Traditional vs. Flexible Modalities in a Data Structures Class

Author Name 1
author1email
Author 1 institution
Author 1 city, Author 1 state, Author 1 country

Author Name 2
author2email
Author 2 institution
Author 2 city, Author 2 state, Author 2 country

Author Name 3
author3email
Author 3 institution
Author 3 city, Author 3 state, Author 3 country

ABSTRACT

This experience report presents results from a quasi-experiment comparing course performance and student-reported survey constructs between two groups of students. One group took a Data Structures course with traditional, in-person modality. The second group took the same course with flexible, online modality. The work was motivated by the rapid adjustments computer science instructors made due to remote learning during the Covid-19 pandemic. In a response to these forced changes, this study was set up to investigate the differences between pre- and post- pandemic modalities. There are 212 students in the study, which took place in Fall 2021 at an R1, minority serving institution in the Midwestern United States. The study found that students in both groups performed similarly on common course components like projects, labs, homework, and a final Data Structures assessment. There were not significant differences in their self-reported ease of learning, enjoyment, belongingness, attitude, mindset, and self-efficacy. However, when taking into consideration gender, we found that women's performance was lower than men's in the traditional modality course. Women's performance in the flexible modality course was on par with men. Lastly, we present some feedback from students relating to assessments and modality.

CCS CONCEPTS

• **Applied computing → Interactive learning environments**.

KEYWORDS

modality, data structures, assessments

ACM Reference Format:
Author Name 1, Author Name 2, and Author Name 3. 2024. Traditional vs. Flexible Modalities in a Data Structures Class. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION AND INTRODUCTION

At our institution, our traditional, pre-pandemic modality was in-person with rewarded but not required attendance to lectures and mostly required attendance to labs. Exams were in-person, high stakes, and proctored. Throughout the pandemic, our classes moved to flexible, online, and asynchronous modality. Many of us learned that it was nearly impossible for our online exams to resemble anything like the exams we were used to [2]. Proctoring online high stakes exams increased student anxiety [25] and unproctored online high stakes exams were riddled with academic misconduct [7]. Courses using flexible, online modality had a variety approaches to assessment including specifications grading [6, 9], eliminating deadlines [22], peer and self assessments [24], low stakes assessments [1], oral assessments [15], and mastery based assessments [13].

In this experience report, we present results from a quasi-experiment conducted during Fall of 2021 in a Data Structures course. The goal was to formally investigate the differences between our new and old modalities. The research theme for this work is to *study the differences, if any, that exist in students who take a Data Structures course in our traditional versus flexible modalities.*

The research questions are:

(1) What differences, if any, exist between students who take a Data Structures course in traditional and flexible modalities when comparing their performance on graded course components (homework, labs, projects) and a standardized Data Structure Concept Inventory assessment?
(2) What differences, if any, exist between students who take a Data Structures course in traditional and flexible modalities when comparing their self-reported responses about mindset, belongingness, attitude about CS, and self-efficacy?
(3) What happens if we take into consideration gender?

2 BACKGROUND

Many computer science faculty have been experimenting with modality. The Covid-19 pandemic has made educators everywhere rethink their approach to their courses. The traditional and flexible modalities vary in a few ways: lecture modality, lab modality, office hours modality, and assessments. The modality of lecture, lab, and office hours is straight forward in both the traditional and flexible modalities. The choice of assessment is more complex, so we will provide a summary of the approaches to assessment used in this study.

One approach is oral assessments. Oral exam are usually live, but can also be submitted via a student recording. Oral exams are primarily spoken and often accompanied by the student writing or typing out their thought process in a shared way. A post-pandemic study [19] used oral exams as a replacement of traditional, written exams in an introductory computer science course. Oral exams are particularly useful in remove learning and have been found by some to be a more accurate method to assess students' learning [5]. For a small class of about 50 students, Ohmann found that instructors spent about the same amount of times on oral exams as they would

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

on written exams, however, students spent more times preparing for oral exams compared to written exams [12]. In a long-term twin study, Rimfeld et al. found that teachers' assessment was able to predict exam performance after compulsory education finished [16].

