract

Research has demonstrated that Peer Instruction (PI) is an attractive pedagogical practice in computer science classes. PI has been shown to improve final exam performance over standard lecture, reduce failure rates, contribute to increased retention, and be widely valued by students. In addition, a recent study using isomorphic (same-concept) questions found that students are learning during peer discussion and not merely copying from neighbors. Though this prior work is useful for evaluating peer discussion, it does not capture learning that takes place after peer discussion when the instructor further expands on the concept through a whole-class discussion. In the present work, isomorphic questions were used to determine the value of a PI question from start to finish: solo vote, group discussion, group vote, and instructor-led classwide discussion. The analysis revealed that the value of the instructor-led classwide discussion was evident in increased student performance over peer-discussion alone (raw gains of 22% compared to 14%). Moreover, the instructor-led discussion was highly valuable for all groups of students (weak, average, and strong) and was of particular value for weak students. Importantly, the largest gains were associated with more challenging PI questions, further suggesting that instructor expertise was valuable when students struggled.

Keywords:
Peer instruction, Computer Science 1, instructor intervention, clickers, active learning

Email addresses: `daniel.zingaro@utoronto.ca` (Daniel Zingaro), `leo.porter@skidmore.edu` (Leo Porter)

1. Introduction

Peer Instruction (PI) is a pedagogical technique developed in physics that has since been used with considerable success in computing. At the core of this pedagogy is the ConcepTest (Crouch, Watkins, Fagen & Mazur, 2007): a multiple-choice question answered by students typically using clickers. Each ConcepTest sets off a well-defined pedagogical protocol: students first answer the question individually (solo vote), then discuss the same question for several minutes with their neighbors, and finally re-vote on the question in light of the group discussion (group vote). Following the group vote, the instructor facilitates a classwide discussion and explanation of the ConcepTest, and can adjust the remainder of the class to target student difficulties.

In physics, it has been repeatedly demonstrated that PI vastly improves student performance on post-course concept inventories (Crouch et al., 2007; Hake, 1998). In CS, there are few concept inventories, and those that exist have not been widely deployed and established (Tew, 2010). Therefore, CS education researchers have used other metrics to measure the effectiveness of PI. PI in computer science has been found to improve final exam performance (Simon, Parris & Spacco, 2013), reduce failure rates (Porter, Lee & Simon, 2013b), and contribute to improved retention (Porter & Simon, 2013).

In addition to overall student outcomes, the value of PI in the classroom can be measured quantitatively by the shift in student correctness between the solo vote and the group vote (Porter, Garcia, Glick, Matusiewicz & Taylor, 2013a; Simon, Kohanfars, Lee, Tamayo & Cutts, 2010; Zingaro, 2010). Such numeric gains from peer discussion suggest, but do not imply, conceptual gains. Is peer discussion helping students conceptually, or are students largely copying from neighbors? Recent work by Porter, Bailey-Lee, Simon & Zingaro (2011) used isomorphic questions to verify that students are indeed learning from the peer discussion.

Since PI is squarely a student-focused pedagogy, it is unsurprising that the nascent PI-CS literature has focused on determining the value of the peer discussion portion of PI. However, measuring gains *solely* from peer discussion may underestimate the total learning conferred through a PI ConcepTest. Each PI cycle concludes with the instructor providing the correct answer and engaging in a classwide discussion meant to provide students with knowledge uniquely held by a subject matter expert: illuminating each response choice, discussing why the concept is important at large, and aiding students in integrating this concept with other concepts in their construction of increasingly expert-like maps of core disciplinary areas. This "instructor intervention" must be captured if we are to truly evaluate student learning from PI ConcepTests.

In this paper, we offer the first account in CS education of the additional benefits conferred through instructor intervention. We also offer the first account in the sciences of the effect of question difficulty on learning gains conferred through peer discussion or instructor intervention.

We conducted a controlled experiment on a large introductory computer science (CS1) class in order to compare peer discussion alone versus peer discussion combined with instructor intervention. We find statistically significant differences between these two modes, clearly demonstrating the importance of instructor intervention within the peer-based PI framework. In addition, we conduct analyses on student ability groups and find that instructor intervention is particularly useful for low-performing students.

The contributions of this paper include:

- A first CS study that measures both peer learning and instructor-led learning. We compare these results with a similar study in biology (Smith, Wood, Krauter & Knight, 2011).
- Evidence that instructor-led discussion is valuable for weak, average, and strong students alike.
- Evaluation of question difficulty demonstrating that difficult questions are particularly valuable for student learning.

2. Background and Related Work

While our focus in this paper is the PI pedagogy, we note briefly that the CS research community is currently investigating many active and collaborative forms of teaching and learning. For example, the flipped classroom, pair programming, and lectures supported by visualizations (Lockwood & Esselstein, 2013; McDowell, Werner, Bullock & Fernald, 2006; Kaminski, 2008) have all been advanced as alternatives or complements to traditional lecture-based teaching.

2.1. Peer Instruction

As described previously, each cycle of the in-class portion of "classic PI" involves students answering a question on their own (solo vote), discussing with their neighbors, and voting again (group vote); see panel A in Figure 1. Following the group vote, the instructor leads a classwide discussion related to the core concept and its misconceptions. The instructor may lecture briefly, or ask students to explain why specific distractors were compelling ("Why did you choose A? What misunderstanding might have led you to choose B?").

Core to the implementation and evaluation of PI is the use of clickers: small devices, similar to television remote controls, that enable students to transmit responses to the instructor's base receiver (Blasco-Arcas, Buil, HernáNdez-Ortega & Sese, 2013). Clickers provide students a low-risk, "fun" technology with which to commit to response choices and engage with the material (Simon et al., 2010; Knight & Wood, 2005). While others have argued that learning is equivalent whether clickers or flashcards are used (Lasry, 2008), clickers afford the immediate generation of accurate response graphs that are of value to both the students and teacher. For example, teachers can use the graphs to facilitate discussion, and students can use the graphs for purposes of formative feedback (Moss & Crowley, 2011). Clickers also generate interaction between students and their peers and instructor, leading to active learning and engagement (Blasco-Arcas et al., 2013). Naturally, clickers as a technology must be paired with an effective pedagogy in order for the positive effects of clickers to be realized. One such pedagogy, and the pedagogy used in the present work, is PI.

