



Viewpoint

What's Your Placebo?

The dangers of participation bias in educational studies.

ATALE OF WOE: Our institution, like many others, has a high attrition rate in introductory computer science (CS1), our first programming class for majors. We often use the term “DWF rate,” as those students who earn Ds or Fs, or who withdraw from the course are ineligible to continue further in taking other CS classes as part of the major. Beyond the DWF rate, students who earn Cs in their CS1 course, while technically allowed to continue taking CS classes, tend to struggle in those later classes. We do offer a pre-CS1 programming class to help students who are not ready to jump directly into CS1. Many researchers have shown students with prior programming experience tend to do better in the initial CS1 course. We have published research on this,³ as have many others.

We were tasked with developing an assessment our department could use to recommend whether students should sign up for our pre-CS1 course or our CS1 course. We decided to use an instrument that combined two recently validated instruments that measured a student's programming ability. One was a computing concepts inventory¹ and the second was a programming comprehension inventory.⁴

Our university's institutional review board (IRB)-approved study allowed us to have CS1 students optionally take these instruments at the start of the semester. And by agreeing to participate in this study, the students allowed their final grades in the course to be shared by the instructors of the course with us at the end of the



semester. Instructors were to offer extra credit for those students who chose to participate in the study. We note that a student's final grade for a course may well not be the best measure of a student's knowledge about the material learned in the course. However, the department had a problem with its students' DWF rate in CS1 and wanted us to use final grade in the course as a proxy for student mastery of the material in the course, so this is what we did.

In fall 2020, 459 students were enrolled on one of the department's CS1 courses. Of these 459 students, 202 students signed up and completed the assessment instrument that made up

our study. After the semester ended, we conducted the post analysis. And the significant results were ... nothing! It did not seem to matter what students scored in the pretest at the beginning of the semester. All students were just as likely to pass or fail the course. The average grade for students was approximately a B and the median grade was an A-.

As Reich notes, it is possible our assessment instrument does not completely or correctly measure a student's programming ability due to what he calls the “reification fallacy.”⁵ In other words, just because a student scores highly on our multiple-choice instrument does not automatically mean the

student is a strong programmer. Reich also notes that an exam written in English (as ours was) also tests a student's ability to understand English as much as the student's ability to understand the content. Furthermore, our instrument was fixed rather than adaptive. In other words, all students were given the same questions rather than being given more difficult problems when getting an earlier question correct and easier problems when getting an earlier question wrong.

Despite these valid concerns as noted by Reich, we were confident that our instrument was reasonable. Both sub-instruments from which this instrument was constructed had worked previously when used with a similar student audience. Stronger students scored higher on the exam while weaker students scored lower. There had to be something else.

Participation Bias

The one strange piece of information we noticed was that a smaller than ex-

Mean and median letter grade of students who participated and did not participate.

	Mean Grade	Median Grade
Participated (N = 202)	3.1 (B)	3.7 (A-)
Non-Participated (N = 257)	2.3 (C+)	3.0 (B)

It did not seem to matter what students scored in the pretest at the beginning of the semester. All students were just as likely to pass or fail the course.

pected percentage of the students in our study were part of the DWF group. This study was voluntary, and so the obvious question was whether the students who participated were representative of all the students in the class.

