Principles of Programming Languages @ Scale: The Value of Student Collaboration

Spencer Wilson

University of Colorado Boulder

Abstract. With the growing promise of high income for advanced education in IT fields, many higher academic institutions continue to observe higher enrollment in their computer science programs and related fields [1] [2] [3]. Over time, this has led to higher enrollment in courses, limiting the units of one-on-one support available to students, prompting a change in methods to teach effectively at a larger scale. Accordingly, many courses have gone to focus on teaching at scale, or continuing their previous teaching methods with limited exploration in adaptation to scale. In this paper, we propose the use of peer-to-peer interviews to scale effective teaching in the Principles of Programming Languages course based on lessons learned from review of literature on effective education at scale.

Keywords— Education at Scale, Computer Science, Principles of Programming Languages, Peer Grading, Ungrading, Interview Grading

1 Introduction

As is standard in academia, there is a lag between increased enrollment and funds for the hiring of additional staff. Over time, we have seen the number of staff resources available per student decreases. With this student to staff interaction resource decreasing, a key value proposition disappears and an important question arises: How do we

provide an effective learning experience to our students at scale? In this paper we explore the use of peer-to-peer interviewing in an ungraded and self-reflective model for pair assessment on complex lab assignments for the Principles of Programming Languages.

We propose a system for measuring effectiveness of education based on student self-reported ability and weight this with their exam performance. We define a method of peer-to-peer reflective interviewing to engage students in a highly scale-able manner that improves student agency in learning. Finally, we report on the benefits suggested by the study for peer-to-peer interviewing compared to TA interviewing found in this experiment.

We explore this value of peer-to-peer interviews by four metrics:

1. What impact does this have on students' completion of the course?
2. What impact does this have on students' ability to correctly assess their own performance?
3. What impact does this have on student performance?
4. What impact does this have on student satisfaction with the course?

2 Background

2.1 Effectiveness

For this paper, effectiveness in learning refers to providing experiences that engage the student and enables high levels of cognition as defined by the Bloom's taxonomy.

Bloom's Taxonomy In this paper we center on the Bloom's taxonomy of learning as the measure of student achievement in mastering the course material. The popular 2001 revision to Bloom's taxonomy suggests a linear progression of cognition from "remember", "understand", "apply", "analyze", "evaluate", and "create" [4]. Here, "remember" is the lowest level of cognition that a student can achieve, in which they know a few seemingly disparate facts. On the other end "create" is the highest level of cognition, in which students can build on all they have learned to form well-reasoned

solutions to complex problems which are novel to the learner. While "create" is rarely the goal of an assignment, it is often a good goal for the course as a whole.

Interview Grading A tool for effective instruction explored at various institutions is to give students an oral assessment of their work called "interview grading". In interview grading, students **evaluate** their mastery of the course material with an oral review of their written assignments. Interview grading has been shown to hold value for students being accountable to their own learning. It works best in a small class setting where the instructors can manage all of the interviews [5] [6]. However, it can be done at scale, by offloading the effort to support staff such as graduate teaching assistants and graders [7]. It is important to note that as proposed, this doesn't continue to scale well as more students means more time for grading by "expert" course staff.

Ungrading Next, we explore ungrading models, by which we move away form a model of grading out of one hundred points and toward a model of "X", "✓-", "✓", "✓+", or some other naming model to represent a distinction from work that is unacceptable (X,✓-) versus "good enough" (✓), or even exceptional (✓+). In various un-grading models such as reflective un-grading, contract grading and standards based grading we move the staff focus away from time obsessing over the difference in grade from an 85% to a 88%, and instead state, that's a "✓". This allows us to instead focus on providing substantive feedback to our students [8] [9] [10]. While this requires constant buy-in from the course staff and students to ensure success across the term as students become co-conspirators in this different educational model, the model has proven effective in many college courses including upper division topics [10] [11][12][13]. This concept can be leveraged effectively in interview grading to emphasis formative feedback over a course grade for the student, helping to move students toward intrinsic learning rewards over extrinsic ones [14].

