Peer Instruction: Do Students Really Learn from Peer Discussion in Computing?

Leo Porter, Cynthia Bailey Lee, Beth Simon
Computer Science and Engr. Dept.
University of California, San Diego
La Jolla, CA USA
+1 858 534 5419
{leporter,clbailey,esimon}@ucsd.edu

Daniel Zingaro
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
daniel.zingaro@utoronto.ca

ABSTRACT
Peer Instruction (PI) is an instructional approach that engages students in constructing their own understanding of concepts. Students individually respond to a question, discuss with peers, and respond to the same question again. In general, the peer discussion portion of PI leads to an increase in the number of students answering a question correctly. But are these students really learning, or are they just "copying" the right answer from someone in their group? In an article in the journal Science, Smith et al. affirm that genetics students individually learn from discussion: having discussed a first question with their peers, students are better able to correctly, individually answer a second, conceptually-related question. We replicate their study, finding that students in upper-division computing courses (architecture and theory of computation) also learn from peer discussions, and explore differences between our results and those of Smith et al. Our work reveals that using raw percentage gains between paired questions may not fully illuminate the value of peer discussion. We define a new metric, Weighted Learning Gain, which better reflects the learning value of discussion. By applying this metric to both genetics and computing courses, we consistently find that 85-89% of "potential learners" benefit from peer discussion.

Categories and Subject Descriptors
K.3.2 [**Computer Science Education**].

General Terms
Algorithms, Human Factors

Keywords
Peer instruction, Clickers, PRS, Classroom response, Active learning.

1. INTRODUCTION
Based on observations that students learn very little in traditional physics lectures, Mazur and colleagues [4] developed and refined the active learning Peer Instruction (PI) pedagogy [2,4,7]. The core feature of PI is the multiple-choice question (MCQ): students begin by individually answering an MCQ (the individual vote), then discuss it with peers, and finally re-vote in light of that discussion (the group vote). Clickers–small, wireless keypads with buttons corresponding to response choices–are often used to allow a measure of student anonymity and accurate estimates of class ability by the instructor [2]; such estimates are useful for determining the direction of instructor-led discussion following each question.

The majority of early educational research in PI was done in physics. In that work, a commonly-reported metric is normalized gain (NG), which measures the improvement of students as a fraction of the total possible improvement. For example, if a student scores 60% on a pre-test and scores 80% on a post-test, their NG is 50%, meaning they learned 50% of what "remained for them to learn". On the Force Concept Inventory (FCI), a standard physics concept inventory test, NG from pre- to post-course in PI classes were double that of traditional classes [4].

In computing, we have no widely-available concept inventories, so the above approach cannot be used to measure PI effectiveness. Instead, researchers have calculated NG on a per-question basis, and have found positive results. For example, one study found an NG of 41% in CS1 and 35% in CS1.5 [18], and a study of a remedial CS1 course reported an NG of 29% [22]. But, by measuring PI effectiveness in this way, the question must be asked: is this gain due to improved student understanding or due to students passively agreeing with neighbors [8]? In this replication and extension of [19], we seek to answer this question in the computing context by examining student performance on consecutive isomorphic questions (questions designed to exercise the same conceptual understanding). To the extent that students can individually answer an isomorphic question correctly, we have evidence of real understanding rather than peer-mirroring.

Isomorphic questions enable instructors to evaluate the learning gains provided by group discussion. Direct learning gains from discussion are demonstrated by students individually answering an initial question *incorrectly*, participating in a group discussion, answering that initial question *correctly*, and then, most critically, answering a new, conceptually-similar question *correctly*. In prior work [19], 16% of biology students in a class achieved these direct learning gains. Our work evaluates these gains in two different computer science classes: computer architecture and theory of computation, each taught by a different instructor. In these classes, the percentages of students who demonstrated these direct gains from the group discussion were 20% and 13% respectively.

In addition to demonstrating direct learning gains for students and thus reproducing results from Smith et al. [19], we identify differences between subject disciplines, define a new metric useful for exploring benefits of peer discussion, motivate the use of isomorphic questions in PI, and provide guidance for the development of isomorphic questions for computer science classes.

ICER'11, August 8-9, 2011, Providence, Rhode Island, USA.

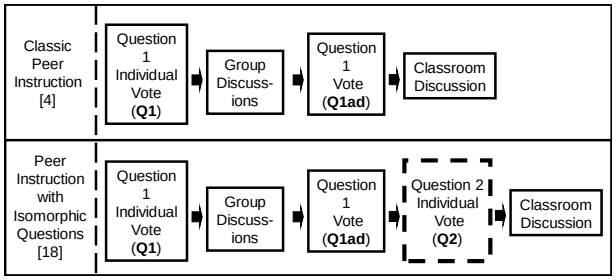

Figure 1. The Isomorphic Testing Process.