Discussion-based pedagogies like PI limit the amount of lecture and increase time spent in discussion and problem-solving. Therefore, to best use available class time, teachers often require students to complete pre-lecture reading (Crouch et al., 2007). In addition, some authors argue that a PI mindset should extend to all aspects of a course, including tutorials and labs (Zingaro, 2010). For this reason, CS researchers have begun a more nuanced inquiry into the effects of PI that go beyond solo-to-group gains. See (Zingaro, 2012; Zingaro, Bailey-Lee & Porter, 2013) for reviews of this work.

Much recent literature suggests that PI is a highly effective pedagogy for teaching CS courses; for example, PI reduces failure rates (Porter et al., 2013b), contributes to increased retention (Porter & Simon, 2013), and yields exam-inferred learning gains compared to traditional course offerings (Simon et al., 2013). PI has been used successfully across the CS curriculum, from introductory courses in C and Matlab (Zingaro, 2010; Lee, 2013) to senior-level courses (Porter et al., 2011). However, there are no studies that examine the instructor's impact on learning in a PI course.

Ideally, peer discussion should engage students in deep processing of material, including comparing and contrasting views with peers as they consider each response choice. We believe that such processing sets up a context in which a follow-on lecture would be particularly effective. This belief stems from constructivist learning theory; specifically, studies have shown that students who have actively compared and contrasted material are better primed to learn from instructor explanation (Schwartz & Bransford, 1998). Through peer discussion, students may come to directly understand the relevant concept. However, if they remain confused, then exchanging perspectives and engaging in argumentation with peers remains educationally important. Bjork has used the term "desirable difficulties" (Bjork, 1994) to characterize student struggles that yield payoff at a later time. We suggest that after engaging with difficult questions, students are particularly prepared and motivated to learn from a more coherent discussion led by the instructor.

2.2. Isomorphic Questions

To examine the impact of peer discussion alone versus peer discussion and instructor intervention, we use a method similar to that used in a recent biology study (Smith et al., 2011). In computing, our work aligns most closely with that of Porter et al. (2011). Each of these studies of interest used **isomorphic questions** to assess the extent to which students' PI-based learning was generalizable across contexts. Isomorphic questions are designed to test the same concept, but use different "cover stories" or parameter values (Smith, Wood, Adams, Wieman, Knight, Guild & Su, 2009).

Porter et al. (2011) used isomorphic questions in their study of learning from peer discussion. Specifically, a first question was presented, on which students voted individually (Q1), discussed, and voted again (Q1$_{ad}$, for "after discussion"). Then, a second question (isomorphic to the first) was presented, on which students voted individually (Q2). Students did not have the opportunity to discuss the first question between the completion of Q1$_{ad}$ and the beginning of Q2, and were not shown the answer to the first question until after Q2.

Note that the terms above will be used throughout the present paper. Q1 refers to the initial vote, Q1$_{ad}$ refers to the vote after discussion, and Q2 refers to the individual vote on the second, isomorphic question. The change in correctness from Q1 to Q2 is representative of the amount of learning that occurred during the PI process.

Porter et al. (2011) highlighted that one particularly important group of students is those who answer Q1 incorrectly, Q1_{ad} correctly, and Q2 correctly. These are students that did not understand the first question, learned from their peers, and were able to apply that understanding to the context of Q2. In the upper-level courses in their study, Computer Architecture and Theory of Computation, respectively, 20% and 13% of students demonstrated these gains.

In addition to these absolute gains, Porter et al. report gains within particular groups of students. They define the Potential Learner Group (PLG) as those students who answer Q1 incorrectly and Q1_{ad} correctly. These students appear to learn from the peer discussion, because they correctly answered the same question following peer discussion. That is, they potentially learned from peer discussion, and Q2 can be used to determine the extent to which this learning was generalizable (rather than, say, a result of uncritically copying peers). In Computer Architecture and Theory of Computation, respectively, 76% and 62% of the PLG correctly answered Q2. The clear majority of learning was generalizable to the isomorphic question.

In a recent study in biology (Smith et al., 2011), the authors varied this protocol slightly in order to compare gains from individual parts of the PI process: peer discussion, instructor, and the combination of peer discussion and instructor. This study was conducted in a genetics course for majors and used three modes of isomorphic question administration:

- Peer: students answered Q1 individually, discussed the question, and answered Q_1ad; the correct answer was shown, and students then answered Q2. (Besides the positioning of the display of Q1's correct answer, this is otherwise identical to the mode used by Porter et al. (2011).)

- Instructor: students answered Q1 individually, the instructor led a classwide discussion, and students answered Q2.

- Combination: students answered Q1 individually, discussed the question, and answered Q1_{ad}; the instructor then led a classwide discussion, and students answered Q2.

Smith et al. used Normalized Change (NC) to compare gains made in the three modes. NC measures the amount of learning as a fraction of available learning. (For example, imagine that 60% of students answer correctly in the solo vote and 80% of students answer correctly in the group vote. This is a 20% gain, but NC divides this 20% by the possible amount of learning following the solo vote. That is, $20/(100-60)\%$ is 50%, so this question has a 50% NC.) These authors found that the combination mode was more effective in terms of NC than the peer and instructor modes. Furthermore, this finding held across ability groups: the combined mode was best for weak, average, and strong students.

3. Hypotheses

From the prior work on biology (Smith et al., 2011), we expect:

- CS students will learn more from the combination of peer discussion and instructor intervention than from peer discussion alone.

Moreover, from our experience with PI, we have observed that the instructor performs critical interventions on difficult questions. On such questions, students struggle on the initial vote and often continue to struggle during peer discussion. As argued earlier, we see peer discussion as setting the context within which students can learn a great deal from follow-up lecture. This may be further heightened when students have just finished grappling with difficult questions. Our second hypothesis is therefore:

- Compared to easy questions, the instructor will have a larger contribution to student learning gains on difficult questions.