Reflective Learning In reflective learning, we ask students to have agency in their own education and continuously reflect on what they have learned, what they are struggling with, and how they could potentially apply what they have learned to reach

their own goals. In fact, there is a model of un-grading built around this concept, sometimes called "reflective un-grading" or "big-U Un-gadding" [8]. Here we develop a learning environment where students must author self-reflections and even recommend their own grade for the course. We as course staff might then decide if the students' self-reflection and decided grade is accurate, or how it differs and discuss significant differences with the students. Alternatively, to increase the scale-ability of this model, the course staff can trust the validity of the student assessment and instead analyze the student reports to understand what students are doing well in and use that knowledge to improve future lectures and readings based on the student experience.

2.2 Scale-ability

For this paper, scale-ability in education refers to providing consistent learning opportunities to as many students as possible. Some obvious places to look for scale-able education tools are the use of artificial intelligence in the classroom, and the world of online learning [15] [16] [17]. While Ai in the classroom is promising, it is currently burdensome to implement, so we'll focus more on tools from online learning. What is found to be most important in scaling education online is encouraging collaboration between students in peer-to-peer interactions. After all, more students in the classroom means more students that can interact with other students. Beyond technology integration's, this is the most scale-able resource for the course as enrollment increases. Let us explore two key tools in improving peer-to-peer interactions.

Peer Grading Having students grade each other is considered a must for effective online education at scale. While many students are resistant to peer grading and do not believe it to be as helpful as feedback from their course staff, it has been shown to be effective[15]. This scales infinitely, as more students yields more people to perform the reviews. Perhaps the most important aspect of doing this effectively at scale is to have a way of assessing the students review capabilities. The literature suggests an effective method to ensure effective peer grading is to have some kind of training assignment. Here, students complete an assignment to demonstrate acceptable knowledge

of the peer review process early in the semester [18]. This method has been employed extensively in the online learning environment where scale is potentially limitless.

Discussion Forums Additionally, to increase a sense of belonging and community in a large class - be it online or in person - we see a recommendation for online discussion forums such as Slack, Discord, Piazza, and Zulip [15] [19]. Here many students are able to engage with the material and start discussions with their peers. It is best practice to have course staff monitor and collaborate on this forum as well. While this requires some time from staff to manage the forum, this is often worth the effort for larger sized classes as it engages students on some semi-synchronous forum where they can ask questions and discuss topics beyond the confines of class time.

2.3 The current syllabus

The current course syllabus has seen continued decrease in effectiveness over the years as course enrollment increases - anecdotally. The following assessments are used to construct a course that in practice is shown to be highly effective with seventy students. However, it is struggling to stay effective at one-hundred-fifty students and does not look promising for three-hundred students:

1. Participation: a formative assessment in which students **analyze** information through discussions during class sessions.
2. Labs: a formative assessment in which students **analyze** topics of interest and serves as the basis of student learning. All students complete the same lab in teams of two to three students and use their findings in the assignment to engage class discussions on the related topics. The lab is auto-graded for correctness against a set of pre-defined tests which are partially shared with the students.
3. Grading Interviews: a formative assessment in which students **evaluate** their mastery on the lab material with twelve minute one on one interviews with the course staff in an ungraded X/✓+ style score returned with limited personalized feedback and a score out of one-hundred percent. This interview is graded on the basis of

student's ability to correctly answer the questions in the interview within the time provided.

4. Exams: the summative assessment in which students **create** novel solutions to relevant problems in a timed assessment that is manually graded by the course staff and returned to students with some limited qualitative feedback.

3 Experiment

In this experiment we propose one core change to the course syllabus. Here, interviews are not graded based on the students' correct answers to the interview questions, but instead purely on the students completion of the interview. We emphasize the formative nature of the interview and focus on giving students qualitative feedback on their performance in an ungraded model. We go on to explore two different methods of implementing the interview process, one where students perform the interviews in peer-to-peer interviews with self-reflective components, and the other in which students select themselves to interview with a Teaching Assistant (TA).

Each TA is as member of the course staff with an assured "**analyze**" level of learning on the material. In this course we had eight TAs comprised of two "**create**", three "**evaluate**", and three "**analyze**" level of course mastery.

The course is comprised of six labs which build off the knowledge of the previous lab. In the first lab, students perform both a peer-to-peer interview, then an additional interview with a teaching assistant. By the beginning of the second lab, students choose to either spend the semester in interviews with a TA, or in peer-to-peer interviews.

