Narrowing Participation and Achievement Gaps in Computer Science through Engagement with Authentic Problems and Real-World Data

Prof. Ada Lerner, Computer Science

Description of the project and pedagogical goals

The field of computer science experiences significant participation, belongingness, and achievement gaps for students from traditionally underrepresented backgrounds. However, research shows that these gaps can be narrowed with specific interventions. In the work of this grant, I propose to implement one such intervention -- the use of realistic and socially relevant examples, data, and questions -- in order to narrow these gaps at one specific point in the computer science department's introductory curriculum, our data structures course (CS 230).

This proposal aims to enhance student learning, participation, and belongingness through the integration of real-world datasets into several learning and assessment experiences throughout CS 230. Specifically, the final project and two programming problem sets will be modified, such that students must use real datasets drawn from governmental, private, and academic sources to answer socially relevant questions. Examples of such datasets and questions might include analyzing questions of equity in housing and employment using census and housing data from cities engaged in open government, or to determine the relationships between advertisers and trackers online using open source data from academic measurements of online privacy. These datasets will be chosen to require and reinforce knowledge and skills related to data structures which students have learned in the class, and to provide realistic examples of the kinds of questions which require data structures to answer.

This innovation is designed specifically to increase performance, participation, and belongingness among groups traditionally underrepresented in STEM and computer science, such as students from racial or ethnic minorities, students with disabilities, first-generation college students, and students from low income backgrounds. As with many pedagogical enhancements, while the largest beneficiaries of the intervention are expected to be these underrepresented groups, it is expected that all students in the course should benefit from the changes. Data throughout the field of computer science and in STEM fields more generally broadly show gaps in performance, engagement, and persistence in the major for students from underrepresented or marginalized backgrounds[1,2,3]. In recent years, the Wellesley CS department has collected data showing that we are not immune to these challenges, with significant achievement, participation, and belongingness gaps in our introductory sequence and throughout the major. This proposal is one step to attempt to address and counteract these issues in order to make our field fully inclusive. Thus, goals for the intervention include a shrinking in the performance gap between majority and minority group students (and a general increase in learning among all students); an increase in feelings of belongingness (both among all students and, particularly, among those from underrepresented and

---

[1] Barker, L. J., McDowell, C., & Kalahar, K. (2009, March). Exploring factors that influence computer science introductory course students to persist in the major. In *ACM SIGCSE Bulletin* (Vol. 41, No. 1, pp. 153-157). ACM.

[2] Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. *Psychological Science*, 21*(8)*, 1051–1057. https://doi.org/10.1177/0956797610377342

[3] Hurtado, S., Newman, C. B., Tran, M. C., & Chang, M. J. (2010). Improving the rate of success for underrepresented racial minorities in STEM fields: Insights from a national project. *New Directions for Institutional Research*, *2010*(148), 5-15.

marginalized groups); and increased belief in the relevance and importance of the work of computer scientists (and of the students as potential computer scientists) to the world.

In order to accomplish these goals, this project proposes to increase the authenticity of the CS230 final project and of several problem sets, since authentic course material and problems have been shown to enhance student learning and engagement[4]. In the context of this proposal, such authenticity will include:

- The use of real-world datasets, and the asking of questions, drawn from settings relevant to the real-world experiences of our society and of our students their communities.
- The use of short reflective exercises as a part of the new problem sets which encourage students to think about the social relevance of the skills they are learning.
- The incorporation of real-world examples of the use of data structures into lecture materials for the course, in order to make clear the relevance and importance of the learning students do about specific and general data structures concepts, analysis, and implementation.

The Context of CS230

CS230 (Data Structures) is the second course in the introductory sequence of the computer science major. It focuses on developing students into mature programmers, including skills related to abstraction, modularity, and object-oriented design. Additional learning objectives for the course include the analysis of the complexity of algorithms and the effective representation of data using abstract data types.

The CS230 Final Project

In its current form, CS230 has a final capstone project component which acts as a coming-together moment which illustrates the extent to which students have developed maturity in software design and development. The project acts as the first large-scale programming project in the curriculum in which students have the opportunity to design a software artifact of significant size and execute on its implementation independently. It is a collaborative group project in which each group works on an individualized topic of their interest. This Mellon Grant proposal aims to retain and enhance those highly valuable aspects of the Final Project. The work of this proposal will emphasize students' ability to work with realistic data, focusing on their ability to answer those questions through their design, specification, and implementation of algorithmic solutions which use data structures. In doing so, a key goal is not to add additional curricular elements to the project: specifically, students in this project will not be asked to independently formulate statistical and data science questions, though they may do so if they wish to push beyond the requirements of the project. Overall, the redesign will retain the basic structure of the project while engaging students with data and questions designed to connect with their motivations and which naturally exercise the full power of the data structures they have learned to analyze and use throughout the semester.

