A Multi-institution Exploration of Peer Instruction in Practice

Cynthia Taylor
Oberlin College
Oberlin, OH, USA
cynthia.taylor@oberlin.edu

Andrew Petersen
University of Toronto Mississauga
Mississauga, ON, Canada
andrew.petersen@utoronto.ca

Jaime Spacco, David P. Bunde
Knox College
Galesburg, IL, USA
{jspacco,dbunde}@knox.edu

Soohyun Nam Liao, Leo Porter
University of California at San Diego
La Jolla, CA, USA
{leporter,snam}@eng.ucsd.edu

ABSTRACT
Peer Instruction (PI) is an active learning pedagogy that has been shown to improve student outcomes in computing, including lower failure rates, higher exam scores, and better retention in the CS major. PI’s key classroom mechanism is the PI question: a formative multiple choice question on which students vote, then discuss, then vote again. While research indicates that PI questions lead to learning gains for students, relatively little is known about the questions themselves and how faculty employ them. Additionally, much of the work has examined PI data collected by researchers operating in a quasi-experimental setting. We examine data collected incidentally by multiple instructors using PI as a pedagogical technique in their classroom. We look at how many questions instructors use in their courses, the difficulty level of the questions, and normalized gain, a metric that looks at increases in student correctness between individual and group votes. We find normalized gain levels similar to those in existing literature, indicating that students are learning, and that most questions, even those developed by instructors new to PI, fall within recommended difficulty levels, indicating instructors can create good PI questions with little training. We also find that instructors add PI questions over the first several iterations of a new PI course, showing that they find PI questions valuable and suggesting that full development of PI materials for a course may take multiple semesters.

CCS CONCEPTS
• Social and professional topics → Computer science education;

KEYWORDS
Peer Instruction, Clicker Questions, CS Education, Active Learning

1 INTRODUCTION
Peer Instruction (PI) is an active learning technique in which students respond to multiple choice questions throughout lecture. It was first shown to be highly effective in promoting student learning in the physics community [2, 7]. Within computer science, it has been shown to improve student learning [16, 21, 27, 28], to lower D/F/Withdraw rates and improve major retention [12, 17], and to be effective for both large and small classrooms [8, 15].

While much work within CS has examined PI’s effectiveness for promoting student outcomes, little has been done to explore how instructors use it. In this work, we look at a collection of over 4,100 posed clicker questions from 26 different offerings of 10 different courses taught by 7 different instructors at 4 institutions, and focus on 3915 questions by 5 instructors. This data was collected in a natural setting, with instructors using PI as a pedagogical technique with little or no research agenda. As such, the set is a foil for PI data collected in quasi-experimental settings.

Contributions of this work include the collection of a large dataset of clicker questions and student response data, the development of a labeling tool for these questions, and a comparison of this data to previously published results. We begin by examining how much time instructors allow for PI questions and how difficult these questions are. While there are differences in how instructors use PI in their classrooms, we find that the majority of our collected questions are between 35% and 70% correct on the initial vote, within the “good question” difficulty guidelines recommended by Crouch et al [2, 3].

Then, we look at normalized learning gain between solo and group questions. Normalized learning gain has been used to measure student learning from PI questions in prior work [15, 20, 22, 24, 27]. We find normalized gain similar to the results published in these works, indicating that students are learning from PI questions [16, 27].

Lastly, we investigate the questions used in the same course over multiple semesters. We find that instructors increase the number of questions used over time when they are either new to PI or developing a new course. We also see that average learning gains are fairly stable for experienced instructors teaching established courses.