3.1 Interview Process

In each model of interviewing the interview process contains four phases

1. Training Phase
2. Interview Phase
3. Reflection Phase
4. Action Phase

Training Phase The training phase is required at the beginning of the semester and is reassigned as needed to students throughout the semester to re-commit the student to this interview grading process. In the training phase, students are given a series of videos on mock-interviews with a grading rubric for the interview using an "X, ✓-, ✓, ✓+" grading system for the topics in the interview. Students are asked to grade the interviewee against the rubric and submit their solutions to an automated grading tool which compares the students' proposed grades to the known grade of the mock interview. While this effort had large upfront cost, this sample of the grading process has been shown in other studies to provide great value in reducing overhead throughout the semester by setting clear expectations for students early in the semester [18].

Interview Phase Consider hypothetical students Ranga and Addison have just completed lab three as a student team.

Peer to Peer Interview Ranga and Addison select a time to meet in-person, or over zoom, and discuss what they learned during the lab. They then download the interview question set for the lab and complete the interview together as a team. They are encouraged to complete the interview within thirty minutes; however, this is at the discretion of the student team. If neither student is able to answer the question, they can reach out on the course discussion forum to seek additional information on the topic. Here, we see more peer-to-peer grading interactions, giving students more autonomy in their learning and freeing the course staff to dedicate time to supporting student learning in other ways.

Teaching Assistant Interview Ranga and Addison each sign up for one on one interviews with a member of the course staff. They attend the interview without prior knowledge of the questions that will be asked, and perform the interview in a twenty-minute slot (twelve minutes for lab 1). At the end of the interview, the course staff, tells the student how they performed on each question in an "X, ✓-, ✓, ✓+" scale and work with the student in the time available to discuss plans for improvement as necessary. The course staff also takes time to celebrate what the students have already

mastered and encourage their continued success. The member of the course staff is able to pivot the interview as needed to ask follow-up questions of the student in the Socratic method that encourages the student to create a more comprehensive understanding of the related topics.

Reflection Phase

Student Reflection and Action Planning Regardless of the interview method used, Ranga and Addison now meet to review their performance on the interview and the lab content as a whole. Students are encouraged to spend about thirty minutes on this exercise. They identify their performance on a selection of key skills used in the lab and develop a personal action plan for what they might focus their efforts on in the next lab, taking advantage of the benefits of reflective learning. While the action plan is personal to the individual, students are meant to discuss these plans together to encourage cross-pollination of ideas. Each student submits this via a survey form that allows for the aggregation of student data.

Staff Reflection and Action Plan Next, the course staff review the student performance from their hosted interviews and enter notes about the student performance into a survey form that allows for the aggregation of student data. The course staff then gathers as a whole to review all the data provided both by the TAs and by the students to identify what students are succeeding with, and where they are really missing the mark. Collectively they discuss how this data can inform a change to the course lecture process, using the stores of knowledge that students have today to assist in filling in those gaps as we move on to new topics. Here the staff also has an opportunity to discuss what common issues and successes were observed during the interviews. In practice, this required a two-hour meeting with the full course staff at the end of each lab, after the completion of the interview phase.

Action Phase In the action phase, the course staff executes on their plan for improving the course lectures based on common findings in students' gaps. In an attempt

to increase transparency of the process and build our students as conspirators to the method, the course instruction includes anonymous quotes from the student reflections and openly recognizes why we are covering certain topics in more depth. The students are also encouraged to act on their own action plans and seek whatever assistant or materials they may need. Toward enabling the students' success, the course staff is listening to students and taking note of what roadblocks exist for the students and actively working at removing those roadblocks wherever staff intervention is necessary.

3.2 Enrollment

In this experiment students' self-selected to TA interviews or peer to peer interviews for the semester. The total course enrollment at the beginning of the term was 300, of which 60 selected to interview with a member of the course staff and 240 students selected peer to peer interviews. By the end of the term, 18 students in the TA interview model withdrew from the course and 12 students in the peer-to-peer interview model withdrew.

3.3 Performance

In an attempt to measure the impact of peer-to-peer interviews on student learning, we ask students to reflect on their performance in the course in comparison to the Bloom's taxonomy. Prior to the midterm and final exam, each student is asked to rank their learning on individual course topics and the course as a whole.