In the project, students will be required to work with at-scale, complex data which cannot be effectively handled or reasonably analyzed without the mature application of data structures. Data

[4] E.g., Allen, Deborah E., Barbara J. Duch, and Susan E. Groh. "The power of problem-based learning in teaching introductory science courses." New directions for teaching and learning 1996.68 (1996): 43-52, Smith, Karl A., et al. "Pedagogies of engagement: Classroom-based practices." Journal of engineering education 94.1 (2005): 87-101.

structures have two major purposes which are studied in CS230: (1) providing human-friendly abstractions for reasoning about data and writing algorithms which process that data, and (2) enabling the modular implementation of efficient algorithms which rely on the properties of the data structures used. In order to make students' engagement with these two aspects of data structures concrete, the final project will be designed in order to require the use of data naturally and powerfully represented by the data structures they have studied (e.g., trees, graphs, hash tables, etc.), and students will be required to use data which is large enough in scale that efficiency concerns become critical. To emphasize these aspects of the project, students will be provided with curated a) datasets and b) questions about that data, so that they can spend their effort analyzing and executing on the use of data structures to answer the questions. Questions provided for the students to answer may include, for example, reproductions of past results on a dataset or the application of questions asked in one context to a new one.

New Problem Sets

The CS230 final project serves as a capstone to the work that students have done throughout their semester studying data structures. To reach the level of skill where students can achieve the objectives of this capstone project requires scaffolding them to develop skills and independence. To do so, this proposal includes the development of two new problem sets, to be used during the semester, which will give students more guided experiences to develop toward the learning objectives assessed by the final project. These new problem sets will, like all CS230 problem sets, involve smaller, more closely guided (but still significantly independent) software design and programming work. Students will be asked to work with specific, at-scale datasets using the data structures they have studied to date.

Additionally, these problem sets will include miniature reflective components in which students reflect on the social relevance of the datasets they are working with, the problems they are investigating, and the skills they are developing. This type of reflection on the relevance of course material has been shown to have a significant ability to reduce achievement gaps, especially for underrepresented minority and/or first generation students[5].

Selection of Datasets, Questions, and The Need for Curation

Datasets will be chosen from government, the private sector, and academic sources. Examples could include census data, transit service data, electricity grid data, wildfire data, weather data, economic data, and climate data; data scraped from private websites with APIs such as Yelp, Google Maps, IMDB, etc.; and data from published papers such as internet measurements, privacy and tracking measurement, genomic data, and data from other fields of science. Data with impact on human communities, such as climate data, census data, and privacy data will be emphasized through the selection of datasets and the questions asked about them, which may include questions related to equity and democracy, environmental factors, and economic factors in order to particularly enhance students' interest in the data and feelings of importance of the work. Wherever possible, data will be as global as possible to include international students' interests.

Cleaning data and inventing scientific questions are complex skills which are not covered in this class. As a consequence, to avoid adding additional learning objectives to the project, students will not be required to invent de novo scientific questions -- suites of interesting questions, such as

[5] Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745.

questions asked of similar data in different settings, will be provided to the students as scaffolding. Ambitious groups are welcome to choose their own questions or even datasets, with clear communication that this goes above and beyond and may be more challenging. Additionally, since realistic datasets are messy, data cleaning is challenging, a major part of the work of this proposal is to mitigate such challenges by curating and preparing the data and tools which students will be using. Prof. Lerner and a student assistant will undertake this project, in order to provide pre-cleaned data for students to analyze. By doing so, this proposal aims to increase learning, engagement, and identity-development through the use of realistic data while holding constant the overall challenge level of the work similar, keeping it similar to the challenge level for CS230 in the past and not requiring new skills of scientific inquiry or complex data cleaning.

What is the research evidence supporting this teaching innovation?