2. BACKGROUND & RELATED WORK

The core of classic PI (see Figure 1) is the in-class process of posing a deep conceptual question to students: having students answer the question individually, discuss it briefly with their seatmates, and vote again in light of increased understanding. However, to be most effective, PI requires other concordant course changes. In order to use class time most productively, many teachers require students to come prepared to discuss conceptual questions. Preparation can be solicited through pre-lecture reading and associated quizzes [4], pre-lecture screencasts [3], or clicker quizzes at the start of class [10]. Others suggest that PI-informed pedagogy should also occur in lab and tutorial sessions, not just lecture [4,22].

Key to the success of PI is the use of challenging conceptual questions that target common student misconceptions and core course concepts. Beatty et al. [1] offer helpful "tactics" for focusing student awareness, evoking cognitive processes, and promoting productive small-group and class-wide discussion. Other work provides best-practice tips for clicker use [2,21].

Several recent studies report the benefits of PI in computing [5,15,16,18,22]. In addition to consistently notable NGs, these studies use surveys to elicit student opinions and attitudes regarding PI. Students overwhelmingly support the effectiveness of PI for their learning, and recommend that it be used in further courses. In addition to student benefits, studies report various instructor benefits including a sharpened focus on student difficulties, an improved ability to adapt lectures, and an ability to involve students in teamwork and collaboration [16,18].

2.1 Isomorphic Questions in PI

Performance on individual MCQs and post-course student surveys tell us little about the long-term effects of answering MCQs. In the context of biology, Smith et al. [19] modified the classic PI format to assess the extent to which students were learning robust concepts, rather than copying peers or obtaining fragile, context-bound representations. Sixteen times throughout the term, students consecutively voted three times with no teacher intervention. Figure 1 contrasts classic PI with this variant where Q2 is a question isomorphic with Q1.

First, students answered a question individually (Q1). Next, they discussed that question in groups and voted again (Q1ad). Finally, and without seeing the answer to Q1, students were presented with a conceptually-similar isomorphic question (Q2) to which they responded individually. Isomorphic questions are questions that "look different" on the surface, but are simply "cover story" variations [19] that require students to adapt and apply the same core concept.

Smith et al.[19] found that the average correctness on Q2 was higher than the average correctness on both Q1 and Q1ad. Additionally, of those students who answered Q1 *incorrectly* and Q1ad *correctly* (i.e. those students who could have "copied" from their neighbors), 77% answered Q2 correctly. Finally, when looking at questions by difficulty, the increase from Q1ad to Q2 was more pronounced for the most difficult questions than for questions of other difficulties.

Similar research can be found in [17], where sequences of "rapid-fire" questions were used to free students from context-specific learning of concepts. The general finding is that on each subsequent question targeting the same concept, the student correctness rate increases.

2.2 The Value of Isomorphic Questions

One can evoke a social constructivist epistemology to understand the source of value in PI discussions [14]. Specifically, individual understanding arises powerfully in a context suffused both with opportunities to be active and opportunities to discuss with like-ability peers. Individually thinking about a question spurs the creation of an initial mental model, which can be refined through the ensuing peer discussion. Of importance is that these mental models be appropriate (i.e. match reality) and that students apply such models consistently in a variety of contexts [13].

Most commonly, teachers use one PI question per concept [17]. Students are expected to be able to apply that concept to a variety of situations. Unfortunately, a single MCQ may not be sufficient to alert teachers and students to potential difficulties in applying and abstracting concepts across situations. Ma et al. [13], for example, found that 41% of sampled CS1 students exhibited both a viable and a non-viable model for the same concept (reference assignment) on a course exam. On the specificity and rigidity of mental models, Jonassen notes: "What often makes human models weak and oversimplified is that they fail to identify relevant factors and are not dynamic, that is, they do not represent change in factors over time" [9]. Researchers have therefore investigated the use of multiple, sequential, isomorphic questions for both giving students more practice-per-concept and assessing mastery of concepts [17].

How are we to interpret the situation where students answer a first question correctly, but then answer a follow-up isomorphic question incorrectly? One hypothesis is that the students lack a coherent and generalizable model of the underlying concept. Research shows, however, that surface feature differences—the very features that we change in order to create isomorphic questions—can themselves substantially impact students' performance, even in the presence of robust conceptual understanding.

For example, several studies have found drastic differences in student performance on various versions of the Tower of Hanoi problem [11], where all versions are isomorphic regarding problem space and transformations between states. Two such isomorphisms include "move isomorphisms" (where discs are physically moved) and "change isomorphisms" (where discs do not move, but are made larger or smaller in-place). The general finding is that change isomorphisms are twice as difficult as move isomorphisms, and that this difference must lie in the way that subjects imagine or model the different situations. Kotovsky et al. [11] suggest several hypotheses; for example, that isomorphic questions differ linguistically, that they differ in the ease with which their "rules" can be applied, that they are more or less consistent with real-world knowledge, that they impose different cognitive demands, and that they elicit differing internal

representations of problem state. This highlights an important caution: questions deemed isomorphic by instructors are not necessarily isomorphic to students.

We are interested in the extent to which the findings of Smith et al. [19] (biology) and Reay et al.[17] (physics) apply to computing. First, in a discipline with no agreed-upon concept inventory, to what extent can we create questions that are isomorphic? Second, if a student answers Q1 incorrectly, they presumably have a fragile understanding of the underlying concept. Given the tenacity of misconceptions and non-viable concepts reported in much computing literature, is it possible for students to grasp a concept in the mere minutes that elapse between Q1 and Q2?