2 BACKGROUND AND RELATED WORK
PI was first developed in physics [2]. Instructors using PI present a mini-lecture introducing a topic, and then ask students a multiple choice question on which students vote, then discuss, then vote again. While research indicates that PI questions lead to learning gains for students, relatively little is known about the questions themselves and how faculty employ them. Additionally, much of the work has examined PI data collected by researchers operating in a quasi-experimental setting. We examine data collected incidentally by multiple instructors using PI as a pedagogical technique in their classroom. We look at how many questions instructors use in their courses, the difficulty level of the questions, and normalized gain, a metric that looks at increases in student correctness between individual and group votes. We find normalized gain levels similar to those in existing literature, indicating that students are learning, and that most questions, even those developed by instructors new to PI, fall within recommended difficulty levels, indicating instructors can create good PI questions with little training. We also find that instructors add PI questions over the first several iterations of a new PI course, showing that they find PI questions valuable and suggesting that full development of PI materials for a course may take multiple semesters.
choice question on the covered concepts. Students first vote individually on the question (the solo vote), frequently by using a “clicker”, a small electronic device which sends the instructor an aggregated view of the class votes. The students then discuss the question in small groups. Each group comes to consensus on the answer, and then votes again (the group vote). Next, instructors facilitate a class-wide discussion on the topic and question, usually showing the class the results of their voting, going over the correct and distractor answers, and answering any student questions. This process is repeated for each topic. Notably, the purpose of the PI questions is student learning rather than assessment; they are either ungraded or graded for participation rather than correctness.

While PI can be implemented using low tech devices such as flash cards rather than clickers [6], the use of clickers allows instructors to collect data on student understanding anonymously and with greater accuracy. If most of the class answers the question correctly in the individual vote, the instructor may skip the group discussion and vote. If most students still answer incorrectly after the group discussion, the instructor may give a broader explanation of the topic. Displaying the histogram of answers to students after the group vote lets students know they are not alone in their misconceptions [5].

PI relies on students preparing for the topic before class. This preparation has traditionally taken the form of students watching a video or reading a book chapter, followed by a quiz on the topic [2]. These quizzes may be administered at the beginning of class using clickers or before class using a web interface [11]. Reading quizzes have been found to be helpful to students in CS classes using PI [11, 26]. However, some CS classes do not use reading quizzes at all [19], or may assign exploratory homeworks instead [4].

PI has been studied in computer science education since 2010 [20, 24]. It has been shown to be well-liked by both students and instructors [13–15, 20]. Using isomorphic questions, several studies [16, 27] have shown that students learn from the discussion phase. PI has been shown to improve student efficacy [25], to improve student retention [12], and to lower fail rates [17]. It has also been used to identify struggling students early in a course [9, 18].

### 3 METHODOLOGY

In this section we describe how our data was collected and labeled, as well as establishing our research questions.

#### 3.1 Setting

Our study uses over 4,100 question responses from 7 instructors teaching 26 courses at 4 North American institutions over the past 3 academic years, shown in Table 1. Institutions are described in Table 2. All instructors were using iclicker software, which automatically collected student responses and images of questions as a side effect of instructors using PI in their classes.

#### 3.2 Labeling

Each question was labeled using a tool built by one of the authors. This allowed us to: (1) mark solutions, (2) link solo and group votes, and (3) classify the questions into categories.

Here are the labels used to categorize questions:

- **Paired** These are questions asked twice, with a solo vote followed by discussion and group vote. We matched each set of paired votes together, allowing us to compare solo and group votes. In the rest of this work, we count a set of votes as a single PI question.

- **Single** These votes are taken only once, likely because the results indicate that the class already understands the concept or because the instructor failed to close voting between rounds.

- **Quiz** These are votes taken as a reading quiz at the beginning of the period. Unlike the other votes, they are graded for correctness rather than participation.

- **non-MCQ** These non-multiple choice questions represent deliberate use of the clicker for purposes other than standard questions, such as using the timer for in-class discussions or activities, attendance taking with zero-content questions (“Are you here?”), and administrative polls (“When would you prefer the review session?”). We also counted questions as non-MCQ if they relied on information outside of the slide, such as a handout not captured by the software.