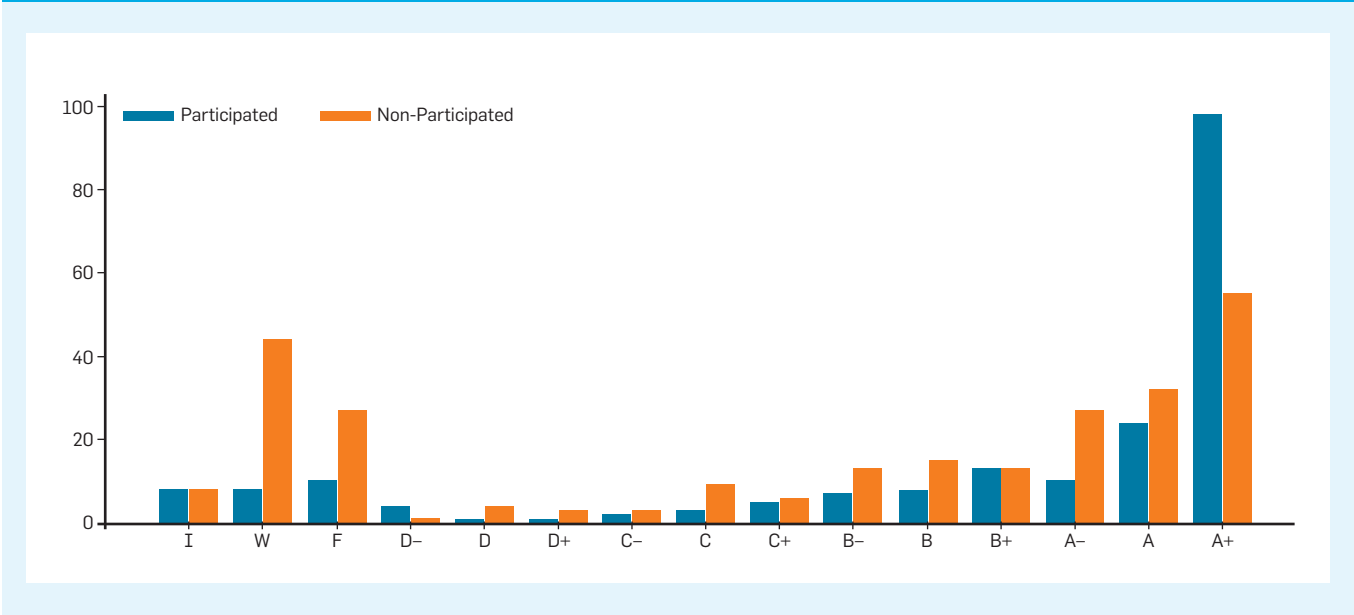
Student grades in the U.S. are protected under the Family Educational Rights and Privacy Act (FERPA), a federal law that protects the privacy of student education records. However, Section 99.31(6)(i) approves of disclosing student data if the “disclosure [of students’ grades without consent] is to organizations conducting studies for, or on behalf of, educational agencies or institutions to: (A) Develop, validate, or administer predictive tests; (B) Administer student aid programs; or (C) Improve instruction.” Because

we were in the process of developing and administering a predictive test, our university’s IRB allowed us to access grades from all students in the course, not just from students who signed the consent form.

After obtaining all students’ final letter grade, we converted them into a number based on the grading scale: A+ = 4.0, A = 4.0, A- = 3.7, B+ = 3.3, ... F = 0. We counted student withdrawals as a 0. Ultimately, we do not know how well those students were performing when they withdrew. While it is likely that the majority of withdrawals are from students who are underperforming and want to withdraw before failing, some withdrawals could be by students who withdraw to focus on other classes even though they are doing fine in this one, or by students who were doing fine but are forced to withdraw due to significant personal or family emergencies. While it could be argued that treating a withdrawal as a failure is thus unfair, the small number of withdrawals we saw does not affect the overall averages regardless how we decided to ultimately treat a withdrawal.

We then organized the students into two different groups. Group 1 consisted of students who participated in our study, and group 2 was students who did not participate. Group 1 had 202 students and group 2 had 257 students. The mean final grade for students who participated was 3.1

Grade distribution of students who participated and did not participate.



points (approximately a B average) and their median final grade was 3.7 points (approximately an A- average). The mean final grade for students who did not participate was 2.3 points (C+) and their median final grade was 3.0 points (B). The resulting differences were statistically significant. Students who volunteered to participate in the study earned on average one entire letter grade higher than students who did not volunteer.

