Gender-balanced TAs from an Unbalanced Student Body

Amir Kamil
University of Michigan
akamil@umich.edu

James Juett
University of Michigan
jjuett@umich.edu

Andrew DeOrio
University of Michigan
awdeorio@umich.edu

ABSTRACT

Increasing participation of women and underrepresented minorities is a key challenge in the field of Computer Science Education. Balanced representation of these groups among teaching assistants in Computer Science courses influences recruitment and retention of underrepresented students. At the same time, the status-quo reduced participation of these students makes it more difficult to hire instructional staff from underrepresented groups.

In this paper, we describe our experience evaluating candidates with teaching-demonstration videos, followed by in-person interviews, to hire a gender-balanced set of undergraduate TAs for a large-scale CS2 course. Our research goal is to quantitatively assess gender balance throughout the hiring process.

Our initial applicant pool is just one-sixth women, but we found that women applicants perform better in our application process than men, resulting in a gender-balanced course staff without making hiring decisions based on the gender of applicants. We show that our approach results in a more gender-balanced teaching staff than hiring based on applicant GPA. We also use course-evaluation data to demonstrate that women perform as well as men as teaching assistants in CS2, and that the overall quality of our teaching assistants has remained high after the hiring-process change.

CCS CONCEPTS

• **Social and professional topics → Computing education**;

KEYWORDS

gender diversity, women in computer science, CS2, undergraduate TA, teaching staff, hiring practices, learning at scale

ACM Reference Format:
Amir Kamil, James Juett, and Andrew DeOrio. 2019. Gender-balanced TAs from an Unbalanced Student Body. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19), February 27-March 2, 2019, Minneapolis, MN, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3287324.3287404

1 INTRODUCTION

Enrollment in undergraduate computer science programs and workplace demand for computer scientists are at an all-time high. Yet, women are underrepresented in computer science at all levels. Women account for only 23% of AP computer science test-takers, 18% of computer science degrees earned at major research universities, and 26% of professional computing occupations [7].

This work in particular is concerned with representation of women as undergraduate teaching assistants in CS courses. Undergraduate teaching assistants (TAs) are common in large-scale, modern computer science courses [8, 16, 17]. TAs form the front line of our courses, interacting with students in a variety of settings, including labs, discussions, office hours, in-class exercises, and online forums.

The underrepresentation of women in computer science programs presents a challenge for identifying and hiring a gender-balanced staff. At our university, the CS2 population is approximately 25% women. Furthermore, some of the same barriers to women's participation in CS generally may also discourage applying for TA positions. In our experience, the initial applicant pool for our CS2 course is approximately 16.5% women. As we describe in detail below, simple hiring schemes based on GPA or previous grade in the course result in a staff with the same gender-imbalance as the initial applicant pool.

In this paper, we describe an application and interview process we have used to consistently hire a gender-balanced staff over each of the past 5 semesters. Crucially, this has not required hiring less-qualified students to meet diversity goals. Rather, women who apply tend to perform better in teaching-demonstration videos and in-person interviews, which leads to a gender-balanced hiring ratio despite an imbalanced applicant pool.

The research questions we address in this work are:

- What gender balance is present at all phases of the application/interview process?
- Do men and women perform differently in evaluative measures used in the hiring process?

1.1 Background and Related Work

Evidence shows several factors contribute to underrepresentation of women in CS. At the college level, a lack of previous experience with CS is a key contributor; one large-scale study at Stanford University [15] found 42.4% of women entering college reported previous experience with CS, compared with 66.3% of men. Lack of previous experience with CS puts students at a disadvantage [1], and this imbalance can manifest in stronger positive attitudes toward computing among men entering college [2, 21]. Confidence also plays an important role in students' self-concept in their field, yet the Stanford study [15] found that among CS majors, women report much less confidence than men about their choice of major entering college. This is likely related to differing levels of exposure to CS before college.