4. Method

4.1. Study Context

The course for this study was an introductory computer science course (CS1) taught in Fall 2012 at an undergraduate campus of a large Canadian research-intensive university (131 students wrote the final exam). Since 2007, the course has been taught using the Python programming language. Python has recently gained traction as a language for introductory instruction because of its clean syntax and extensive library of functions (e.g. graphics, Internet applications, user interface design, etc.) The course covers traditional CS1 topics in imperative programming, and also spends one week each on sorting, complexity, and object-oriented programming. The course took place over 12 weeks, with 3 50-minute lectures per week. Prior to each lecture, students completed a reading quiz; the instructor read the responses before class to help shape the lecture. The reading quizzes were marked based on completion (not correctness) and were worth 4% of students' final grade; in-class clicker participation accounted for a further 5% of students' grade. The course instructor was a senior education graduate student with significant PI and CS teaching experience, and had taught CS1 using PI several times. The course instructor developed new PI materials for this course offering; we discuss one question later in the paper (Figure 5), and all questions are available at (anon).

4.2. Question Administration

Each lecture contained an average of three PI cycles, with one of the cycles augmented with a follow-up isomorphic question. It was necessary to gain confidence that questions were actually isomorphic and that the difficulty of questions within pairs was comparable (Porter et al., 2011). To this end, the course instructor sent proposed questions to a colleague who is an experienced CS1 PI instructor. This colleague read each pair of questions; when concerns were raised relating to the questions' isomorphic nature or relative difficulty, changes were made by the course instructor. Then, to further eliminate within-pair difficulty variance, the course instructor generated a random number prior to each lecture that determined the order in which to present the isomorphic questions.

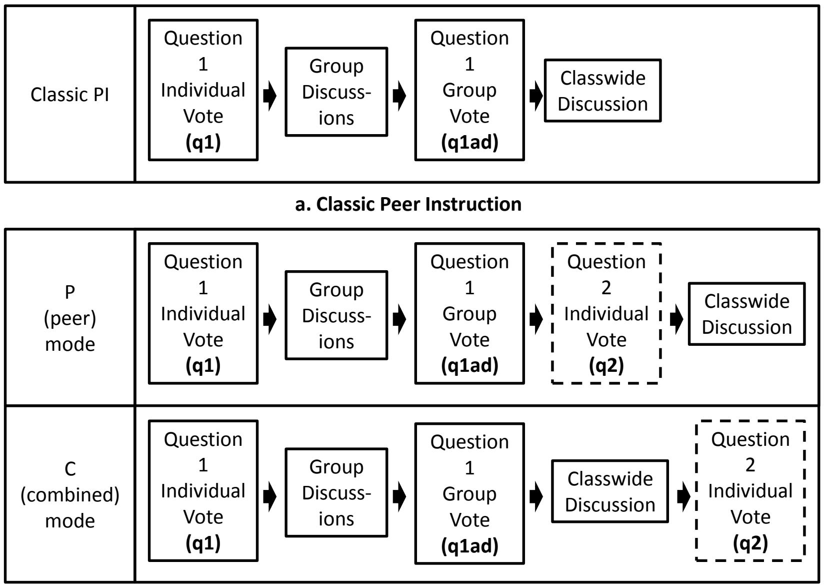
For each isomorphic pair, three student votes were taken; we use the notation introduced above and refer to these votes as Q1, $Q1_{ad}$, and Q2, with the first two votes taken on the first question and the third vote taken on the second. Note that the histogram of responses was never shown between Q1 and $Q1_{ad}$.

Our two question modes are similar to the Peer and Combined modes used by Smith et al. (2011); we did not introduce an Instructor mode as our primary interest here is measuring the instructional value of the full Peer Instruction process including both peer discussion and instructor intervention. These two modes were administered as follows (see panel B in Figure 1):

- **Peer**: students were shown Q1, individually answered, engaged in peer discussion, and answered Q1$_{ad}$. Then, students were immediately presented with Q2 (the second question of the isomorphic pair) and voted individually. Note that between Q1$_{ad}$ and Q2, there was no instructor intervention at all, and that the correct answer to the first question was not displayed. This mode exactly matches the protocol used in Porter et al. (2011).

- **Combined**: the treatment of Q1 and Q1$_{ad}$ in this mode is the same as in the Peer mode. Specifically, students individually answered Q1, engaged in peer discussion, and answered Q1$_{ad}$. Then, the instructor displayed the histogram for Q1$_{ad}$ and proceeded to explain the question, its distractors, and its correct answer. The instructor was careful not to "give away" Q2, even though solid teaching would likely argue for this very comparison (we return to this in the discussion section). Following instructor intervention, Q2 was shown on which students voted individually.

When responding to Q1 and Q2, students worked by themselves, with no help from their peers in the discussion or the chosen response. Therefore, we use Q1 and Q2 as measures of what students individually know, independent of the influence of their group.


Classic PI
Question 1 Individual Vote (q1)
Group Discuss-ions
Question 1 Group Vote (q1ad)
Classwide Discussion
P (peer) mode
Question 1 Individual Vote (q1)
Group Discuss-ions
Question 1 Group Vote (q1ad)
Question 2 Individual Vote (q2)
Classwide Discussion
C (combined) mode
Question 1 Individual Vote (q1)
Group Discuss-ions
Question 1 Group Vote (q1ad)
Classwide Discussion
Question 2 Individual Vote (q2)

a. Classic Peer Instruction

b. Modified PI Formats for this Experiment

Figure 1: The P (Peer) and C (Combined) administration modes used in the study.

At the beginning of the semester, the course instructor generated a stream of random numbers that was to be used each lecture to determine the mode (Peer or Combined) for the isomorphic pair. The instructor checked the random number just prior to class (after slides had been finalized) to minimize mode bias.

4.3. Data Analysis

As noted by several studies (Simon et al., 2010; Zingaro, 2010; Crouch et al., 2007), PI questions that are too easy have limited learning potential for students. There is no standard cutoff for "too easy", though Q1 correctness above 70% (Crouch et al., 2007) and above 80% (Smith et al., 2011) have been proposed. For purposes of comparison with the Smith et al. (2011) study, we chose to drop questions where the Q1 correctness was 80% or above. (The rationale here is that we seek to develop challenging questions for our students; when 80% of students answer correctly prior to peer discussion, we have produced a poor question.) Two Peer questions and four Combined questions were dropped according to this criterion, leaving 12 Peer and 12 Combined question pairs as our dataset.