3. METHODOLOGY

Isomorphic questions were tested in two classes under various administrative and experimental controls. In addition, students self-reported their beliefs and experiences with PI through surveys.

3.1 Courses Included in this Study

Two courses were included in this study. Both are upper-division required majors courses, taught at a large R1 institution during the summer of 2010.

Introduction to Computer Architecture (N=51) focuses on instruction set and processor design. The instructor, a senior graduate student in the area, had previously taught the course at a different institution, had been a teaching assistant (TA) for this course multiple times, but had not used PI. The course met for six contact hours per week (four 1.5-hour lectures), for five weeks. Class participation (answering at least 80% of the PI questions in each meeting) was worth five percent of the course grade and reading quizzes were worth three percent of the course grade.

Introduction to the Theory of Computation (N=45) focuses on automata and proof-writing skills. The instructor had some prior teaching experience (four courses), but no prior experience using PI, nor teaching this specific course. The course met for six contact hours per week (two three-hour lectures), for five weeks. Class participation (answering at least 80% of the PI questions in each meeting) was worth six percent of the course grade and reading quizzes were worth four percent of the course grade.

Both instructors had observed PI as adopted in an introductory computing course at their institution and received 1-2 hours of advice in creating questions from the instructor of that class. Recent evidence has been reported showing inexperienced teachers whose instructional practices are grounded in research-based methodologies can be highly effective [6].

3.2 Methodology for Creating Isomorphic Questions

To create isomorphic questions, instructors initially identified several core concepts for each course. Then they identified a common misunderstanding or key element of that concept to emphasize in an MCQ.

After designing one MCQ, a second was created, patterned after the first in both the concept being tested and the targeted misconception or key idea related to that concept. Typically, several iterations of edits to both questions were required to bring them into alignment in terms of equal difficulty, as well as to remove similarities that would render one answerable based superficially on the other.

Each instructor enlisted a colleague with experience in the course, either as instructor or TA, who reviewed final drafts of the questions for accuracy and equal difficulty.

3.3 Question Administration and Experimental Controls

In order to assess the learning value of peer discussion, in isolation from other aspects of the classroom experience such as instructor explanation, the following restrictions were observed:

First, the question order of Q1 and Q2 was determined by random coin toss. Although the instructors endeavored to create equally difficult questions, this precaution further eliminated potential for instructor bias toward presenting an easier or harder question first.

Second, between the start of Q1 and the conclusion of collecting student responses to Q2, the instructor did not provide guidance or explanation.

Third, students were not shown the correct response to Q1 prior to responding to Q2, and, moreover, were not shown graphs of the class' responses, from which they might deduce the likely correct response.

Fourth, student consultation with peers during Q1 and Q2 was prevented by enforcing strict silence at those times. No group was permitted to begin group discussion until all Q1 responses had been collected.

Finally, using experienced instructors from outside of our local context, a post-hoc analysis was performed to confirm the equal difficulty and content of the question sets in both courses.

3.4 Student Survey

At the conclusion of the term, students were asked to respond to a questionnaire about their experiences using PI, and their perceptions of its usefulness for learning. The survey was required for course credit. In the architecture course, all but three students participated; in the theory course, all students participated. In both cases, students were advised that instructors would not see their responses until after final grades were submitted.

The questions that were adopted verbatim from the replicated study [19] used a five-point Likert scale. Several additional questions were included, and these used a six-point forced-choice Likert scale (i.e, no neutral option).

4. RESULTS

There are three components of our results. The first is a motivating example to show how two similar individual and group votes can mask different learning gains which can be elucidated through a second isomorphic question. The second is an analysis that compares the isomorphic results from the two computing courses and the work of Smith et al. [19]. Lastly, an analysis of student perspectives is provided.

4.1 Value of Isomorphic Questions

Classic PI has a single individual vote followed by a group discussion and group vote. Typically, the instructor then adapts the classroom discussion and follow-on material based on the percentage of students correctly answering the group vote [16]. For example, although some instructors encourage students to explain their reasoning (not just the answer) for both correct and

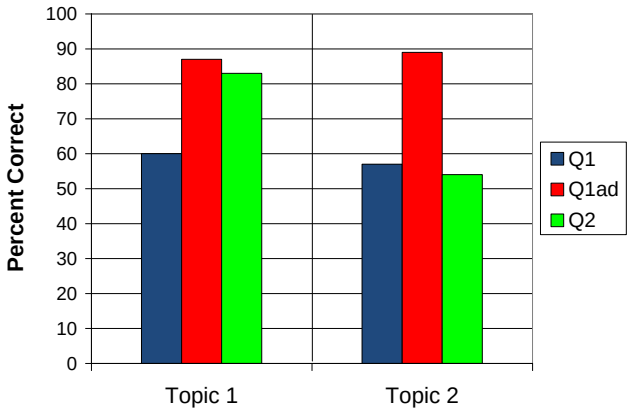

Figure 2: Percentage of students answering correctly for Q1, Q1ad, and Q2 for two topics from the architecture class.