From these categories, we count the paired and single votes as PI questions. Although single votes do not follow the entire PI process (solo vote, discussion, group vote), some instructors skip the last two parts when the class does very well in the solo vote, indicating the question is too easy to lead to useful peer discussion. Instead, the instructor leads a classroom-wide discussion immediately after the solo vote.

We also marked votes that were not part of a course, such as those triggered by accident, class preparation, or “ghost votes” to help students identify their clickers’ unique ID. These were removed from our data and do not appear in our results.

This tool is available for other researchers at: https://github.com/jspacco/iclickerviewer

#### 3.3 Threats to Validity

This data was collected in a natural setting, and as such suffered from both human and computer error. We have done our best to label...
When this was not possible. Occasionally the clicker software did not save data correctly; these questions were also labeled as errors. As with any human labeled dataset, it is possible there are flaws in labeling. Finally, data collection was voluntary and instructors who contributed their data may represent "better" instructors (or ones more comfortable with PI).

### 3.4 Research Questions

This work looks at the following research questions:

1. How are instructors actually using clickers? What percentage of questions use the entire process (solo vote, followed by group discussion and group vote)?
2. How many PI questions are used?
3. How much time do instructors allow for student discussion?
4. How difficult are the questions that instructors design?
5. How much do students learn from PI questions, reflected by gain in correct answers between the solo and group vote?
6. How does instructor use of PI change over time within a course?

These questions were designed to compare our real-world dataset to previous work and advice about PI.

### 4 RESULTS

In this section we look at how our research questions are answered by this dataset, and compare our results to that of existing work.

#### 4.1 Quizzes, Surveys and Discussions: How Instructors Are Using Clickers

To answer our first research question, we labeled clicker questions using the labels described in Section 3.2. As shown in Table 3, all instructors used clickers for more than just paired questions. In the non-MCQ category, we saw clickers frequently used for course surveys and as a timer for non-multiple choice exercises. Clickers are used for administering reading quizzes by two instructors, both at the same institution, and in one course, make up nearly 50% of the questions. We should note that in all classes, the majority of questions were ungraded, discussion-eligible, questions: this is important as grading clicker questions for correctness appears correlated with lower levels of student satisfaction with PI [14].

We also see a wide variance in the percentage of paired versus single peer instruction questions. The two outliers, instructors F and G report using clicker software in unintended ways due to unfamiliarity with the tool. Instructor G describes his usage thusly: “While I probably didn’t push for group discussion enough, I did use it regularly. I tended to push for individual votes quickly, identified an issue to discuss, and then moved to group discussion ... without closing the poll.” It is worth noting that while their clicker data over-reports single questions, the students still regularly engaged in the group discussion required for PI. We exclude data from these instructors from further metrics as this makes their data uninterpretable, leaving 3915 clicker questions from 5 instructors.

The range of usage for clickers, including quiz questions, single questions, and survey type questions, indicates that researchers may want to dig deeper into self-reported PI classes, since fewer discussion questions may result in less learning gains for students[16, 27].

#### 4.2 Timing and Question Quantity

For our second and third research questions, we looked at the amount of time the clicker software recorded the question as being open for voting, as well as the number of questions recorded in each session. We compare this to the existing research and advice on timing in PI. Beatty et al. [1] argue for "Question-driven Instruction", in which topics are introduced by questions discussed "a few minutes" before a vote followed by a class-wide discussion, and then optionally a mini-lecture on the topic. They recommend three or four discussion questions per 50 minute lecture. As shown in Table 4, we see three of our instructors use a bit over 3 questions per class and two of them use around 6. This is in the same range as Porter et al. [14], which looked at 7 courses and found 4–7 clicker questions per class.

Table 5 shows the average time each instructor gave for PI questions.