For each remaining question pair, we required that each student vote all three times (Q1, Q1_{ad}, and Q2); we removed partial data resulting from students answering only a subset of the votes. In addition, we completely removed a student's data if they answered two or fewer Peer questions or two or fewer Combined questions. We did this in an effort to obtain more reliable data from students. For example, if a student answered only one Peer question, then both 0% and 100% are unlikely to be reasonable estimates of the students' knowledge. Our final dataset contains data for 127 students.

Using Q1 percentage correct, we compared the difficulty of Peer and Combined questions. We found no significant difference in the difficulty of the questions used in these modes (paired $t(126) = 1.7, p = .09$), suggesting that our Peer-versus-Combined randomization was effective.

In order to assess gains between student ability groups, we require that our questions adequately separate weak from strong students. For each question, we expect strong students to answer correctly most of the time and weak students to answer incorrectly most of the time. We used a measure of question discrimination from Item Response Theory (IRT) that lies between 0 and 1, with higher values indicating greater levels of discrimination. Research argues that questions with discrimination indices above 0.3 adequately separate weak students from strong students (Matlock-Hetzel, 1997). The average discrimination for our Peer questions was 0.34; the average discrimination for our Combined questions was 0.38. As we are above the cutoff of 0.3 in both cases, we are justified in using Q1 correctness as a measure of student ability.

We divided students into three groups based on Q1 correctness (Smith et al., 2011). Weak students are defined as those students answering up to 33% of Q1 correctly, average students answered between 33% and 66% correctly, and strong students answered more than 66% correctly. This resulted in 13 weak, 81 average, and 33 strong students. These group sizes are significantly unbalanced; we use them only for comparison with Smith et al. (2011) and for data visualization. Rather than use this arbitrary split in a statistical analysis, we instead use the percentage of Q1 answered correctly as each student's baseline ability. We used the `nlme` R package (Pinheiro, Bates, DebRoy, kar & R Core Team, 2013) to test a random-intercept linear mixed-effects model of the effect of Peer and Combined questions on Q2 performance, using Q1 performance as a covariate. A mixed-effects model was used because the data are hierarchical, not independent: each student answers both Peer and Combined questions, so questions are grouped within students. We set $p = .05$ for determining statistical significance.

5. Results

5.1. Potential Learners in CS1

Prior work has examined the use of isomorphic questions in upper-division computing courses (Porter et al., 2011). As the present paper is the first such study of isomorphic questions in a CS1, we begin by extending the main findings of Porter et al. (2011) related to Peer questions.

Recall that the Potential Learner Group (PLG) are those students who answer Q1 incorrectly and Q_1ad correctly. We find a raw Q2 correctness of 73% for the potential learners which is comparable to the percentages reported by Porter et al. (76% and 62%). That is, over three quarters of potential learners learned from peer discussion.

Next, consider those students who answer both Q1 and $Q1_{ad}$ correctly. This is our control group: we would expect these students to answer Q2 correctly as well. However, they may not, and it is therefore advocated by Porter et al. (2011) that we weight the performance of the PLG based on the performance of this control group. When we do this, we find that the PLG is 92% as likely as the control group to correctly answer Q2. This result is similar to those reported by Porter et al. (2011) (85% and 89%). This is strong evidence that the potential learners are learning from the peer discussion; they are almost indistinguishable after the PI process from (i.e. 92% similar to) those students who understood the concept before any discussion.

In this section, we have demonstrated that CS1 students learn from peer discussion. The similarity of our measures to those of Porter et al. (2011) suggests that the conclusions on two upper-level computer science courses (Computer Architecture and Theory of Computation) may be generalizable to lower-division CS1 courses. We now move to an analysis of the PI process as a whole.

5.2. Comparison of Peer and Combined Modes

In this section, we investigate the first of our two hypotheses. Do students learn more from the combination of peer discussion and instructor intervention compared to only peer discussion?

Figure 2 shows the gain from Q1 to Q2 for each of the three student ability groups and all students. We intentionally provide raw values for these results to allow for more meaningful comparisons. For the error-bars: the Peer

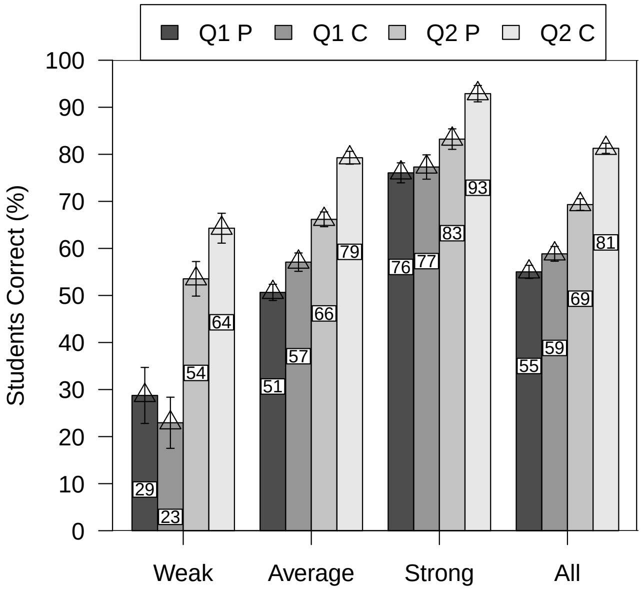


Figure 2: Student performance on P (Peer) and C (Combined) questions. From left to right, the groups of bars represent weak, average, strong, and all students. Error bars show the within-group standard errors.

and Combined scores are repeated measures on the same students, so traditional between-group error bars are not appropriate. Therefore, we have used within-group error bars (Cousineau, 2005; Morey, 2008) to eliminate between-participant differences. Error bars roughly represent the 68% confidence interval; visually, significant differences between two conditions are represented by error bars that are separated by the average size of the two error bars in question (Cumming & Finch, 2005).

This figure first provides the Q1 data for Peer and Combined modes which demonstrates that Q1 scores were not significantly different. This can be seen by comparing “Q1 P” and “Q1 C” in the figure for each pair (Weak, Average, Strong, All).