Another approach is asynchronous assessments. Unlike synchronous assessments, which are done in a proctored environment, asynchronous assessments are completed by students in the environment of their choice. They can choose when to start the assessment within a specified time frame. A study by [20]found that students performed on average only 3% better on an asynchronous exam compared to students taking the same exam synchronously [20]. The score advantage was even less when using randomized questions.

Another approach is mastery based assessments. This is where students are given multiple attempts on an assessment. Therefore, the assessment must be graded and the grade returned in between attempts. A review by Garner et al. of a mastery based approach in CS found that the research in this area was sparse and not conclusive [4]. For large classes, mastery based exams require assessments to be autograded and in most cases questions need to be pooled, randomized, or something similar so subsequent attempts are not trivial. Rusak and Yan developed an assessment framework that generates unique and identifiable exams for each student in a probability course [18]. Fully mastery based courses in CS have not always been shown as successful. The study of a fully-fledged mastery model in a first-year CS course resulted in decreased task completion and long periods of inactivity [13]. A common issue with this work is that researchers often are not able to include students who fail out of the course. Additionally, many experiences in using a mastery-based approach report small class sizes so results may not scale to larger classes [21].

Lastly, we will discuss the concept of low stakes versus high stakes. We define high stakes components as those worth a significant portion of the grade. If a student were to do poorly or miss a high stakes component, their grade would be greatly impacted. We define low stakes components as the opposite.

3 STUDY DESIGN

This work is a quasi-experiment that took place at an R1, minority serving university in the Midwest during Fall 2021. The control group was made up of students who were enrolled in our traditional modality version of the course. In this group, lectures were in-person and recorded, attendance was rewarded but not required. Labs were in-person, attendance was required. Office hours were both in-person and online. Assessments were in-person, high stakes, and proctored.

The intervention group was made up of students who were enrolled in our flexible modality version of the course. In the group, lectures were online, asynchronous videos plus an optional weekly synchronous session. Labs were required and offered both in-person and online (students could choose). Office hours were both in-person and online. Assessments were online, low stakes, asynchronous, and included both oral and mastery-based approaches.

The study presented here was approved by the Institutional Review Board. All participants of this study consented to their grades and survey responses being used for research purposes.

3.1 Overall course

The course component grade weights were the same for both groups. The grade weights for all students were: participation (5%), homework (10%), labs (15%), projects (35%), midterm assessments (30%), final survey and assessment (5%). Both groups completed the same curriculum (homework, labs, projects) and followed the same deadline schedule. Both groups used the same textbook, IDE, and submission system. Both groups also shared the same TA staff, which was made up of about twenty undergraduate and graduate TAs.

3.2 Lectures

The lecture sections were taught by two different instructors. Since gender is a topic of this paper, we will point out that the control group was taught by a male professor and the intervention group was taught by a female professor. The lecture schedule was the same for both groups, however, each instructor's delivery was unique. Both groups had access to all lecture materials and recordings (control group could access intervention group's lectures and vice versa).

3.3 Midterm assessments

The component called midterm assessments differed between the two groups. The control group's exams were in the form of three in-person, handwritten, closed notes, proctored exams. Each exam was weighted 10%. The exams were comprised of short answer, concept questions, as well as short coding questions. The exams were graded by the instructor. The intervention group's exams were in two categories: mastery quizzes and oral exams. Students completed seven mastery quizzes, due every other week. These quizzes were online and students were scored best out of three attempts. Questions came from a large pool, so each attempt included different multiple choice questions. Quizzes were open notes, unproctored, and students had 48 hours to complete. Students also completed three oral exams. Each oral exam was a different format. The first was solving a problem, for which they could prep in advance, live in front of a small group of students and a TA (10 minutes). The second was solving a problem they were given live in front of a small group of students and a TA (20 minutes). The third was solving a problem they were given live in front of only a TA (30 minutes). Grades for oral exams came from autograder functionality tests combined with rubrics filled out by their peers and TAs. Rubric items included time management, communication, clarity of presentation, testing code, incremental development, and handling errors.