At all levels, cultural norms and stereotypes about computer science can be a barrier to women's participation. Learning environments that match these stereotypes decrease women's interest in pursuing computer science [4, 14]. Many women in CS also face

SIGCSE '19, February 27-March 2, 2019, Minneapolis, MN, USA

ACM ISBN 978-1-4503-5890-3/19/02…$15.00
https://doi.org/10.1145/3287324.3287404

stereotype threat [4, 23], which occurs when an individual feels at risk of confirming negative stereotypes about a demographic or social group with which they identify. Empirical works show stereotype threat negatively impacts women's performance and feeling of belonging in STEM and CS fields [11, 20, 22].

These barriers are compounded by the existing underrepresentation of women in computer science. In a study of undergraduate students at Stanford University [15], 84% of women CS majors self-reported feeling like a gender minority, vs. 52% for women overall. Underrepresentation of women in teaching and leadership roles may reinforce this feeling and leads to fewer same-gender role models for women in computing courses. In CS courses, TAs often act as role models for students. Haller and Fossum [9] suggest students may have an easier time identifying with a successful TA who is just a few years ahead of them than with a professor or "famous" individual whose position seems out of reach. Roberts et al. [18] describe a "stepping-stone" model which stresses the importance of women at all levels of leadership in a course.

The presence of women as role models maintains the interest of women who already have exposure to computer science [6] and combats the shrinking pipeline of women in CS [3]. Contact with in-group experts is known to enhance women's self-concept and motivation in pursuing STEM fields, and can inoculate them against negative pressures of CS stereotypes [24], improving retention of women in CS programs. TA experience is an influential and motivating factor in encouraging women to pursue careers as CS teachers [19].

Several descriptions of TA hiring are in the literature. Decker et al. [5] describe a multi-phase hiring process in use at the University at Buffalo, which involves initial paper applications, in-person interviews, presentations on a course topic, and additional panel-style interviews with current TAs. Leyzberg et al. [13] discuss a formal interview process at Princeton in which faculty meet with students and evaluate them according to a common rubric, which the authors report has led to a fairer and more effective hiring process. We examine the impact of our TA hiring processes on the gender balance of our course staff.

2 METHODS

We proceed to describe the course that uses our interview process and its context in the curriculum. We then provide details about our interview process, as well as the statistical methods we used to analyze the data we collected through the process.

2.1 Description of the course

Our data set comes from a second-semester computer programming course at a the University of Michigan, a large, public, research institution. Students take an introductory CS1 course, followed by CS2, "Programming and Introductory Data Structures". While many students are aspiring computer science majors and minors, the course also has a diverse group of students from other majors who want a second computing course.

CS2 covers major computer science concepts including functional abstraction, data abstraction and dynamic resource management. Students are exposed to C++ features like arrays, structs, classes, pointers, inheritance, polymorphism, and recursion.

Each week, students attend two 80-minute lectures and a 2-hour lab. Labs are led by TAs and contain short assignments that reinforce the lecture material. There are two exams and 5 large programming assignments. A student's final grade is comprised by 40% projects, 5% labs, and 55% exams.

Programming assignments are substantial pieces of work and many student solutions reach 1,000 lines of code. A few examples include a machine-learning tool that predicts the subject of message-board posts, and a simple image-processing algorithm. Students work in optional partnerships on programming assignments. TAs hold regular office hours that primarily support questions about the projects and studying for exams.

2.2 TA Hiring Process

In this section, we describe the interview process we have used for the last five semesters, beginning in Fall 2016, based on the method used by Dr. Mary Lou Dorf's CS1 course. At a glance, our hiring process involves advertising our position, soliciting first-round applications (including a short teaching-demonstration video), reviewing first-round applications, and bringing in top candidates for a second-round of in-person interviews. During the interviews, each candidate participates in an interactive teaching demonstration and is evaluated based on a common rubric.

Prior to this TA hiring process, our method for hiring was more ad-hoc. There was little advertising, and interested students approached faculty to inquire about the position. We only interviewed a small subset of applicants, and the interview consisted of checking a student's knowledge of the material and their past experience with teaching or mentoring.