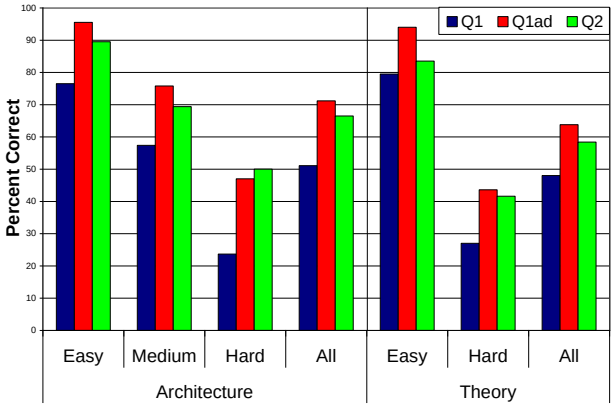

Figure 3: Percentage of students answering correctly for Q1, Q1ad, and Q2.

incorrect choices, it can be enticing for an instructor to hurry discussion in the presence of a compelling correct response rate.

Taken from the architecture class, two sets of questions with a high correct response rate on the group vote serve to illustrate the value of asking a follow-on isomorphic question. Figure 2 shows the average percentage of students responding correctly in Q1, Q1ad and Q2, for two different topics. In both cases, the percentage of students answering Q1 and Q1ad are comparable. The large percentage of students answering Q1ad correctly could indicate to the instructor that the students now understand the topic and the class-wide discussion can be brief. However, the Q1ad vote can be misleading. In Topic 1, the students demonstrate that their understanding of the material after the group discussion is strong enough that it can be applied to new topics. In the case of Topic 2, student understanding was not only fragile, but peer discussion may have confused some of the originally-correct students, or students may have been able to come up with the right answer without a deep understanding of the core issues. Without an isomorphic question, an instructor may struggle to determine whether a situation like Topic 1 or Topic 2 has occurred.

4.2 Student Learning Results

As presented in Section 4.1, the second isomorphic question is valuable in determining whether the group discussion resulted in generalizable learning. In that vein, Figure 3 provides the average percentage of students responding correctly to Q1, Q1ad, and Q2, in each of the two courses. Averages across Easy, Medium, and Hard question categories are provided. The classification of questions as Easy, Medium, or Hard was done according to the percentage of students responding correctly to Q1, as was done by Smith et al. [19]. All question categories in both courses show improved correctness from Q1 to Q1ad. The critical component—the component that demonstrates that the group discussion was helpful—is the gain from Q1 to Q2, which is positive for all classes. The difference between Q1ad and Q2 is also important as it may indicate fragile student understanding; i.e. they could understand Q1, but could not apply that understanding to Q2 [12]. In nearly all cases, there was some decrease in the percentage responding correctly between Q1ad and Q2, though Q2 was still higher than Q1. (The one exception is the Hard question category from architecture; on average, more students responded correctly

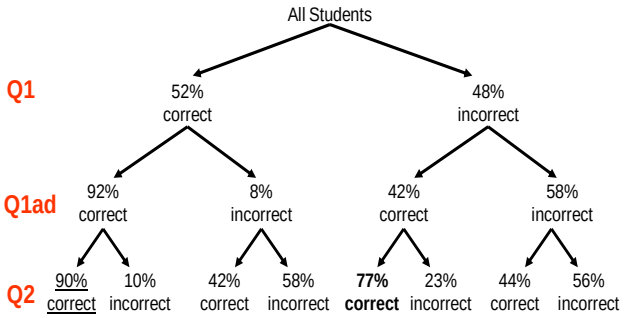

Figure 4: Flow chart from Smith et al.

to Q2 than Q1ad. Further discussion is in Section 5.) These data, demonstrating improvement between Q1 and Q2, are consistent with the finding that students learn from group discussion, and that not all of the Q1ad improvement was due to "copying" from a knowledgeable group member. The next set of diagrams further examines from which student populations these gains stem.

Figures 4-7 trace student response patterns over three-question isomorphic sequences. For each tree, the top two branches correspond to the percentage of all students responding correctly (left branch) and incorrectly (right branch) to Q1. The next layer down are the percentages (for each of the previously split groups) who answered correctly (left) and incorrectly (right) to Q1ad. Percentages are relative in these figures. In Figure 4, for example, 92% *of the 52% of students who answered Q1 correctly* went on to answer Q1ad correctly again following the group discussion. Figure 4 contains the biology class results from Smith et al. [19]. Percentages are the average across all questions from the term that were reported in that study. Figure 5 provides the results from the architecture course, broken out into the averages of the Easy, Medium, and Hard questions, respectively, and the average of All questions. Figure 6 presents the average results for Easy, Hard, and All questions from Theory of Computation.