3.4 Final assessments

Students in both the control and intervention group took a final survey and assessment. This took place during final exam week and was a required course component worth 5% of the final grade. Students answered questions from a validated survey instrument relating to self-efficacy, mindset, attitudes [17], as well as, belongingness [23]. The final assessment used questions from a validated

	Control	Intervention	All
Total	122	90	212
Male/Man	83	61	144
	68.0%	67.7%	67.9%
Female/Woman	21	21	42
	17.2%	23.3%	19.8%

Table 1: Demographics of participants. The sum of each demographic item does not add up to the totals listed due to some subgroups being too small to include, student did not fill out the survey, or student responded something like "other", "prefer not respond", etc.

concept inventory of basic data structures [14], which were all multiple choice. There were also questions created by the instructors. These questions were a combination of short answer, hot spot, fill in multiple blanks, and numeric. The questions were pooled by topic, so each student got slightly different questions. All questions were autograded. Students completed this final survey and assessment in an online, synchronous, unproctored format. Only one attempt was permitted.

4 STUDY POPULATION

The highest enrollment for the class was recorded at the start of week 3 (the end of our add/drop period). At that time there were a total of 307 students enrolled, 172 in the control group's section and 135 in the intervention group's section. By the end of the term, 255 students were enrolled in the course, 113 from the intervention group's section, and 142 from the control group's section. We had 235 students complete the final assessment and survey (a required course component). There are 212 students included in the research study based on those who consented during Week 3 combined with those who either filled out the survey or received a final grade in the course. The demographics of the participants can be found in Table 1.

5 METHODS

Survey results, assessment results, and grades were analyzed. Comparison between various groups were made, for example comparing the control group versus the intervention group. First, a Jarque-Bera test [8] was run to determine if the data being tested comes from a normal distribution. If the data is normal, a two-sample t-test is run on the data in the two groups. If that data is not normal, a two-sided Wilcoxon rank sum test is run on the data in the two groups. We report p-values and the mean of each group, where significant differences are marked *** ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$)

6 RESULTS

6.1 RQ 1: Course Performance

The students completed 13 homework assignments, 11 labs, and 7 projects. We averaged each grade component for each student and then we compared the average homework, lab, and project scores between the control and intervention groups. Table 2 summarizes those results. The performance on labs and projects was similar between the two groups. The students in the intervention group performed better on homework than the students in the control group. Both groups performed similarly on the final assessment, which was a standardized Data Structure Concept Inventory.

	Control	Intervention	p-value
Projects	67.64	65.54	0.6198
Labs	91.81	93.34	0.6584
HWs	93.89	97.22	0.0225*
Final Assessment	72.00	70.68	0.8651
Midterm Assessments^{o}	54.15	77.28	0.0000***

Table 2: Mean scores for control group vs intervention group, where significant differences are marked * ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$). ^{o} The midterm assessments in each group were different.**

It is interesting to look at the distribution of the final assessment scores for the control group versus the intervention group. Figure 1 shows how the two groups performed broken down into A, B, C, D, F, and zero performance. Grade boundaries for A are >=90, Bs are >=80 and < 90, Cs are >=70 and <80, Ds are >=60 and <70, Fs are >0. and < 60, Zeros include only scores of 0. This distribution is similar on the average project scores. There are a similar percentage of A performers in both groups, a higher percentage of B performers in the intervention group, a much higher percentage of C performers in the control group, a similar percentage of D performers in both groups, and a much higher percentage of F performers in the intervention group.