<table>
<thead>
<tr>
<th>Instructor</th>
<th># courses</th>
<th>Avg/Class</th>
<th>Avg/Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>75.3</td>
<td>3.2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>108.3</td>
<td>3.4</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>86.3</td>
<td>5.6</td>
</tr>
<tr>
<td>D</td>
<td>8</td>
<td>208.9</td>
<td>6.4</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>119.0</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Table 4: Question statistics for each instructor. We include paired, single, and non-MCQ questions. For paired questions, each pair of questions counts as a single question.

Table 5: Number of seconds instructors wait for students to answer PI questions.

G report using clicker software in unintended ways due to unfamiliarity with the tool. Instructor G describes his usage thusly: “While I probably didn’t push for group discussion enough, I did use it regularly. I tended to push for individual votes quickly, identified an issue to discuss, and then moved to group discussion ... without closing the poll.” It is worth noting that while their clicker data over-reports single questions, the students still regularly engaged in the group discussion required for PI. We exclude data from these instructors from further metrics as this makes their data uninterpretable, leaving 3915 clicker questions from 5 instructors.

The range of usage for clickers, including quiz questions, single questions, and survey type questions, indicates that researchers may want to dig deeper into self-reported PI classes, since fewer discussion questions may result in less learning gains for students[16, 27].

#### 4.3 Question Type by Instructor

Table 3: Question type by instructor. * indicates that the instructor reported keeping the poll open across the individual and group votes.

Table 3: Question type by instructor. * indicates that the instructor reported keeping the poll open across the individual and group votes.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Paired</th>
<th>Single</th>
<th>Quiz</th>
<th>Non.MCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75.7%</td>
<td>14.2%</td>
<td>0.0%</td>
<td>10.2%</td>
</tr>
<tr>
<td>B</td>
<td>37.9%</td>
<td>3.2%</td>
<td>45.9%</td>
<td>13.0%</td>
</tr>
<tr>
<td>C</td>
<td>83.4%</td>
<td>12.2%</td>
<td>0.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>D</td>
<td>70.5%</td>
<td>1.1%</td>
<td>26.7%</td>
<td>1.7%</td>
</tr>
<tr>
<td>E</td>
<td>94.1%</td>
<td>5.9%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>F</td>
<td>22.2%</td>
<td>77.8%*</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>G</td>
<td>15.7%</td>
<td>80.6%*</td>
<td>0.0%</td>
<td>3.7%</td>
</tr>
</tbody>
</table>

Table 4: Question statistics for each instructor. We include paired, single, and non-MCQ questions. For paired questions, each pair of questions counts as a single question.

Instructor # courses | Avg/Class | Avg/Single |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>67.0</td>
<td>110.4</td>
</tr>
<tr>
<td>B</td>
<td>34.4</td>
<td>79.0</td>
</tr>
<tr>
<td>C</td>
<td>65.4</td>
<td>126.2</td>
</tr>
<tr>
<td>D</td>
<td>52.8</td>
<td>65.7</td>
</tr>
<tr>
<td>E</td>
<td>64.6</td>
<td>72.9</td>
</tr>
</tbody>
</table>
Table 6: Fraction of easy, medium, and hard questions by instructor, using difficulty levels from Smith et al. [23]. Includes both solo votes and single votes.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15%</td>
<td>49%</td>
<td>36%</td>
</tr>
<tr>
<td>B</td>
<td>21.2%</td>
<td>51.9%</td>
<td>26.8%</td>
</tr>
<tr>
<td>C</td>
<td>10.1%</td>
<td>51.2%</td>
<td>38.7%</td>
</tr>
<tr>
<td>D</td>
<td>21.4%</td>
<td>59.3%</td>
<td>19.3%</td>
</tr>
<tr>
<td>E</td>
<td>45.4%</td>
<td>35.3%</td>
<td>19.3%</td>
</tr>
</tbody>
</table>