In Figures 4-7, the number in each tree which deserves the most attention is the one in bold font. This result depicts students who incorrectly answer Q1, then correctly answer Q1ad, and then correctly answer Q2. These students are gaining appreciable benefit from the group discussion as they did not understand the

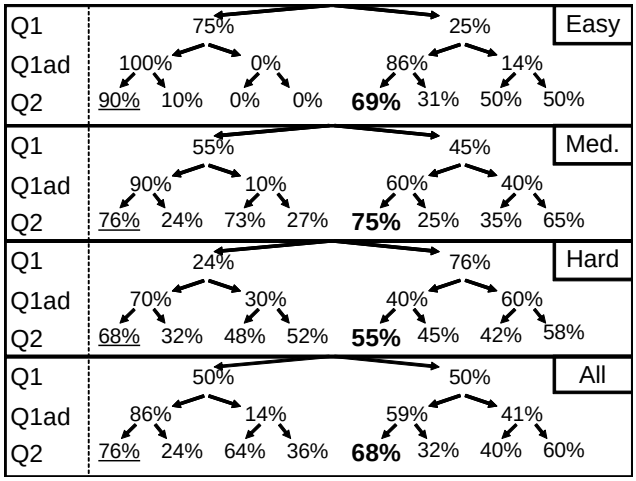

Figure 5: Flow chart for architecture grouped by easy, medium, hard, and overall average. Similar to prior results, left indicates correct, right indicates incorrect.

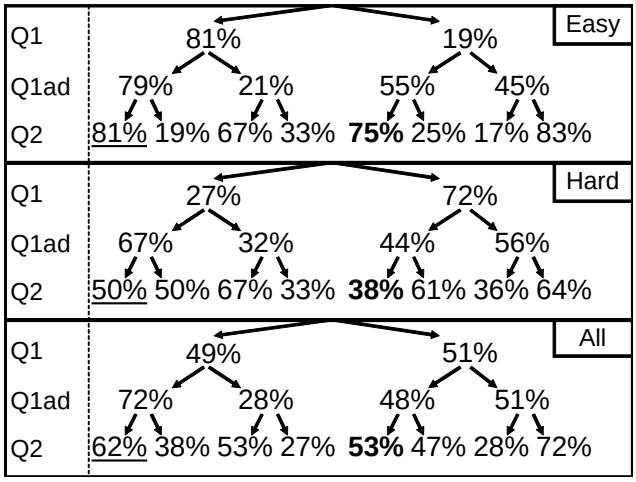

Figure 6: Flow chart for theory of computation grouped by easy, hard, and overall average. Similar to prior results, left indicates correct, right indicates incorrect.

concept at the first individual vote, learned from their group during the discussion, and then were able to apply that understanding to the next question. In each category, learning gains in computing are comparable to those provided for biology in Smith et al. [19].

Rather than using relative percentages, Figure 7 compares the all-question average for the three courses using absolute percentages. This helps capture the effect of the initial difficulty of the question–the fewer students getting Q1 incorrect, the greater potential benefit there is for discussion to be beneficial for students.[1] The most critical group (Q1 incorrect, Q1ad correct, Q2 correct) is again bolded. We can see that in each class (Smith's, architecture, and theory), somewhere between 13% and 20% of the students had demonstrable gains in learning from their discussion. One other notable difference between Smith's results

[1] There is no hard and fast rule for what defines a "good clicker question" based on how many students should get it right in the first vote, but Mazur recommends between 35-70% [4] and a faculty handbook advises against too easy of questions [21].

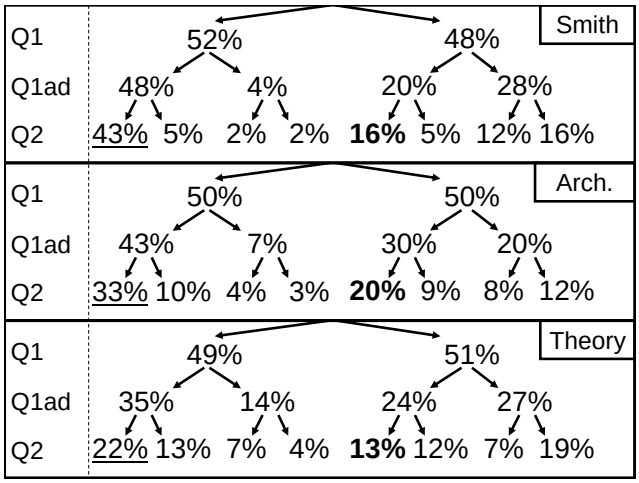

Figure 7: Flow chart for each class with percentage of students computed cumulatively.

in biology and those in computing is the percentage of students who answer Q1 correctly, Q1ad correctly, and Q2 incorrectly. The percentage of students in that group for Smith et al.'s biology class, the architecture class, and the theory class are 5%, 10%, and 13% respectively. These results will be discussed in more detail in Section 5.

4.3 Student Survey Results

Students broadly believed that "clickers with discussion is valuable for my learning" (80% and 86% agree for architecture and theory, respectively). They also believed that peer discussion caused them to learn more during lecture than they would otherwise. Responding to the statement: "If we completely took away the option for you to talk to your classmates during clicker questions, it would increase (or not change) the amount I learn during lecture," 73% in architecture and 82% in theory disagreed. Representative student comments include "I have learned different ways at approaching a question by the way group members explain it. To me, it has been highly valuable," and "Explaining why we think one answer is right over another helps us formalize the logic of why we chose that answer."