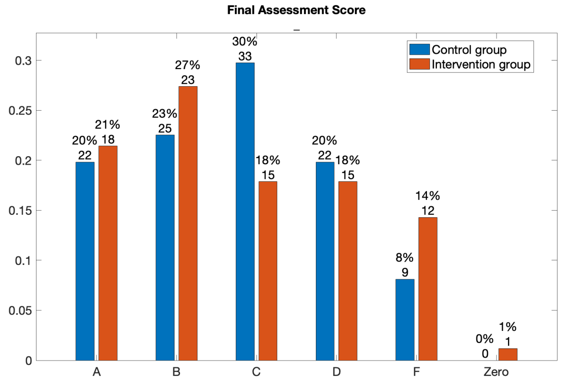

Figure 1: Final Assessment score distribution between groups. Y-axis is plotted using the percentage of students who received each grade (total number of students also included).

Lastly, it is important to make a few comments about the midterm assessment performance between the two groups. The mean score for the control group was 54.15% and the mean score for the intervention group was 77.28%. This illustrates the striking difference between the assessments. Looking at Figure 2, the more complete distribution of the two can be seen. These results are the raw scores, not curved. Actual end of term grades for students were curved to account for the differences between the groups midterm assessment scores.

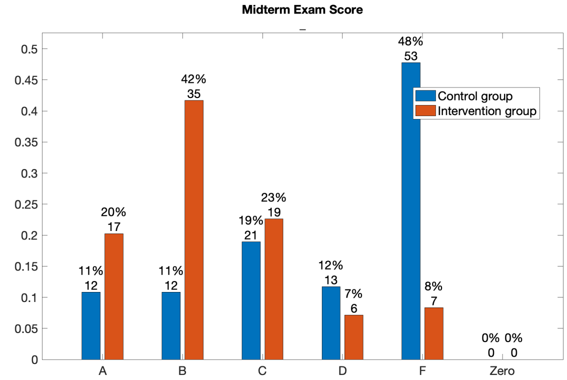

Figure 2: Midterm assessment score distribution between groups.

6.2 RQ 2: Attitudinal Constructs

Our survey questions can be broken down into four factors including: belongingness, mindset, attitude, and self-efficacy. There were no statistical differences in the two groups' responses to these questions. The average enjoyment of the course was 3.56/5.00 (Control) and 3.58/5.00 (Intervention), 5 is "most enjoyed". The average rating on ease of learning was 2.11/4.00 (Control) and 2.13/4.00 (Intervention) where 4 is "found most easy". We didn't find any significant differences between the control and interventions groups for any of questions relating to belongingness, mindset, attitude, and self-efficacy. However, when controlling for gender, we did see differences which are presented in the next section.

6.3 RQ 3: Considering Gender

When we take into consideration gender, we found differences in male students/men and female students/women. We did not have enough students who identified as any other gender to include other genders in this part of the analysis.

There were differences in course performance between men and women. Women in the intervention group scored better than women in the control group in every course component, however, none of those differences were significant. Women and men performed more similarly in the intervention group, whereas women and men performed quite differently in the control group. Most concerning, women in the control group scored significantly lower than men in the control group on both the midterm assessments and the final assessment. Both of these differences were significant. Women in the control group also scored more than 10% lower on their projects than the men in the control group, however, that was not significant. This is interesting because women in the intervention group scored about 3% higher than men in the intervention group. The details of these comparisons can be found in Table 3.