Table 7: Percentages of correct answers for each vote of a paired question, and for single questions.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Paired, Solo Vote</th>
<th>Paired, Group Vote</th>
<th>Single Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41.3%</td>
<td>55.7%</td>
<td>70.8%</td>
</tr>
<tr>
<td>B</td>
<td>49.6%</td>
<td>61.8%</td>
<td>75.1%</td>
</tr>
<tr>
<td>C</td>
<td>40.5%</td>
<td>53.5%</td>
<td>63.6%</td>
</tr>
<tr>
<td>D</td>
<td>53.3%</td>
<td>62.9%</td>
<td>61.9%</td>
</tr>
<tr>
<td>E</td>
<td>60.5%</td>
<td>70.8%</td>
<td>72.1%</td>
</tr>
</tbody>
</table>

Figure 1: Percent correct on the solo vote of paired questions and single questions, binned by deciles.

Figure 2: Paired questions, with solo vote correctness on the x-axis, and group vote correctness on the y-axis. The dashed line is the xy-axis (points above this line show gains from solo to group vote), and the solid line is the regression trend line.

4.3 Question Difficulty

To answer our fourth research question on how difficult instructor-developed PI questions are, we look at the percentage of students who correctly answer the first vote, a difficulty metric which is often used within existing PI literature. There is no hard and fast rule on what percentage is best, but Crouch et al. [2, 3] recommend between 35% and 70% correct on the initial vote, and Smith et al. [22] recommend below 80%. Zingaro and Porter [27] divide questions into “easy” and “difficult”, where hard is below 50% correct on the initial vote, and easy is above 50%. They argue for more difficult questions, pointing out that in their study, difficult questions lead to a 25% student improvement on isomorphic questions after group discussion, and 42% after instructor lead discussion, suggesting that more difficult questions provide the greatest student learning gains.

Smith et al. [23] establishes cutoffs for easy, medium, and hard questions, where easy is more than 70% correct on the first vote, medium is 35–70% correct (roughly aligning with Mazur’s guideline), and hard is less than 35%. These cut-offs are also used by Porter et al. [15, 16] and Simon et al. [20]. Table 6 shows that for most instructors, above 50% of the questions fall in the medium, good question range, indicating that instructors, even when beginning PI, develop clicker questions with the suggested difficulty level.

Figure 1 shows percent correct for the initial, solo vote of paired questions, binned by decile across all instructors. We see a clear normal distribution centered around the 50% percent correct range. The vast majority of the questions fall in 35-75% recommended range.

Table 7 shows question correctness broken down by solo, group and single vote. We see that single votes have the highest percent correct, which follows the intuition that instructors will omit group discussion on questions that a large majority of their students already get correct. We also see that the group vote is higher than the solo vote, indicating that students are learning from group discussion.

4.4 Normalized Learning Gains

Our fifth research question asks how student correctness changes from the solo vote to the group vote. Figure 2 provides each paired question plotted with the percent correct in the solo vote as the x-axis, and the percent correct on the group vote as the y-axis. First, as
one might expect, we see that there is a strong correlation between the two, with most points clustering around a line slightly above center. Second, we see that most questions have students perform better on group vote than solo (as we might hope).

To aggregate the results of multiple questions, we use two measures of this improvement. Raw learning gain (RG) for a single question is the difference between the percents correct on the solo and group votes. Since the RG achievable on a question depends on the correctness of the solo vote, normalized learning gain (NG) is commonly used instead [15, 20, 24]. We use the version of the NG metric defined in Marx and Cummings [10], shown below:

$$NG = \begin{cases} 
100 \times \frac{\text{group} - \text{solo}}{\text{solo}} & \text{if group} > \text{solo} \\
100 \times \frac{\text{group} - \text{solo}}{\text{group} - \text{solo}} & \text{if group} \leq \text{solo}
\end{cases}$$

When averaging the gain over multiple questions, we compute NG per question and then average the results. Normalized gain scales the difference between group and solo vote by the percentage of students who answered the initial vote incorrectly, allowing for comparisons of improvement between both hard and easy questions. However, it may penalize hard questions, as hard questions need more improvement to achieve the same normalized gain as easy questions. To avoid this, we compare the normalized gain of easy, hard and medium questions separately, as well as providing a single normalized gain metric for comparison with other work.