Researchers have proposed and studied several models of student motivation, belongingness and persistence in STEM in general and computer science in particular, and have validated those models in a variety of experiments and studies. For example, "Communal goal affirmation theory" suggests that goals such as working with others or working for communal, professional, or social purposes are often motivating for students from underrepresented groups, and that the use of assignments which are directly relevant to such goals can be beneficial to the learning and engagement of those students[6,7]. Similarly, for example, research under the framework of "Utility value theory" has shown that increased personal relevance of course materials and reflective exercises about the relevance of such course materials can decrease performance gaps, with a treatment size effect in this case of 0.51 grade points.[8] This research motivates the inclusion of short reflective exercises as part of the new problem sets, in order to emphasize the authenticity of the course material as much as possible without significantly increasing student workload. The goal of not increasing overall student workload is key since research in computer science pedagogy in particular has shown that student persistence in the major is significantly correlated to their perception of whether the workload of introductory classes is fair and achievable[9].

How will I know if students have achieved the key learning outcomes/objectives?

- Grade data from past semesters will be compared with data from this semester.
- Achievement gaps in grade data will be analyzed for past semesters and for this semester.
- Qualitative and quantitative data on student engagement and beliefs about the value, relevance, and realism of the work and skills of the class will be collected at various points in the semester.

---

[6]Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. *Psychological Science*, 21*(8)*, 1051–1057. https://doi.org/10.1177/0956797610377342

[7]Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., & Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to stem careers: Evidence for a goal congruity perspective. *Journal l of Personality and Social Psychology*, 101*(5)*, 902–918. https://doi.org/10.1037/ a0025199

[8] Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745.

[9] Barker, L. J., McDowell, C., & Kalahar, K. (2009, March). Exploring factors that influence computer science introductory course students to persist in the major. In *ACM SIGCSE Bulletin* (Vol. 41, No. 1, pp. 153-157). ACM.

    - When possible, qualitative data collected will use validated measures of student identity, belongingness, self-efficacy, and other measures.[10,11,12]
    - Similar data may be collected from students who took CS230 in the past, for longitudinal comparison.
- Quantitative data related to the workload and amount of time spent on the course will be collected and analyzed with respect to demographics of the students.

Basic information about the project

This is an enhancement for an existing course, CS230 Data Structures, to be implemented in Spring of 2019. CS230 is the second course in the Computer Science introductory curriculum. The proposal involves the introduction of two new programming problem sets, a new design for the final project, and the development of small, modular lecture units to be inserted into existing lectures. Three sections of CS230 will be taught (two by Prof. Lerner, one by Prof. Orit Shaer), and the course is closely integrated with its lab (taught by Senior Lab Instructors Stella Kakavouli and Jean Herbst). Each lecture section of the course has a cap of 20 students, suggesting a total enrollment of 60 students. A key goal of this proposal is to ensure that the materials and innovations designed as part of this grant, including student learning activities, assessments, and evaluation strategies, benefit future iterations of the course,. This will enable the work funded by this proposal to be used going forward to continue to study and enhance student performance and reduce the size of the gaps in the introductory sequence of the major.

Availability to develop the project

Prof. Lerner plans to work full time during two weeks of Wintersession and one week of Winter Break on the design of the project, the problem sets, and the creation of lecture materials which can be integrated throughout the semester. They plan to work with Sarah Pociask during Wintersession, and to meet with co-instructors for the class during January in order to discuss and integrate the new material.

Prof. Lerner will have particular availability to work on these enhancements to the course during the semester since they are teaching two sections of CS230, reducing the preparation time they require for teaching. This additional time will be used to conduct data collection related to evaluating the efficacy of the project, and to cover additional responsibilities during the semester in order to ensure that the project and new assignments run smoothly. For example, they plan on visiting labs, holding extra office hours, providing supplemental tutor training, and providing additional resources to students throughout the course to get them on track with the new assignments without overly burdening their co-instructors.

---

[10] Trujillo, G., & Tanner, K. D. (2014). Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity. *CBE—Life Sciences Education*, *13*(1), 6-15.

[11] Hazari, Z., Sadler, P. M., & Sonnert, G. (2013). The science identity of college students: Exploring the intersection of gender, race, and ethnicity. *Journal of College Science Teaching*, *42*(5), 82-91.

[12] Williams, M. M., & George-Jackson, C. (2014). Using and doing science: Gender, self-efficacy, and science identity of undergraduate students in STEM. *Journal of Women and Minorities in Science and Engineering*, *20*(2).