	Cont-M	Cont-W	Inter-M	Inter-W
Projects	72.50	61.41	67.93	70.83
Labs	93.33	94.98	95.37	95.80
Homework	93.79	99.65	99.68	99.67
Final	78.25	70.37^{*}	76.75	71.62
Midterms	61.31	41.51^{**}	81.19	76.76

Table 3: Mean scores for control group vs intervention group by Gender, where significant differences are marked *** ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$) **for comparison of men versus women in each the control and intervention groups.**

	Cont-M	Cont-W	Inter-M	Inter-W
Self-efficacy				
Q1	4.41	4.00^{**}	4.34	4.43
Q2	3.98	3.57	3.90	3.90
Q3	4.29	3.86^{**}	4.30	4.29
Q4	4.29	3.76^{*}	4.20	4.38
Q5	4.08	3.86	4.00	4.24
Q6	4.31	4.05	4.15	4.38

Table 4: Self-efficacy responses for control group vs intervention group by Gender, where significant differences are marked *** ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$) **for comparison of men versus women in each the control and intervention groups.**

The first factor was self-efficacy, which means it had the strongest variability in our sample. The results for self-efficacy are summarized in Table 4. The statistics show that for all six question on self-efficacy, men in the control group reported a stronger sense of self-efficacy than women in the control group. On three of the questions, those differences were significant. On the other hand, the women in the intervention group reported a stronger sense or similar sense of self efficacy than the men in the intervention group. None of those differences were significant.

The second factor was mindset, which means it has the second strongest variability in our sample. However, there was not consistent or significant differences across genders in the responses to questions relating the mindset.

The third factor was attitude. For all eight question related to attitude, the men in the control group reported a better attitude towards CS than the women in the control group. Two out of those eight differences were significant. Similarly, the women in the intervention group reported a better attitude towards CS than the men in the intervention group. None of those eight differences were significant. One example of a significant difference is how students responded the following attitude statement "I think CS is interesting". A response 5 means strongly agree. The mean response of men in the control group is 4.42/5.00 and the mean response of the women in the control group 3.81/5.00. That difference is significant with a p value 0.0023. The mean response of men in the intervention group is 4.30/5.00 and the mean response of the women in the control group 4.67/5.00. That difference is not significant. Additionally, the difference between the women in the control group versus the women in the intervention group is also significant with a p-value of 0.0006. For conciseness of the paper, more details on attitude are not provided.

	Cont-M	Cont-W	Inter-M	Inter-W
Enjoyment	3.69	3.00*	3.69	3.29
Ease of Learning	2.24	1.62*	2.21	1.81*
Interested in CS	4.30	4.14	4.43	4.57
More CS soon	4.58	4.24	4.61	4.71
More CS later	4.59	4.38	4.56	4.81

Table 5: Survey responses for control group vs intervention group by Gender, where significant differences are marked *** ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$) **for comparison of men versus women in each the control and intervention groups.**

The fourth factor was belongingness. There were five questions relating to belongingness. Only one question showed a significant difference. The question was "When something bad happens, I feel maybe I don’t belong in CS", where 4 means strongly agree. The mean response from men in the control group was 2.45/4.00 and the mean response from women in the control group was 2.90/4.00, which is a significant different with a p-value of 0.03. This is an important result considering low midterm assessment scores in the control group.

Finally, Table 5 shows some other relevant responses by gender. Women in the control group reported enjoying the course the least and significantly less than the men in the control group. In both the control and intervention groups, men reported finding CS easier to learn than women. In both cases, the differences were significant and the difference was larger in the control group than in the intervention group. Students were asked if they were interested in a CS major/minor, if they want to take more CS classes next year, and if they want to take more CS classes in the future. In all three questions, men in the control group were more interested in CS than women in the control group and women in the intervention group were more interested in CS than men in the intervention group. However, none of the differences were significant.

6.4 Student Preferences

We asked students about their preferences of different assessment strategies. Figure 3 summarizes their responses broken down by group. Interestingly, each group, on average, preferred the type of assessment they had. Most of the means are close to the middle but slightly favor the intervention group assessment style. The strongest preference among all students is for mastery-based and open notes assessments. Secondary preference appears to be low stakes and typed.