This figure also provides the improvement between Peer and Combined. The improvement between Q1 and Q2 in each mode represents how much was learned between the start of the initial question and when students are tested on Q2. For example, for Weak students in Peer, 29% (on average) initially respond correctly (Q1). After discussing the question with their peers, 54% respond correctly. This represents a 25% improvement in student correctness. For Weak students in Combined, 41% (64%-23%) of students

Table 1: Linear Mixed-Effect model of Q1, mode, and the interaction, on Q2. $^{**}p < 0.01$

Predictor	β	SE
(Intercept)	0.44^{**}	(0.04)
Q1	0.46^{**}	(0.07)
Admin_Mode	0.19^{**}	(0.05)
Q1:Admin_Mode	-0.15	(0.09)
Log Likelihood	124.97	

improve from the combination of peer discussion and instructor intervention.

Evaluating each student ability group and the course overall, the figure shows the significant differences between gains for Peer and Combined. Weak and average students show the largest gains in Combined compared to Peer; this difference remains highly significant with strong students but is less defined than with the other two subgroups. The rightmost group of bars shows that, across the class, students learn significantly more in Combined compared to Peer, suggesting that instructor intervention confers additional learning beyond peer discussion alone.

To evaluate the statistical relevance of the results from Figure 2, we performed multilevel regression analysis on the model we created to determine which inputs (question mode, student ability level, etc.) statistically impact Q2 performance. The results from this prediction appear in Table 1. The predictor variables are students' Q1 performance, the administration mode (Peer or Combined), and the interaction of Q1 performance and mode; the outcome variable is students' performance on Q2. Each predictor occupies one row in the table, which contains the predictor's name, beta coefficient, statistical significance, and the standard error of the beta coefficient. Each beta coefficient provides the increase in Q2 performance for a one-unit increase in the predictor, keeping all other predictors constant. Beginning with the Q1 predictor, we see that an increase of 1% in Q1 score yields a 0.46% increase in Q2 score, and that this relationship is statistically significant. This means that Q2 performance increases as Q1 performance increases. This is not surprising: we expect that students who are more likely to answer Q1 correctly will also be more likely to answer Q2 correctly.

Next, we move to the mode predictor: the table shows that there is a significant relationship between administration mode and Q2 performance; since the coefficient 0.19 is positive, we know that as mode "increases," so does performance on Q2. In carrying out this analysis, our baseline administration mode was Peer, so we can interpret the mode "increasing" as changing the mode from Peer to Combined. That is, compared to Peer, Combined yields performance increases on Q2. This confirms that adding instructor intervention to peer discussion has a significant effect compared to peer discussion alone.

Finally, we consider the interaction of Q1 and mode. The effect is non-significant ($p = .09$), offering no evidence that the interaction significantly predicts Q2 performance, and that our interpretations of the main effects above are warranted. However, as the interaction does approach significance, it is worth considering what significance of the Q1-mode interaction would have meant. Were it significant, such an interaction might suggest differential effects of the administration mode based on Q1 performance. For example, Combined might help some subgroups of students more than others. Indeed, we see evidence of this tendency in Figure 2, where Combined was helpful for all students (main effect) but particularly helpful for weak and average students (evidence of an interactive effect). As such, the near significance suggests that possible interactions are worthy of future evaluation.

In summary, our statistical model confirms what we observe in Figure 2. The combination of peer discussion and instructor intervention is superior than peer discussion alone, and this finding holds independent of student Q1 scores. These results support our first hypothesis.

5.3. Comparison of Normalized Change between Modes

Normalized change (NC) remains a standard metric in the PI literature for evaluating student learning. As such, in Figure 3, we provide the NC from Q1 to Q2 for each of the student groups and the class overall. Note that we have not displayed error bars in this figure, and that we have not statistically analyzed these NC scores. The reason is that NC is a nonlinear computed quantity, not the raw data itself (Marx & Cummings, 2007). It is therefore inappropriate to treat these values as raw data in a statistical model.

Similar to the results for raw gains, NC shows that students demonstrate substantially larger learning gains in Combined compared to Peer, again supporting our first hypothesis. However, compared to the raw gains in Figure 2, it appears that strong students learn proportionally more than the

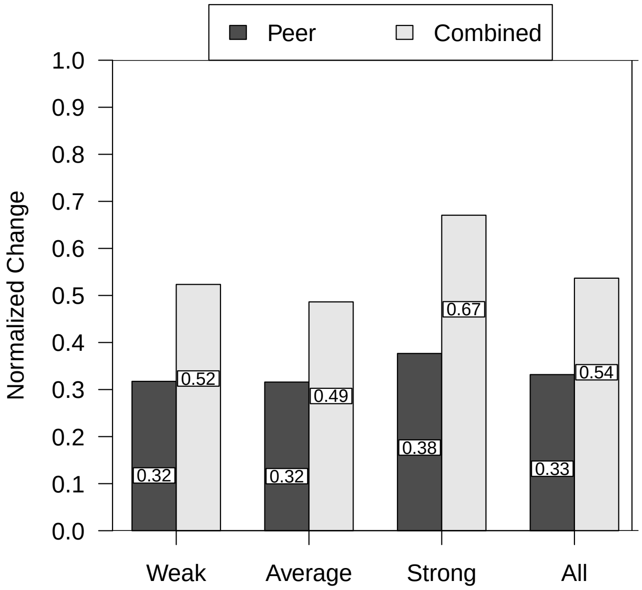


Figure 3: Student normalized change on Peer and Combined questions. From left to right, the groups of bars represent weak, average, strong, and all students.

other groups of students; i.e. weighted by the maximum amount of possible learning, strong students learned the most. We return to the difference between raw gains and NC in the discussion.

5.4. Question Difficulty

So far, we have compared Peer and Combined for all students and for student ability groups, finding clear learning improvements for Combined. We now move to our second hypothesis to further analyze our findings by question difficulty.