2.2.1 First-Round Applications. We advertise our TA positions widely and accept first-round applications from any interested students. A mass email soliciting applications is sent to all declared computer science majors and all students who have taken our CS2 course in the past year. We ask current TAs to advertise the position via word-of-mouth to their peers and to encourage current students to apply. A student group hosts a "What is it like to be an EECS TA" panel in which current TAs and instructors answer questions.

Our first-round application is available online. Information provided with the application describes the responsibilities of the position. Students are asked to provide their class standing, intended major, and current GPA. They also give free-response descriptions of why they would like to TA for our course and any previous teaching experience they have. Students are also allowed to submit a resume, but this is not required. The application is open for roughly one month. In the most recent semester, we received approximately 150 applications.

Students are also required to submit a 5-minute demonstration video with their initial application. Students are allowed to choose any topic related to course material and are free to style the video any way they like (e.g. a sample lesson plan, a mock office-hours session, exam review, etc.). We emphasize that the video need not be a polished, high-quality production, but should showcase their ability to convey ideas and their teaching style.

2.2.2 Selecting Finalists for In-Person Interviews. Students' initial applications are reviewed to select finalists for second-round, in-person interviews. The primary component of this process is evaluation of the teaching-demonstration videos. A single faculty member watches and evaluates each of the videos, assigning a score from 1-5. Videos that receive a score of 3.5 or higher (about 40-50% of applications) are split among remaining faculty to receive a second 1-5 score. Reviewing videos can be a significant time investment, but viewing at 2x speed works well and an accurate impression of quality can generally be formed within a few minutes. Overall, reviewing 100-150 videos requires 5-10 hours of faculty time.

The average score for each video is used as a primary component in determining who to invite for second-round interviews. While the top few candidates with near-perfect video scores (i.e. average 4.5 to 5) are almost always selected for interviews, there are inevitably several students essentially "tied" in the next echelon of video scores. We break the ties by considering the rest of the students' application materials, previous teaching experience, and personal experience faculty have with the students, if any.

The total number of finalists we interview depends on our anticipated staff needs. Roughly one-third to half of our in-person interviews yield candidates we would be happy to hire.

2.2.3 In-Person Interviews and Evaluation. Each finalist is invited to a 20-30 minute interview with two of our faculty. The first half of the interview consists of a few standard questions (e.g. "tell us about your motivation for teaching", "what works well about our course, and what would you change to improve it?", etc.) as well as free-form conversation intended to get to know the student and allow them to feel comfortable. We also ask each candidate a question about diversity and inclusion in our courses. The second half of the interview is a teaching demonstration on a fixed topic, which we tell the candidates in advance. The interviewers act as students, and the candidate is tasked with helping the mock students understand the topic at hand. Each interviewer ranks the candidate on a five-point scale on four categories: clarity, technical proficiency, use of whiteboard, and responsiveness to student questions and needs.

Once all the interviews are completed, the instructors rank candidates based on a combination of their numerical scores on the categories described above and their responses to the free-form questions. We do not consider the gender of applicants or TA diversity – the ranking is based solely on the candidates' performance in the in-person interview. The preference list is then passed on to the department, which hires staff from this list in order as positions become available. Overall, interviewing 20 students (yielding about 8 potential hires) takes 20 hours of faculty time.

2.3 Statistical methods

We analyzed three data sets: application-video scores, evaluations of in-person interviews, and student evaluations of course staff.

Scores for application videos were collected over five semesters, from Winter[1] 2016 through Winter 2018. We received a total of 459 applications from 397 unique candidates over that period. Our analysis considers only the first application from each candidate, using the average evaluation score from each reviewer of the candidate's video. We used the university's Learning Analytics Data Architecture (LARC) [12] system to determine applicants' genders, GPA at the time of application, and grade in our course. We applied a two-sided Student's t-test to determine if men and women candidates are distinguishable by video evaluation score, GPA, or grade in the course. Grades were converted to a numerical value (A+ = 4.3, A = 4.0, A- = 3.7, and so on) for analysis.