7 LIMITATIONS AND THREATS TO VALIDITY

This is a quasi-experiment where students were self-selecting into these sections. Some students might have selected the section because they preferred the modality or assessment method. As a result, we asked students why they chose to register for their section. Their responses are summarized in Figures 4 and 5. The top three reasons students in the control group selected their section were socially motivated, struggled with self-motivation online, and accountability. The top four reasons students in the intervention group selected their section included flexibility, prefer online to rewind and rewatch, minimized my commute time, and health/safety reasons. No students selected that they picked their section due to the assessment type (except for 10 students had preferences relating to oral assessments).

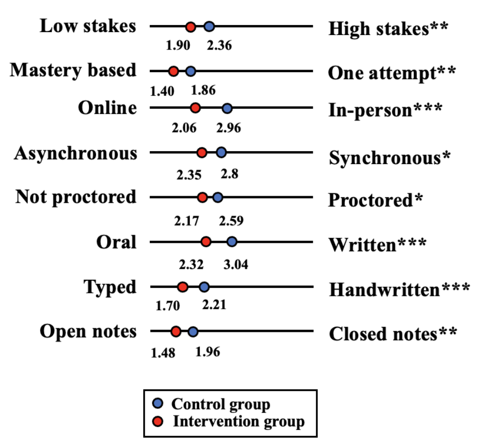

Figure 3: Mean assessment style preference for control group vs intervention group, where significant differences are marked *** ($p < 0.001$),** ($p < 0.01$), * ($p < 0.05$)

.

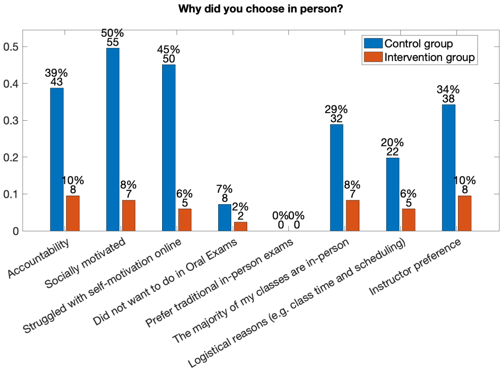

Figure 4: Reasons students picked in-person.

Since students self-selected into the section they were enrolled in, we know that the two groups are not random. Therefore, we asked students in the final survey some questions to help us better understand our two groups. There are many uncontrollable factors about each group. In both groups, students are able to experience some aspects of the course in an online or in person way (e.g. lectures, office hours, etc.). Figures 6 and 7 summarize how the students experienced the course. The students in the intervention group had more of an online experience, and the students in the control group had more of an in person experience, on average. We also asked students more details on what lecture content they were consuming. The results show that, on average, the students in the control group watched about 25% asynchronous lecture videos that were part of the intervention groups’ section and attended, on average, about 61% of their section’s lectures in person. On the

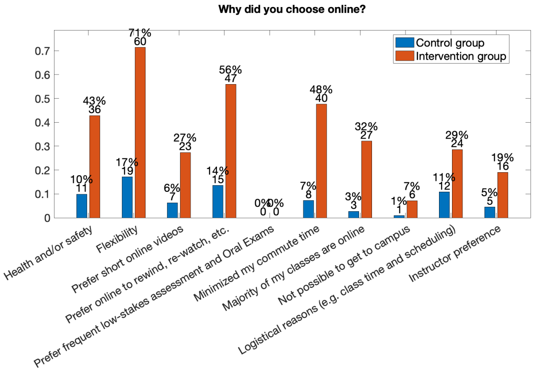

Figure 5: Reasons students picked online.

other hand, the students in the intervention group watched, on average, about 7% of the in-person lecture recordings that were a part of the control groups' section and watched, on average, 80% of their own asynchronous lecture videos.