To investigate this hypothesis, we change focus from examining groups of students to examining groups of questions. We define difficult questions as those where Q1 correctness is below 50%, and define easy questions as those where Q1 correctness is at least 50%.[1] This yielded 7 easy and 5 difficult Peer questions, and 9 easy and 3 difficult Combined questions. There was

[1] Ideally, we would have preferred to use an easy, medium, difficult split where difficult questions are those whose Q1 correctness is below 35% (Crouch et al., 2007). However, we had only one Combined and two Peer questions that would have been classified as difficult according to this split; such small numbers preclude useful analysis.

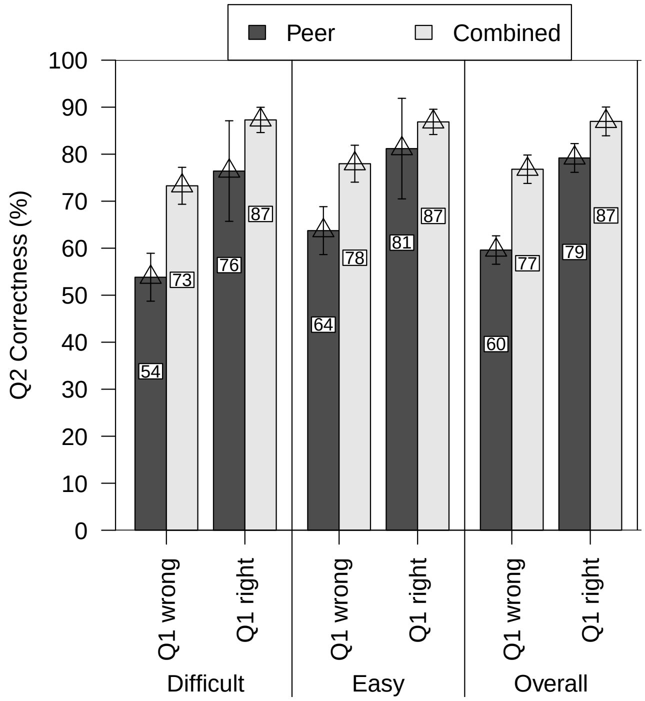


Figure 4: Student performance on Peer and Combined questions. From left to right, the groups of bars represent Q2 performance on difficult, easy, and all questions. Error bars show the within-group standard errors.

no statistical difference between the Q1 scores on easy Peer and Combined questions ($t(13.24) = 0.33, p = .75$), or between the Q1 scores on difficult Peer and Combined questions ($t(4.64) = 0.17, p = .87$). That is, as for questions overall, our randomization was effective in balancing both easy and difficult Peer and Combined questions.

For each of the two question difficulties, we examined performance on Q2 for two subgroups of students: those who answered Q1 incorrectly and those who answered Q1 correctly (see Figure 4). Most striking in this data is the benefit of Combined when students initially answer Q1 incorrectly or when questions are difficult. On difficult questions, students who answer Q1 incorrectly score 54% on Q2 in Peer compared to 73% in Combined. This is a benefit of almost 20% for Combined. Smaller but substantial gains associated with Combined are found on difficult questions when students answer Q1 correctly; in this case, Combined outperforms Peer on Q2 by 11%. Indeed, the only case where Peer and Combined are comparable on Q2 is when

Table 2: Comparison of Q1, Q1$_{ad}$ and Q2 for Peer and Combined.

Mode	Difficulty	Q1 (%)	Q1$_{ad}$ (%)	Q2 (%)
Difficult	Peer	38	60	63
	Combined	36	56	78
Easy	Peer	68	82	76
	Combined	66	84	84

questions are easy and students correctly answer Q1. (In this case, students are over 80% likely to answer Q2 correctly whether or not instructor intervention is provided.) In all other cases, the instructor significantly contributes to learning as measured through Q2 performance.

Surprisingly, peer discussion was also highly useful on difficult questions (see Table 2). In Peer, we see a gain from Q1 to Q2 of 25% for difficult questions and 8% for easy questions. As expected from Figure 4, these values are substantially larger for Combined; nevertheless, progress is made by peer discussion even when questions are difficult. These results provide support for our second hypothesis: compared to easy questions, the instructor has a larger contribution to student learning gains on difficult questions.

6. Discussion

In this study, we have demonstrated the learning gains conferred to students through peer discussion alone (Peer) and through the combination of peer discussion and instructor intervention (Combined). For all students, Combined resulted in significantly larger learning gains than Peer, and a near-significant interaction suggests that Combined may have an even greater effect on weak and average students. In terms of NC, however, we found that strong students proportionally gained the most. How can these results be remedied? It is our opinion that raw gains are a more useful measure in this context, since NC can be inflated by high Q1 scores. For example, if 10 out of 12 strong students answer Q1 correctly, and one of the two incorrect students then proceeds to answer Q2 correctly, the NC for the strong students will be 50%. This, however, does not communicate the effectiveness of the mode of administration, which in this example helped only one strong student. We see raw gains, at least when group comparisons are concerned, as a more pure measure of the effectiveness of each mode.

Continuing with an examination of question difficulty, we found that both administration modes led to learning gains on both easy and difficult questions. On difficult questions, however, we found that Combined offered significant benefits over Peer, especially for those students who incorrectly answered Q1. When students struggle with Q1, instructor intervention is argued to be extremely valuable.

These findings, and the design of the study itself, lead to several areas of discussion: the quality of instructor explanation, the usefulness of difficult questions, and the use of Q1 to $\mathrm{Q1}_{ad}$ as a proxy for "total learning" in a PI cycle.

6.1. Quality of Instructor Explanation

In Combined, the instructor discussed Q1 following the $\mathrm{Q1}_{ad}$ vote. Naturally, the instructor was aware that Q2 would directly follow the explanation, and reflection suggests that this constrained his teaching. In standard lectures, we seek to engage students in the types of cognitions that are valued by CS professionals: comparing-and-contrasting, generating multiple solutions, studying multiple examples (Beatty, Gerace, Leonard & Dufresne, 2006). Frequently, the instructor desired to do this, but often the "what if" or "what would happen when" was exactly represented in the impending Q2. (In fact, several questions were removed from the dataset due to the instructor inadvertently giving away the answer to Q2.)