Additionally, we normalize the students' summative assessment scores to the Bloom's taxonomy. In review of the final exam content and grade distributions we categorized the student score as Bloom levels as follows (visualized in figure 1):

1. "create": the second and third positive deviation from the higher distribution
2. "evaluate": the first deviation from the higher distribution
3. "analyze": all scores below the first negative deviation of the higher distribution and above the first positive deviation of the lower distribution
4. "apply": the first deviation of the lower distribution

5. "understand": the second negative deviation of lower distribution
6. "remember": the third negative deviation of the lower distribution

Fig. 1. Student grade distribution on the final exam of the course color coded to the assigned Bloom's taxonomy level.

We then weight the student self-reported Bloom level against the normalized summative assessment scores to construct a suggested true Bloom level of cognition for each student.

3.4 Satisfaction

Finally, prior to the midterm and final exam, the students are asked to rank their satisfaction level with the course from "very unhappy", "unhappy", "neutral", "happy" and "very happy".

4 Results

4.1 Case Study

In lab 2 students completed an assignment on authoring an interpreter for a subsection of JavaScript. In the interviews all students are asked about which operators were "overloaded". Among the correct answers, the "+" operator is quickly identified by most students. However, when asked about the expression "'hello' + 2 * 5" and its evaluation, students were not able to arrive at the correct solution. This is identified as

an inability to accurately parse expressions in the language. Accordingly, we adapted lectures during lab 3 to further emphasize visual parsing skills, while talking about the new topic of inference rules. Not surprisingly, during the reflection phase of lab 3, we observed that students had an increased mastery or parsing. In future semesters, we have an action planned to carry this lesson forward, and consider new approaches to lecturing during lab 2 which attempt to resolve this confusion earlier in the semester.

4.2 Enrollment

What impact does peer to peer interviewing have on students' completion of the course? Upon review of the collected data, we found a much higher completion rate of students in peer-to-peer interview grading as shown in figure 2. Observe that only five percent of students in the peer-to-peer model dropped the course while thirty percent of students in the TA interviews dropped the course. While this data can be biased by the students' self-selection to the different models of interviewing, this suggests a positive impact from peer to peer interviewing on the students decision to complete the course.

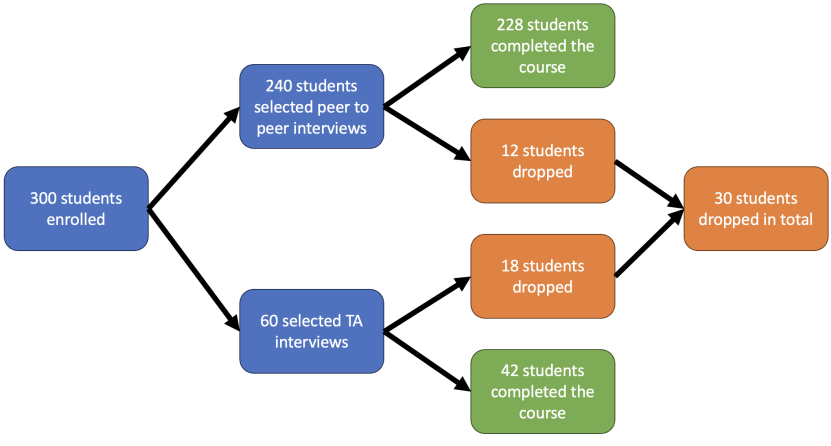

Fig. 2. Visualization of student enrollment, selection to peer to peer interviewing or TA interviewing and the respective course completion rates.

4.3 Performance

What impact does this have on students' ability to correctly assess their own performance? As shown in figure 3, students completing peer-to-peer interviews have more successes along the diagonal matching of Bloom levels with the highest density at "evaluate" and next highest at "analyze". By contrast, the students completing TA interviews have very sparse matches along the diagonal matching of Bloom levels. In further review the student estimation of their own ability from the TA interviews is almost random compared against the true value. This provides clear evidence that the students completing peer to peer interviews have a better understanding of their own cognitive mastery of the material when compared to students completing the TA interviews.