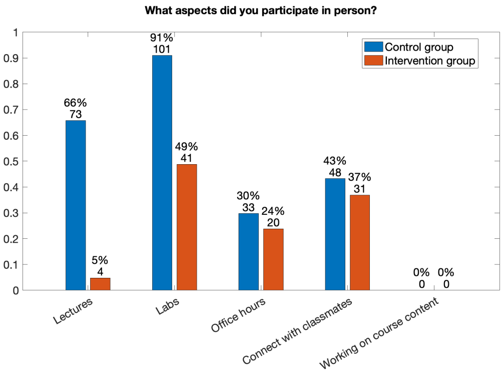

Figure 6: In-person participation broken down.

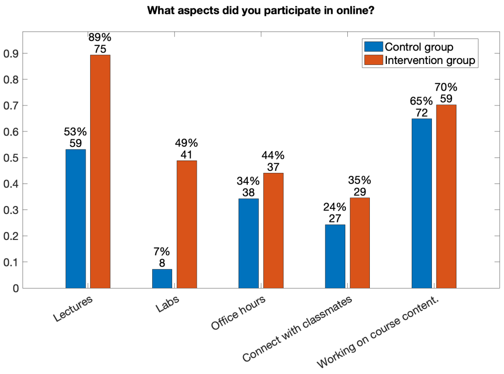

Figure 7: Online participation broken down.

There was selection bias in who consented to be included in this work. The distribution of the consent data had some differences from the actual data. Other differences between the groups include: different instructors, different genders of the instructors, different section timing, different section sizes, different lab instructors (all TAs), and more. Based on Figures 4 and 5, we know that a small percentage of students in both the control and intervention groups chose their section based on instructor preference.

8 DISCUSSION

We have found that teaching Data Structures using a flexible, online modality appears to be as effective as teaching it using a traditional, in-person modality. Subsequently, we can say that we do not have any evidence that teaching our flexible, online Data Structures is any less effective than our traditional, in-person Data Structures course. We think this is the primary conclusion to draw from this paper and an important one. We believe that the uncontrolled variables across the two groups like instructor, gender of instructor, lab instructor, and session timing all strengthen this result.

We think that in courses like Data Structures, which are largely project based, the modality may not be the most important aspect of the course. There are successful variations of Data Structures that are primarily project-based [10]. Since assessment design is tied closely to modality, we do not see any negative affect of students taking the Data Structures course with online, low stakes, un-proctored, mastery-based and/or oral assessments. Therefore, it may be beneficial for CS educators to focus their time and effort (and grade weights) to course components other than assessment.

A less consequential result of this study relates to the experiences of women in the Data Structures course. We have found that there are differences in the experiences of women and men when comparing modalities. We have two potential explanations for this effect on women. The first is that since the intervention group was online, the women in that group were not exposed to as much "bro" culture [11]. The flexible modality allows students to choose how they experience the course, which could include opting out of any potentially, negative experiences. The second is that the instructor of the intervention group was a woman and the instructor of the control group was a man [3]. We cannot conclude that women will be more successful in flexible modality courses. However, we would like end the paper with some open questions to CS educators. How important is the effect of a female instructor on female students in CS? With the limited number of female CS educators, could flexible modality help us give more female students access to female CS instructors? Could this extend to any other underrepresented groups?

ACKNOWLEDGMENTS

The authors would like to thank the students and teaching assistants involved in the Data Structures course during the Fall of 2021. They would also like to thank the Department of Computer Science for their support in running the classroom experiment. This research was supported by the xxxxx.

REFERENCES

[1] Matthew D Casselman. 2021. Transitioning from High-Stakes to Low-Stakes Assessment for Online Courses. In *Advances in Online Chemistry Education*. ACS Publications, 21–34.

[2] Ted M Clark, Christopher S Callam, Noel M Paul, Matthew W Stoltzfus, and Daniel Turner. 2020. Testing in the time of COVID-19: A sudden transition to unproctored online exams. *Journal of chemical education* 97, 9 (2020), 3413–3417.

[3] Tara C Dennehy and Nilanjana Dasgupta. 2017. Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences* 114, 23 (2017), 5964–5969.