For example, consider the isomorphic pair in Figure 5. These questions address algorithmic complexity in computer science. That is to say, given a problem size of n, what is the upper bound on the runtime of the algorithm with respect to n. In the question on the left, the nested *for* loops cause the runtime to be n multiplied by two constants (10000 and 50). As constants do not grow with n, this runtime is Linear (n). In the question on the right, we see that the runtime is n multiplied by n multiplied by a constant (50). Again, the constant does not grow with n so the answer is Quadratic (n^2).

Following peer discussion of the first question (the question on the left), it would be natural for the instructor to ask, "what would happen if I replaced one of the numeric constants with n?" However, this is exactly what happens in Q2 (the question on the right). Therefore, rather than explore a concept to its full generality, the instructor often stayed focused on Q1 itself, only expanding the focus following Q2. Once Q2 was complete, the instructor was free to compare Q1 and Q2, add further examples, and otherwise highlight similarities in the isomorphic questions that may not have been apparent.

```
def mystery(n):
  total = 0
  for i in range(n):
    for j in range(10000):
      for k in range(50):
        total += 1
  return total
```

This algorithm is:

- **A. Linear (n)**
- B. Quadratic (n^2)
- C. Cubic (n^3)
- D. Not one of these three

```
def mystery(n):
  total = 0
  for i in range(n):
    for j in range(n):
      for k in range(50):
        total += 1
  return total
```

This algorithm is:

- A. Linear (n)
- **B. Quadratic (n^2)**
- C. Cubic (n^3)
- D. Not one of these three

Figure 5: Two isomorphic questions on complexity. Correct answers in bold.

We therefore argue here that the instructor's explanation following Q1 is not a reflection of the best-quality instruction that could be given, but was constrained by the experimental controls. One possible change to our protocol would involve a third-party generating the Q2 questions, of which the instructor would remain unaware until after the explanation was complete. However, we argue that many questions would have been "spoiled" by virtue of the instructor essentially answering Q2 before it was asked.

What we are suggesting is that gains due to instructor intervention may in fact be larger than what we have shown here. In essence, we are measuring the effect of an instructor's intervention when the instructor was constrained in his ability to engage students in discussion of the concept at large.

6.2. Use of Difficult Questions

The Q1 average over all questions in the study was 57%. The instructor sought to create challenging questions that were within the 35%-70% range advocated in the literature (Crouch et al., 2007). At the same time, the instructor hesitated to intentionally create questions where the Q1 performance would likely be very low, for fear that peer discussion would be fruitless.

However, the few extremely difficult Q1 questions in the study do allow us to make tentative suggestions regarding these very questions. Rather than provide little benefit, peer discussion on difficult questions was extremely valuable, leading to an average 25% increase from Q1 to Q2. Adding instructor explanation increased the gains to 42%. For two reasons, therefore, we advocate for using conceptually-difficult PI questions. First, such questions have the potential to lead to large learning gains, since room for improvement is large. (Easy questions may have large NC, but the number of students benefiting is small.) Second, both peer discussion and instructor explanation combine to produce these large gains, so the use of class time is certainly warranted. Faced with a challenging question, we anecdotally observe that students have longer, more in-depth discussions than those observed on easy questions. After grappling with difficult material, the students may be in a position to learn even more with instructor support. Easy questions, on the other hand, may not engender such discussion and do not maximize the benefits of knowledgeable instructor input.

6.3. NC as Proxy for Total Learning

Much current PI literature measures NC between the individual and group vote (what we have been referring to as Q1 and $Q1_{ad}$), and uses this as a measure of student learning. However, those studies that additionally include Q2 sometimes find that Q2 correctness drops below that of $Q1_{ad}$ (Porter et al., 2011). The goal of those studies is to measure student learning from peer discussion, so the comparison between $Q1_{ad}$ and Q2 is appropriate for capturing that fraction of "learning" that may not be sustained. However, their finding has relevance for the wider PI discussion because it suggests that NC between Q1 and $Q1_{ad}$ may be an overestimate of the amount of student learning. Indeed, we see this drop from $Q1_{ad}$ to Q2 in the Peer mode of the present study as well. Here, we find an NC between Q1 and $Q1_{ad}$ of 0.38, but a smaller NC of 0.33 between Q1 and Q2. $Q1_{ad}$ appears to be an overestimate of learning, likely because it aggregates conceptual gains with gains associated with passive forms of peer influence.

However, the picture is rather different in Combined. There, the NC between Q1 and $Q1_{ad}$ is 0.41, and the NC between Q1 and Q2 is 0.54. Contrary to what was found in Peer, NC between Q1 and Q2 is in fact **larger** than NC between Q1 and $Q1_{ad}$.

This result is of important practical value for PI instructors. Low $Q1_{ad}$ results can be disheartening for an instructor if $Q1_{ad}$ is highly indicative of student understanding at the end of the PI process, especially for conceptually difficult questions. However, our results from Combined show that Q1$_{ad}$ is not representative of student understanding after instructor explanation. Therefore, instructors should feel comfortable asking difficult questions despite low Q1$_{ad}$ results, knowing that their classwide discussion is likely to raise student understanding above the Q1$_{ad}$ threshold. That is, they should view Q1$_{ad}$ as an underestimate of student understanding.

7. Future Work

In this paper, we have conceptualized instructor intervention as the combination of providing students' the correct answer along with conducting a classwide discussion of the ConcepTest. It may be worthwhile to disaggregate these effects to determine the relative importance of "providing an answer" (e.g. showing "A" as correct) and "explaining that answer". It has been shown in biology (Smith et al., 2011) that most of the gain results from instructor explanation (not simply giving the correct answer). In addition, our analysis of difficult questions suggests that it is unlikely that the correct answer would have led to such large learning gains (if students do not understand the question, then they may not be able to abstract from a correct answer to a correct interpretation of why that answer is correct). That said, directly measuring the gains from answers alone would be productive follow-up to the present work.

More broadly, future work should further probe the utility of PI in engendering specific skills among CS students. While prior work has shown PI to be overall "better" than traditional lecture (Simon et al., 2013), and the present work shows that the full PI process is advantageous, a remaining goal is a more fine-grained understanding of skill development. Introductory CS students must learn and demonstrate a variety of skills, including code-tracing, code-reading, code-explaining, and code-writing (Venables, Tan & Lister, 2009). The way in which these specific skills relate to conceptual understanding is an important open question and one that deserves future research attention. As the CS community moves forward in its understanding of PI, it is important to consider the specific ways in which PI helps our students develop conceptual knowledge as well as programming-related skills.