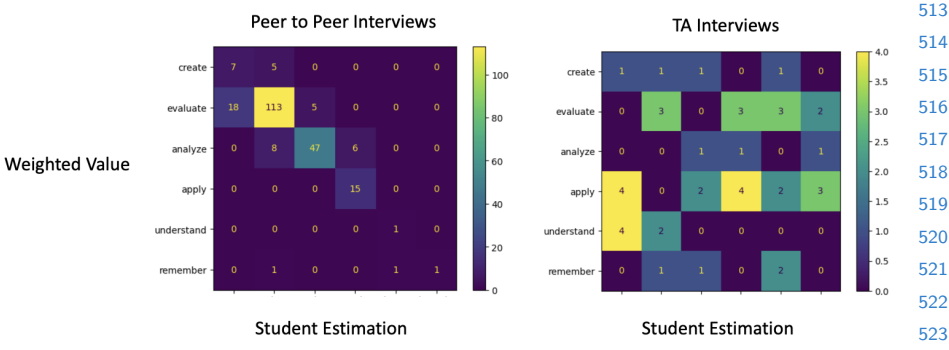

Fig. 3. Confusion matrices of student estimation of performance of Bloom level compared to weighted true Bloom level. Left: Peer to Peer Interviews have low confusion with most density along the diagonal. Right: TA Interviews have high count of false estimations with limited density along the diagonal.

What impact does this have on student performance? As shown in figure 4, we find that students in the peer-to-peer interview have higher overall performance in the course with sixty percent density at "evaluate", while students completing the TA interview have a sixty percent density at "apply". Clearly suggesting a high value for students to complete peer to peer interviews rather than an interviewing with their TA.

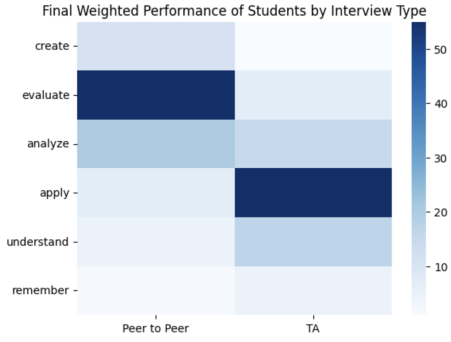

Fig. 4. A comparison of students true Bloom level in completion of peer to peer interviews against TA interviews.

4.4 Satisfaction

What impact does this have on student satisfaction with the course? We found significantly higher satisfaction rates from peer to peer interviewing as visualized in figure 5. Peer to peer interviews have about half of the students happy with the course as a whole, while that number decreases to only twenty-two percent for students in the TA interviews by the time of the final exam.

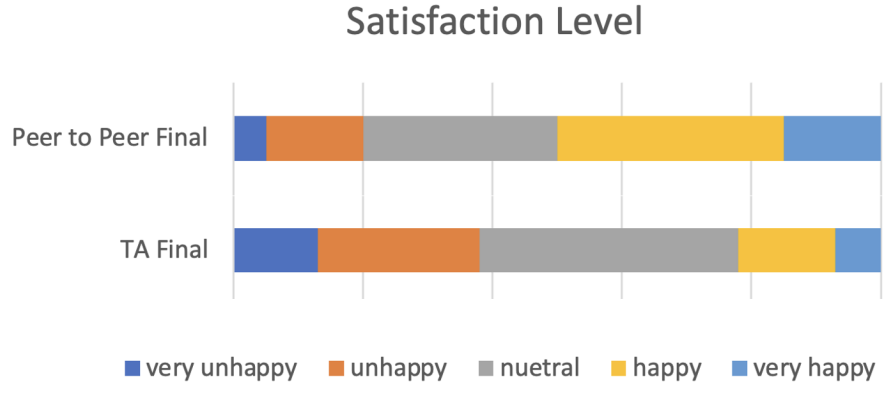

Fig. 5. Comparison of student satisfaction level with the course by the end of the semester.

5 Future Work

This initial case study shows some promise to the value of peer-to-peer interviews, but leaves us with many more question to answer.

Timeline: One important advantage of the peer to peer interviews for the course staff is that the interview data is returned to course staff about one week earlier than it is with TA interviews. Accordingly, if the full class completed peer to peer interviews only, then the course staff would be able to construct their action plan for course improvement based on the student reflections earlier and be able to deliver effective change to the classroom more rapidly. But what impacts would this have on the effectiveness of the review process if TAs had not actually completed an interview with a student and directly observed where students are struggling? Would the staff reflection phase still be as effective?