[4] James Garner, Paul Denny, and Andrew Luxton-Reilly. 2019. Mastery learning in computer science education. In *Proceedings of the Twenty-First Australasian Computing Education Conference*. 37–46.

[5] Hasmik Gharibyan. 2005. Assessing students' knowledge: oral exams vs. written tests. *ACM SIGCSE Bulletin* 37, 3 (2005), 143–147.

[6] Jonathan Hardwick. 2018. Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time. (2018).

[7] Stefan Janke, Selma C Rudert, Änne Petersen, Tanja M Fritz, and Martin Daumiller. 2021. Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity? *Computers and Education Open* 2 (2021), 100055.

[8] Carlos M Jarque and Anil K Bera. 1987. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique* (1987), 163–172.

[9] Jessica Stewart Kelly. 2020. Mastering your sales pitch: Selling mastery grading to your students and yourself. *PRIMUS* 30, 8-10 (2020), 979–994.

[10] Jason King. 2021. Combining Theory and Practice in Data Structures & Algorithms Course Projects: An Experience Report. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 959–965.

[11] Ryan A Miller, Annemarie Vaccaro, Ezekiel W Kimball, and Rachael Forester. 2021. "It's dude culture": Students with minoritized identities of sexuality and/or gender navigating STEM majors. *Journal of Diversity in Higher Education* 14, 3 (2021), 340.

[12] Peter Ohmann. 2019. An assessment of oral exams in introductory cs. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 613–619.

[13] Claudia Ott, Brendan McCane, and Nick Meek. 2021. Mastery Learning in CS1-An Invitation to Procrastinate?: Reflecting on Six Years of Mastery Learning. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*. 18–24.

[14] Leo Porter, Daniel Zingaro, Soohyun Nam Liao, Cynthia Taylor, Kevin C Webb, Cynthia Lee, and Michael Clancy. 2019. BDSI: A validated concept inventory for basic data structures. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*. 111–119.

[15] Scott J Reckinger and Shanon Marie Reckinger. 2021. Oral Proficiency Exams in High-Enrollment Computer Science Courses. In *2021 ASEE Virtual Annual Conference Content Access*.

[16] Kaili Rimfeld, Margherita Malanchini, Laurie J Hannigan, Philip S Dale, Rebecca Allen, Sara A Hart, and Robert Plomin. 2019. Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry* 60, 12 (2019), 1278–1288.

[17] Audrey S Rorrer. 2016. An evaluation capacity building toolkit for principal investigators of undergraduate research experiences: A demonstration of transforming theory into practice. *Evaluation and program planning* 55 (2016), 103–111.

[18] Gili Rusak and Lisa Yan. 2021. Unique exams: designing assessments for integrity and fairness. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1170–1176.

[19] Mihaela Sabin, Karen H Jin, and Adrienne Smith. 2021. Oral exams in shift to remote learning. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 666–672.

[20] Mariana Silva, Matthew West, and Craig Zilles. 2020. Measuring the score advantage on asynchronous exams in an undergraduate CS course. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 873–879.

[21] Ellen Spertus and Zachary Kurmas. 2021. Mastery-based learning in undergraduate computer architecture. In *2021 ACM/IEEE Workshop on Computer Architecture Education (WCAE)*. IEEE, 1–7.

[22] Doreen Thierauf. 2021. Feeling better: A year without deadlines. *Nineteenth-Century Gender Studies* 17, 1 (2021).

[23] Gregory M Walton and Geoffrey L Cohen. 2007. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology* 92, 1 (2007), 82.

[24] Thomas Wanner and Edward Palmer. 2018. Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education* 43, 7 (2018), 1032–1047.

[25] Daniel Woldeab and Thomas Brothen. 2019. 21st century assessment: Online proctoring, test anxiety, and student performance. *International Journal of E-Learning & Distance Education* 34, 1 (2019), 1–10.