8. Conclusion

The current Peer Instruction literature in CS focuses largely on the gains associated with peer discussion. This is to be expected, since peer discussion is a crucial part of PI (Crouch et al., 2007). However, by privileging the discussion portion, other supporting roles in the PI process may not be as well-understood. Most critically, learning gains after peer discussion have not been measured in computing. As such, we have extended previous studies (Porter et al., 2011; Smith et al., 2011) in order to evaluate instructor intervention in PI. In this study, we demonstrate that instructor intervention is indeed crucial to the success of PI. Students evinced larger gains on isomorphic questions when discussing with peers and receiving answers and explanations from an instructor (81% correct) than when only discussing with peers (69% correct). Moreover, the benefit of the instructor was heightened further when analyzing gains made on difficult questions. For these questions, initially incorrect students improved to 54% correct with peer discussion alone whereas they improved to 73% correct with both peer discussion and instructor intervention. Our findings also suggest that instructors should view $Q1_{ad}$ correctness as an underestimate of overall student learning, as that measure fails to capture the value of the full PI process. In summary, this work demonstrates that instructor intervention nicely complements peer discussion while acknowledging the crucial role to be played by the domain expert.

Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics*, *74*, 31–39.

Bjork, R. A. (1994). Memory and meta-memory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT.

Blasco-Arcas, L., Buil, I., HernáNdez-Ortega, B., & Sese, F. J. (2013). Using clickers in class. the role of interactivity, active collaborative learning and engagement in learning performance. *Computers & Education*, *62*, 102–110.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

Crouch, C. H., Watkins, J., Fagen, A. P., & Mazur, E. (2007). Peer instruction: Engaging students one-on-one, all at once. In E. F. Redish, & P. J. Cooney (Eds.), *Research-Based Reform of University Physics*. American Association of Physics Teachers.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *The American psychologist*, *60*, 170–180.

Hake, R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64–74.

Kaminski, M. (2008). Symbolic computations in modern education of applied sciences and engineering. *Computer Assisted Mechanics and Engineering Sciences*, *15*, 143–163.

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4*, 298–310.

Lasry, N. (2008). Clickers or flashcards: Is there really a difference? *The Physics Teacher*, *46*, 242–244.

Lee, C. B. (2013). Experience report: CS1 in MATLAB for non-majors, with media computation and peer instruction. In *Proceedings of the 44th ACM technical symposium on Computer science education* (pp. 35–40). ACM.

Lockwood, K., & Esselstein, R. (2013). The inverted classroom and the CS curriculum. In *Proceedings of the 44th ACM technical symposium on Computer Science Education* (pp. 113–118). ACM.

Marx, J., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, *75*, 87–91.

Matlock-Hetzel, S. (1997). Basic concepts in item and test analysis. annual meeting of the Southwest Educational Research Association. `http://ericae.net/ft/tamu/Espy.htm` (accessed June 25, 2013).

McDowell, C., Werner, L., Bullock, H. E., & Fernald, J. (2006). Pair programming improves student retention, confidence, and program quality. *Communications of the ACM*, *49*, 90–95.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

Moss, K., & Crowley, M. (2011). Effective learning in science: The use of personal response systems with a wide range of audiences. *Computers & Education*, *56*, 36–43.

Pinheiro, J., Bates, D., DebRoy, S., kar, D. S., & R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-108.

Porter, L., Bailey-Lee, C., Simon, B., & Zingaro, D. (2011). Peer instruction: Do students really learn from peer discussion in computing? In *Proceedings of the Seventh international Workshop on Computing Education Research* (pp. 45–52). ACM.

Porter, L., Garcia, S., Glick, J., Matusiewicz, A., & Taylor, C. (2013a). Peer instruction in computer science at small liberal arts colleges. In *Proceedings of the 18th annual joint conference on Innovation and technology in computer science education*. ACM.

Porter, L., Lee, C. B., & Simon, B. (2013b). Halving fail rates using peer instruction: A study of four computer science courses. In *Proceedings of the 44th ACM technical symposium on Computer science education* (pp. 177–182). ACM.

Porter, L., & Simon, B. (2013). Retaining nearly one-third more majors with a trio of instructional best practices in CS1. In *Proceedings of the 44th ACM technical symposium on Computer science education* (pp. 165–170). ACM.

Schwartz, D., & Bransford, J. (1998). A time for telling. *Cognition and Instruction*, *16*, 475–522.

Simon, B., Kohanfars, M., Lee, J., Tamayo, K., & Cutts, Q. (2010). Experience report: Peer instruction in introductory computing. In *Proceedings of the 41st ACM technical symposium on Computer science education* (pp. 341–345). ACM.

Simon, B., Parris, J., & Spacco, J. (2013). How we teach impacts student learning: peer instruction vs. lecture in CS0. In *Proceedings of the 44th*

ACM technical symposium on Computer science education (pp. 41–46). ACM.

Smith, M., Wood, W., Krauter, K., & Knight, J. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE-Life Sciences Education*, *10*, 55–63.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*, 122–124.

Tew, A. E. (2010). *Assessing Fundamental Introductory Computing Concept Knowledge in a Language Independent Manner*. Ph.D. thesis Georgia Institute of Technology.

Venables, A., Tan, G., & Lister, R. (2009). A closer look at tracing, explaining and code writing skills in the novice programmer. In *Proceedings of the Fifth International Workshop on Computing Education Research* (pp. 117–128).

Zingaro, D. (2010). Experience report: Peer instruction in remedial computer science. In *Proceedings of the 22nd World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 5030–5035). AACE.

Zingaro, D. (2012). Peer instruction in computing: What, why, how? In T. Bastiaens, & G. Marks (Eds.), *Proceedings of Global Conference on Technology, Innovation, Media & Education* (pp. 18–24). AACE.

Zingaro, D., Bailey-Lee, C., & Porter, L. (2013). Peer instruction in computing: the role of reading quizzes. In *Proceedings of the 44th ACM technical symposium on Computer Science Education* (pp. 47–52). ACM.