In architecture and theory (respectively), when asked if a knowledgeable group-mate was necessary to have a good discussion, 44% and 44% agreed, 25% and 34% were neutral, and 31% and 22% disagreed. Thus, in regard to learning in discussion, student opinion favored the necessity of having a knowledgeable group-mate. This is in contrast to the result in [19], where, responding to the same question, 47% of students disagreed. A typical comment from students who said it was necessary to have someone in the group who knew the correct answer was: "This is because at least one person is needed in order to explain how to obtain the correct answer." Representative comments from students who did not feel it was necessary include, "Sometimes just discussing the concept is important, even if nobody is sure," and "If neither of us knew the correct answer, we usually each knew enough about the question to collaborate and come up with a reasonable solution together."

Finally, students overwhelmingly supported the adoption of PI by other professors. In architecture and theory (respectively), 90% and 91% of students broadly agree that "I recommend that other

instructors use our approach…in their courses." In both classes, 33% "very strongly agree" with this statement.

5. DISCUSSION

The results and experiences from our study lead to a number of areas for discussion. The first is that isomorphic questions can be valuable to instructors as a tool to test for the fragility of student understanding. The second is that the computer science questions used in our study had different result characteristics than those presented in [19] and potentially require different metrics for evaluation. The third is that student qualitative responses elicited different views of discussion and group member roles which may factor into discussion effectiveness.

5.1 Isomorphic Question Design

As demonstrated in our initial example (Figure 2), a high percentage of students correctly answering Q1ad does not necessarily imply solid conceptual understanding. Students may be copying other student responses [8] or simply have a fragile understanding of the concept [12]. Evidence of such losses from Q1ad to Q2 occurred in questions from all difficulty levels. The losses in student correctness from Q1ad to Q2 for hard questions are likely attributable to fragile knowledge [12]. One hypothesis for the drop in easy questions is that students tended to have briefer, less in-depth discussions on questions self-deemed to be trivial. Further studies would be needed to confirm this hypothesis.

Regardless of the source of the drop in the percentage of students responding correctly, the existence of that drop motivates the use of isomorphic questions. By asking follow-up questions, the instructor can evaluate the level of student understanding. This is particularly important f*or course topics where broad student understanding is crucial.*

Our instructors found that developing isomorphic questions can be difficult. Expert advice was useful as a means of feedback, but may be impractical or unavailable for other instructors adopting this pedagogical technique. In contrast, randomization of the questions was both highly useful and simple. Both instructors noticed an initial tendency to write a slightly easier question first, but when faced with the knowledge that the questions would be randomized, question difficulty became more consistent.

For the purposes of this study, students were not given any explanation by the instructor between Q1 and Q2. However, in a typical classroom, an instructor using isomorphic questions may wish to pose Q2 after leading a class-wide discussion and providing any further instruction as needed. Measuring student learning in this context would be an interesting area of exploration for future work.

5.2 Evaluating Learning Gained

In the context of isomorphic questions, a group of particular interest is the group of students who get Q1 incorrect and Q1ad correct. We define this group as the *potential learners*. This is not to say they are the only students learning from the PI process, but they define the set of students who appeared to learn through the peer group discussion[2].

[2] Also of interest, but beyond the scope of this paper, is the Q1 incorrect, Q1_ad incorrect, Q2 correct group which may have evidenced learning after seeing the concept in another context.

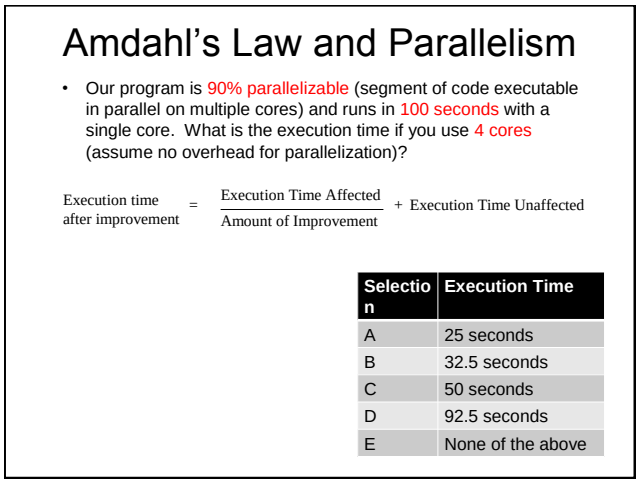

Figure 8: Q1 from an isomorphic pair of "easy" questions in the architecture class.

When examining the results for all three classes, a notable difference appeared (as mentioned in Section 4.2). Smith et al. found that, of those students who correctly answer Q1 and Q1ad, 90% proceed to correctly answer Q2. For the two classes of this study, architecture and theory, those numbers are notably smaller, at 76% and 62% respectively.