Table 8 shows normalized gain broken down by the instructor. We see that in general, hard questions show a lower normalized gain, as expected for the metric. In the CS PI literature, Porter et al. [15] report a NG of 0.65 for easy, 0.55 for medium and 0.31 for hard. Zingaro [24] report a 0.29 NG for a remedial CS course, Simon et al. [20] report a 0.41 NG for CS1 and 0.35 NG for CS 1.5 [20], and Zingaro and Porter [27] report a 0.44 NG for CS1 (0.34 for hard questions, and 0.49 for easy). While our normalized gain is slightly lower, it is in the range of these reports. Previous work has used isomorphic questions to show that normalized gain reflects student learning in the discussion section [16, 27]. Normalized gain in the same range as previously reported work is a promising sign that students are learning from peer instruction "in the wild", not just in the context of courses being taught by researchers studying peer instruction.

### 4.5 Instructor Use of Peer Instruction Over Time

For our sixth research question, we consider how many PI questions instructors ask in classes that they have taught multiple times. Tables 9, 10, and 11 give question statistics for multiple offerings in three different situations.

The data in Table 9 comes from an instructor newly adopting PI; the first offering is their second term using PI. The number of questions increases fairly steadily as the course is converted more completely to the new pedagogy. The data in Table 10 comes from an instructor experienced with PI but new to this course designed the slides. They taught all offerings except Spring 16, when a different instructor used the slides. Fall 16 is the only version this course offered MWF rather than TuThr.
the number of questions per class, which is quite steady, suggesting that the instructor has reached their desired level of questions while the other instructors are still converting the class into the PI format. Although based on only a few instructors, this suggests that instructors new to PI convert their courses to the new format over a period of multiple terms, (note that Avg/Class is still significantly lower for the first instructor than the other two) and that even instructors experienced with PI may spend several terms converting a new course.

While that message may sound discouraging, Table 9 tells a very positive story about adopting PI. In Winter 15, the second term in which this instructor uses PI, they have an average NG of 0.14. However, the following term it doubles to 0.28, and then stays consistently between 0.28 and 0.31. We see this as an instructor who becomes proficient with PI and then consistently uses it well within only two terms. Thus, proficiency with the technique comes relatively quickly even though the course continues to evolve. Also, once proficiency is achieved, all the instructors see a fairly steady average NG.

Overall, Tables 9–11 seem to show that course-level average gain (RG and NG) is fairly steady for repeated offerings of a course once the instructor has become familiar with PI, but it would be interesting to see how stable the gain is at lower granularity (e.g., at the lecture or the question level). Also, different instructors stabilized at different gain levels, which begs the question of whether the differences are caused by how instructors used PI, the course material, the level of the course in the curriculum, or aspects of the student population.

5 CONCLUSION AND FUTURE WORK
In this work, we look at a labeled collection of over 4,000 posed PI questions collected by 7 different instructors over 3 years as a side effect of using PI in their teaching. We use this dataset to examine how instructors use PI in their classes, and how this usage compares to existing literature and best practices. We find that the majority of instructor questions fall within recommended difficulty levels, and that student learning gains between the solo and group vote are similar to those previously reported. We see evidence that new adopters can deploy PI effectively within two terms. Lastly, we find that in repeated courses, instructors include more PI questions over time. Overall, we find that PI “in the wild” behaves similarly as in the quasi-experimental setting where it was first examined by researchers.

This paper offers a first look at this dataset. We plan to investigate this data further, including research into how instructor use of PI questions changes over repeated course offerings, common PI question themes, and characteristics of successful PI questions.

6 ACKNOWLEDGEMENTS
Thank you to everyone who contributed data. Researchers who are interested in contributing clicker data for future work, or using our dataset in their projects are encouraged to contact us.

REFERENCES