Upon graphing the results (as seen in the accompanying figure), the first thing that can be seen, on the high end, is that there is a significant difference of students who received an A+. Students who participated in the study were almost twice as likely to get an A+ in the course. On the low end, the results are even more striking. Students who did not participate were 5.5 times more likely to withdraw (44 students vs. 8 students) and 2.7 times more likely to fail the course (27 students vs. 10 students). The total DFW rate for students who did not participate was 3.3 times greater than those who did participate (79 students vs. 24 students). Students who did not participate were also 1.8 times more likely to receive a C in the course (18 students vs. 10 students). Clearly, trying to determine whether our assessment could differentiate between students who were ready to take CS1 and students who were not ready to take CS1 makes little sense when primarily testing the students who were ready to take CS1.

Implications

We have read about participation bias in medical studies, but we were unable to find interesting papers about participation bias in computing education research. At least for our failed study, participation bias was a real threat. So, what does this mean? Clearly, some students were not motivated to participate in filling out our surveys. Were these students who did not participate in our study facing other barriers? Students, particularly those who are struggling academically, may be facing a variety of obstacles (for example, financial burdens that cause them to work long hours or caregiving responsibilities that mean they have limited time or bandwidth for school) which could

We feel strongly that more can and should be done in our computing education studies to address and ideally mitigate participation bias.

both explain their lower participation rates in the survey and their higher DFW rate. Because these students did not participate, we do not have any data to use to investigate why these students withdrew or did poorly in the course. Will making participation mandatory fix the participation bias problem we faced, or are some students just as likely not to participate regardless? At a minimum, we would certainly expect more than 44% of students to complete our instrument if it were a required part of the course.


We posit the problem of participation bias is broader than our one study. We regularly attend talks on CS education where the speaker promotes a particular intervention and shows it worked with a group of students who volunteered to participate in the study. At best, the speaker/researcher notes the possibility of participation bias as part of a “threats to validity” section if participation bias is mentioned at all. Likewise, we worry about research studies in, say, helping to achieve CS for all in primary and secondary education. By limiting their participants being studied to students in their interventions (rather than considering the much larger population of students choosing not to participate in CS), it raises doubts concerning the applicability of lessons learned to the non-participants.

Participation Bias in Medicine

The medical community has noted that people who are knowingly being studied behave differently than they would were they not aware they

were being studied. Medicine has responded to this “flavor” of participation bias by setting its gold standard of research studies to be those that are randomized double-blind placebo controlled. Neither the researchers nor the participants know whether the participants are in the control group (those who receive the placebo) or the treatment group.

In computing education studies, it is difficult to have the two groups (treatment and control) truly randomized when studying the effect of some intervention. Because of outside factors, it is much more difficult to randomly place students into control or treatment sections. And double blindness is often impossible to realize when the instructors know which section they are teaching. Math education has addressed the participation bias challenge in observational studies through use of propensity score matching.² And several of the social sciences have found other approaches to limit the impact of participation bias in their research studies.

Despite these limitations with respect to what the medical community does, we feel strongly that more can and should be done in our computing education studies to address and ideally mitigate participation bias. When computing education researchers are preparing the design of their next studies, we beseech them to consider the question: What’s your placebo? 

References

1. Bockmon, R. et al. (Re)Validating cognitive introductory computing instruments. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (2029); 552–557.
2. Graham, S.E. Using propensity scores to reduce selection bias in mathematics education research. *Journal for Research in Mathematics Education* 41, 2 (Feb. 2010), 147–168.
3. Moskal, B. et al. Evaluating the effectiveness of a new instructional approach. In *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education*. (Norfolk, VA, 2004). ACM Press, NY; 75–79.
4. Peteranetz, M.S. and Albano, A.D. Development and evaluation of the Nebraska assessment of computing knowledge. *Frontiers in Computer Science* 2, 11 (Nov. 2020).
5. Reich, J. *Failure to Disrupt. Why Technology Alone Can't Transform Education*. Harvard University Press, Cambridge, MA, 2020.

Ryan Bockmon (ryan.bockmon@unl.edu) is a postdoc in the Raikes School at the University of Nebraska-Lincoln, NE, USA.

Stephen Cooper (scooper22@unl.edu) is a professor in the School of Computing and director of the Raikes School at the University of Nebraska-Lincoln, NE, USA.

Copyright held by authors.