One would expect that students who initially understand the concept (answer Q1 correctly) and hold onto that understanding (answer Q1ad correctly) would be capable of answering Q2 correctly. The relatively lower numbers of students in this category for the computer science classes caused the instructors to wonder—were our questions truly isomorphic in tested content? It was true that the computing instructors were both novice users of PI, teachers with less than five years of experience, and the theory instructor was teaching that course for the first time. However, our questions had been judged isomorphic by two outside experts (with significant related teaching experience) and delivery order had been randomized compared to the order of question development. This leads us to posit that, in computing, having questions that "test the same knowledge" may be different than what is meant in biology. Anecdotally, computer science instructors hesitate to ask students "plug and chug" questions; e.g., questions where a basic algorithm or set of steps can be applied to find the solution. Although PI questions in other STEM fields are commonly labeled "ConcepTests", both our inexpert review of the isomorphic questions in [19] and our understanding of Smith's explanation of them [20] leads us to believe that differences exist in the kinds of questions computing instructors identify as isomorphic compared to those used by Smith et. al.

Examining the questions asked in computer science, the two sets of "easy" questions from architecture were more similar to the biology questions in that they asked students to apply an algorithm to solve a problem. For those easy questions, the percentage of students who answer Q1 and Q1ad correctly, but then answer Q2 incorrectly, was the same as Smith et al.'s overall percentage of 10%. For example, an atypically easy isomorphic question from the architecture class was on the topic of Amdahl's Law (Q1 appears in Figure 8). In this question, students are asked to apply the Amdahl's Law equation (provided). The correct response is 90 / 4 + 10 = 32.5 seconds (B). Q2 changes the

question to 2 cores rather than 4. Q2 is entirely the same as Q1 except for the possible responses and the number of cores. If students understand the application of Amdahl's Law in Q1, they need only apply the same equation from Q2 to Q1 to achieve the correct result. For this question, 71%, 93%, and 93% of students responded correctly to Q1, Q1ad, and Q2 respectively. 100% of students who answered Q1 correctly and Q1ad correctly also answered Q2 correctly (these students represented 61% of the class). For further reference and for more typical examples, all questions from this study appear in supporting online text ([23]). The potential similarities (and differences) between these disciplines leads us to believe that this may be an interesting area for future study.

Regardless of the potential differences in questions from the two disciplines, the percentage of students who answer each of Q1, Q1ad, and Q2 correctly provides insight into the evaluation methodology and provides a useful metric for instructors. Let us define a *control group (CG)* as those students who seem to have mastered the concept by correctly answering Q1 and Q1ad.

The *potential learner group (PLG)* are those who learned the material from the group discussion (Q1 incorrect, Q1ad correct). We can use the CG's ability to answer Q2 correctly to assist in normalizing our expectation for the PLG to correctly answer Q2. That is, if the CG did not do well in correctly answering Q2, we should reduce (or weight) our expectations for the PLG to answer correctly.

We can use the percentage of the CG who correctly answer question Q2 as a measure of the maximum of available learning for the PLG. This metric provides us with Weighted Learning Gain (WLG)—weighted by the CG—expressed in Equation (1).

$$\text{WLG} = \frac{PLG\ \%\ correct}{CG\ \%\ correct} \quad (1)$$

Using this metric, the three classes tell remarkably similar stories. The WLGs for Smith et. al's biology class, the architecture class, and the theory class are 86%, 89%, and 85% respectively. We believe this metric to be more representative of the value of discussion than the raw percentage increase.

5.3 Student Beliefs on the Nature of Discussion

As noted in the Student Survey Results (section 4.3), we found many students believed "having someone in the group who knows the correct answer is necessary in order to make the discussion productive" (69% and 78% agreed in architecture and theory respectively). We contrast this with the results in [19], where only 53% agreed. This disparity may be partially attributable to differences in the difficulty or style of questions used in each class. It may also be partially attributable to students' beliefs about what is normative in the nature of group discussion and roles within the group. Open-ended responses yielded two main categories of beliefs:

1. Students who believed it was necessary to have someone who knows the correct answer in the group often appeared to endorse a view that, *within the group, students replicate hierarchical teacher-student roles*. Representative comments include: "Someone [who] knows the answer will help others to learn," "Having someone that knows the concept is better to teach others in the group," and "…they corrected us."
2. Students who believed it was not necessary to have someone who knows the correct answer in the group often described *organizational structure and processes in which group member roles are undifferentiated*. Representative comments include: "Working together can also get you to the correct answer" and "Even if nobody knows the answer, each person might know different pieces of the answer that, together, could lead the group to the right answer". These responses were also more likely to attach value to the process, independent of the result: "There were a number of times that our group reached a wrong conclusion, but the discussion itself resulted in a better understanding of the relevant concepts" and "When people are throwing around ideas, you learn different approaches to a problem."

Many students expressed simply that they felt they could not progress without someone who knows the correct answer ("If nobody knows the answer, you get nowhere. Oftentimes, we just sit there and wait for the instructor because we're both confused"). There was not enough information to suggest whether these students endorse belief 1 or belief 2. Because this question was not designed specifically to elicit this information, more data would be needed to adequately characterize the prevalence of different beliefs about intra-group roles (e.g., targeted survey questions and follow-up interviews). This is a potential topic for further inquiry.