Inclusion The collected data includes demographic data of students that may embed information about how students from traditionally marginalized and underrepresented communities are impacted by this course change. We are curious to see what information could be inferred from the existing data and consider further changes that better support these students.

Why does it work: The current proposed method of peer to peer interviewing has clear value when implemented correctly, but why exactly does this work well for so many students? What aspects of the student learning environment exist in this modified course structure that could be leveraged in other aspects of the course? How would that change be implemented? When would it not be wise to make such a change?

6 Conclusion

We have demonstrated the value proposition of using peer-to-peer interview grading over TA interview grading. This method leverages students themselves as a scale-able source of effective education actors in the learning environment with promising results

at a time when student to staff interactions continue to decrease as course enrollment grows. This method, as implemented led to higher completion rates for students, a better ability for students to assess their own mastery of the material, higher mastery of the material, and higher overall satisfaction in the course. This model comes with an added benefit that course staff spends less time conducting interviews with students, allowing more time to review student performance and adapt teaching methods to meet the students' needs. While there is more work to be done, we hope that this structure continues to see adaptations that better enable our students' success.

References

1. Zweben, S., Bizot, B.: 2015 taulbee survey continued booming undergraduate cs enrollment; doctoral degree production dips slightly Accessed: 2023-09-20.
2. Zweben, S., Bizot, B.: 2018 taulbee survey undergrad enrollment continues upward; doctoral degree production declines but doctoral enrollment rises Accessed: 2023-09-20.
3. Zweben, S., Bizot, B.: Cra 2022 taulbee survey: Record doctoral degree production; more increases in undergrad enrollment despite increased degree production Accessed: 2023-09-20.
4. Anderson, L.W., Krathwohl, D.R.: A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc. (2001)
5. East, J.P., Schafer, J.B.: In-person grading: An evaluative experiment. SIGCSE Bull. **37**(1) (feb 2005) 378–382
6. Ruehr, F., Orr, G.: Interactive program demonstration as a form of student program assessment. J. Comput. Sci. Coll. **18**(2) (dec 2002) 65–78
7. Grunwald, D., Boese, E., Hoenigman, R., Sayler, A., Stafford, J.: Personalized attention @ scale: Talk isn't cheap, but it's effective. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education. SIGCSE '15, New York, NY, USA, Association for Computing Machinery (2015) 610–615
8. Flaherty, C.: When grading less is more Accessed: 2023-09-15.
9. Stommel, J.: How to ungrade Accessed: 2023-09-15.
10. Owens, K.: A beginner's guide to standards based grading Accessed: 2023-09-06.
11. Chen, L., Grochow, J.A., Layer, R., Levet, M.: Experience report. In: Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1, ACM (jul 2022)
12. Mittell, J.: Rethinking grading: An in-progress experiment Accessed: 2023-09-15.
13. Stommel, J.: Ungrading: A bibliography Accessed: 2023-09-15.
14. Shepard, L.A.: Ambitious teaching and equitable assessment Accessed: 2023-11-12.

15. Martin, F., Ritzhaupt, A., Kumar, S., Budhrani, K.: Award-winning faculty online teaching practices: Course design, assessment and evaluation, and facilitation. The Internet and Higher Education **42** (2019) 34–43
16. Berge, Z.L.: Changing instructor's roles in virtual worlds. Quarterly Review of Distance Education **9**(4) (2008) 407–414
17. Alam, A.: Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: Reforming education in the age of artificial intelligence. In Shaw, R.N., Das, S., Piuri, V., Bianchini, M., eds.: Advanced Computing and Intelligent Technologies, Singapore, Springer Nature Singapore (2022) 395–406
18. Gehringer, E.F.: Electronic peer review and peer grading in computer-science courses. SIGCSE Bull. **33**(1) (feb 2001) 139–143
19. Smith, A.J., Boyer, K.E., Forbes, J., Heckman, S., Mayer-Patel, K.: My digital hand: A tool for scaling up one-to-one peer teaching in support of computer science learning. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. SIGCSE '17, New York, NY, USA, Association for Computing Machinery (2017) 549–554