We collected scores for in-person interviews for the four categories of clarity, technical proficiency, use of whiteboard, and responsiveness to student questions and needs over the same five semesters. We applied a two-sided Student's t-test to the scores from each category, averaged over the reviewers for each candidate, to determine if men and women candidates perform differently.

Finally, we examined data from course evaluations conducted by the university over six semesters, from Fall 2015 through Winter 2018, to determine the overall effectiveness of each of our undergraduate teaching assistants. Effectiveness is on a five-point scale, and the evaluation system reports the interpolated median over the responses for each assistant. The group of evaluated assistants includes both TAs hired under a prior process, as well as those hired through the process described above. We applied two-sided Student's t-tests to determine if men and women received different ratings, as well as to compare the performance of TAs before our new process was in place to those hired under the new process.

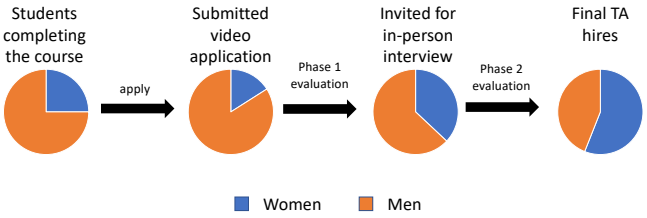

Figure 1: Hiring process overview, showing gender distribution at each step. Among students who complete the course, 25% are women. Of those, students who decide to apply for a TA position include 16.5% women. After the first interview phase which involves a teaching-demonstration video, the candidate pool contains 37% women. Finally, after an in-person interview, TA hires include 56% women.

3 RESULTS

Figure 1 shows a high-level overview of the application process and the resulting gender breakdown. The student population in the class is around 25% women, but our applicant pool has averaged 16.5% women over the course of this study. After our evaluation and interview process, 56% of the new course staff we hire are women.

In the rest of this section, we break down how women and men perform in each step of the application process. We also present data from course evaluations for staff we hire.

3.1 Impact of Initial Evaluation

Table 1 shows the results of evaluating applicants' video submissions. We only considered the first time an individual applied to our course, resulting in a data set with 18% women (though the

[1]The Winter term is January-April, and the Fall term is September-December.

	Women	Men
Count	72	325
Score Mean	3.89	3.58
Score P-Value	0.000103	
GPA Mean	3.65	3.66
GPA P-Value	0.403	
Grade Mean	3.68	3.78
Grade P-Value	0.149	

Table 1: Data from initial application, which includes a teaching-demonstration video. The data include evaluation scores of the videos, overall GPA of applicants, and grade in the course.

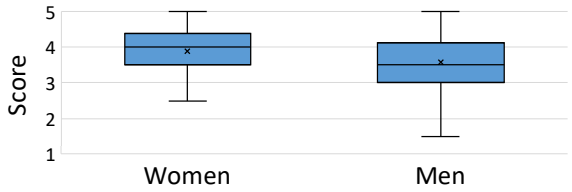

Figure 2: Initial evaluation scores of application, which includes a teaching-demonstration video. Boxes denote the quartiles, with an "X" indicating the mean.

applicant pool itself is 16.5% women). On average, women score about 9% higher than men, and the difference is statistically significant with a p-value of 0.000103. Figure 2 compares the distribution of women's scores with that of men.

We also examined how women and men applicants compare with respect to overall GPA and grade in the course. As the data in Table 1 shows, there is no statistically significant difference between the two applicant populations.

3.2 Impact of Interview

We interviewed a total of 95 candidates over five terms, with 37% women. Table 2 shows how women and men performed in the four categories of clarity, technical proficiency, use of whiteboard, and responsiveness to students. On average, women performed better than men in all four categories, and the difference was significant in every category except technical proficiency. Figure 3 illustrates the distribution of scores for all four categories.