6. CONCLUSIONS

This paper replicates a key finding about the value of peer discussions for learning in the context of courses adopting the Peer Instruction pedagogy. Using matched-ability (isomorphic) questions, Smith et al. demonstrated that students learn from peer discussion in biology [19]. We replicate that result and find strikingly similar learning gains in two computing courses—architecture and theory of computation. During this study, we identified a phenomenon where questions deemed isomorphic by instructors were experimentally shown to be of different difficulty for students. To properly adjust for this phenomenon, we define a new metric, Weighted Learning Gain, that better measures student learning gains for isomorphic questions. We suggest that the use of this metric in future studies of isomorphic questions may not only lead to more comparable explorations of the value of peer discussion, but also illuminate issues surrounding fragile knowledge regarding specific disciplinary concepts. Lastly, we recommend that isomorphic questions be used for critical course concepts, where timely detection, acknowledgment, and discussion of misconceptions are most important.

7. REFERENCES

[1] Beatty, I. D., Gerace, W. J., Leonard, W.J., and Dufresne, R. J. Designing effective questions for classroom response system teaching. American Journal of Physics 74, 2006.

[2] Caldwell, J. E. Clickers in the large classroom: Current research and best-practice tips. CBE-Life Sciences Education 6, 2007.

[3] Carter, P. An experiment with online instruction and active learning in an introductory computing course for engineers: JiTT meets CS. 14th Western Canadian Conference on Computing Education, 2009.

[4] Crouch, C. H., and Mazur, E. Peer instruction: Ten years of experience and results. American Journal of Physics 69, 2001.

[5] Cutts, Q., Carbone, A., and van Haaster, K. Using an Electronic Voting System to Promote Active Reflection on

Coursework Feedback. In Proceedings of Intl. Conf. on Computers in Education, Melbourne, Australia, 2004.

[6] Deslauriers, L., Schelew, E., and Wieman, C. Improved Learning in a Large-Enrollment Physics Class. Science 332, 2011.

[7] Hake, R. R. Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics 66 (1), 1998.

[8] James, M.C., and Willoughby, S. Listening to student conversations during clicker questions: What you have not heard might surprise you! American Journal of Physics 79, 2011.

[9] Jonassen, D. H. Externally modeling mental models. In Learning and Instructional Technologies for the 21st Century, L. Moller, J. B. Huett, and D. M. Harvey, Eds. Springer US, 2009.

[10] Knight, J. K., and Wood, W. B. Teaching more by lecturing less. Cell Biology Education 4, 2005.

[11] Kotovsky, K., Hayes, J.R., and Simon, H.A. Why Are Some Problems Hard? Evidence from Tower of Hanoi. Cognitive Psychology 17, 1985.

[12] Lister , R., Adams, E.S., Fitzgerald, S., Fone, W., Hamer J., Lindholm, M., McCartney, R., Moström, J.E., Sanders, K., Seppälä, O., Simon, B., and Thomas, L. A multi-national study of reading and tracing skills in novice programmers, ACM SIGCSE Bulletin 36(4), 2004.

[13] Ma, L., Ferguson, J., Roper, M., and Wood, M. Investigating the viability of mental models held by novice programmers. In Proceedings of the 38th SIGCSE technical symposium on computer science education, 2007.

[14] Nicol, D. J., and Boyle, J. T. Peer Instruction versus Class-wide Discussion in Large Classes: a comparison of two interaction methods in the wired classroom. Studies in Higher Education 28(4), 2003.

[15] Pargas, R. P., and Shah, D. M. Things are clicking in computer science courses. In Proceedings of the 37th SIGCSE technical symposium on Computer science education, 2006.

[16] Porter, L., Bailey-Lee, C., Simon, B., Cutts, Q., and Zingaro, D. Experience Report: A Multi-classroom Report on the Value of Peer Instruction. In proceedings of the 16th Annual Conference on Innovation and Technology in Computer Science Education, 2011.

[17] Reay, N., Li, P., and Bao, L. Testing a new voting machine question methodology. American Journal of Physics 76, 2008.

[18] Simon, B., Kohanfars, M., Lee, J, Tamayo, K., and Cutts, Q. Experience report: Peer instruction in introductory computing. In Proceedings of the 41st SIGCSE technical symposium on computer science education, 2010.

[19] Smith, M., Wood, W., Adams, W., Wieman, C., Knight, J., Guild, N., and Su, T. Why Peer Discussion Improves Student Performance on In-Class Concept Questions. Science 323, 2009.

[20] Smith, M., Personal Correspondance, 2010.

[21] Wieman, C. and the staff of the CU and UBC Science Education Initiatives. Clicker Resource Guide, http://cwsei.ubc.ca/resources/clickers.htm.

[22] Zingaro, D. Experience report: Peer instruction in remedial computer science. In Proceedings of the 22nd World Conference on Educational Multimedia, Hypermedia & Telecommunications, 2010.

[23] http://cs.ucsd.edu/~bsimon/ICER2011_PI/