3.3 Impact on Students

Table 3 presents data from course evaluations submitted by students for each staff member, and Figure 4 compares the distribution of scores by gender. Students are asked at the end of the term to evaluate each staff member on several metrics, including the all-encompassing "Overall, the instructor was effective." Our staff is rated as highly effective, averaging over 4.6 out of 5 possible with an overall student-response rate of 44%, and the data show no significant difference between women and men. This demonstrates that not only does our process result in hiring capable staff members,

	Women	Men
Count	35	60
Clarity Mean	4.01	3.52
Clarity P-Value	0.00293	
Technical Mean	3.93	3.65
Technical P-Value	0.0910	
Whiteboard Mean	4.07	3.51
Whiteboard P-Value	0.00258	
Responsiveness Mean	4.27	3.77
Responsiveness P-Value	0.0110	

Table 2: Interview scores in the categories of clarity, technical proficiency, use of whiteboard, and responsiveness to students.

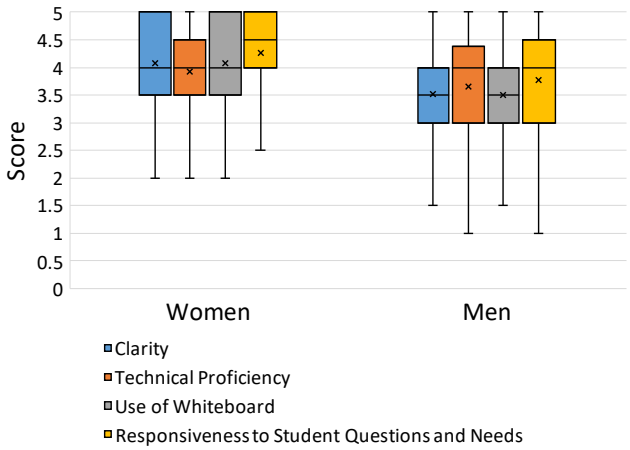

Figure 3: Interview evaluation score, by gender. An interview included a simulated office hours scenario, and interviewees were evaluated on clarity, technical competence, use of whiteboard, and responsiveness to (simulated) student questions. Boxes denote the quartiles, with an "X" indicating the mean.

	Women	Men	F15/W16	F17/W18
Count	52	47	22	41
Effective. Mean	4.65	4.62	4.56	4.63
Effective. P-Value	0.584		0.781	

Table 3: Data for student evaluations of TAs.

but that it results in both women and men staff who are highly effective.

We also compared course evaluations for staff hired before our process was in place to those hired afterwards. Table 3 compares scores in the Fall 2015 and Winter 2016 terms, which had no staff hired under the new process, to Fall 2017 and Winter 2018, where all but one staff member was hired using the process. The data show no significant difference between the two sets of staff members,

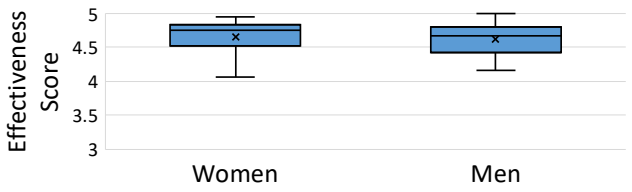

Figure 4: Student course evaluations of TAs, by TA gender. Students in each TA's lab section evaluated the TA in response to "Overall, the instructor was effective." Boxes denote the quartiles, with an "X" indicating the mean.

demonstrating that the diversity that resulted from the new process did not come at the cost of teaching effectiveness.

4 DISCUSSION

4.1 Qualitative Observations of Our Process

4.1.1 Initial Applications and Teaching-Demonstration Videos. By far, the most influential component in evaluating initial applications and deciding whom to interview is a candidate's teaching video. We have found that this provides invaluable insight into a candidate's ability to communicate clearly, appropriate use of aids such as a whiteboard or slide deck, and whether candidates explain at a pace and level that is appropriate for our students. The amount of work candidates put into their video is telling. It is surprising that some candidates that otherwise look good on paper submit a poorly-prepared, single-take video. The choice of format also informs our evaluation. For example, an interactive whiteboard session with real students is almost always more successful (and more appropriate for real teaching/learning) than a monologue. After experience with this process, judging candidates based only on paper applications would feel like being completely in the dark.

Application videos also allow our hiring process to scale to a large number of applicants. By dividing the work of evaluation between the faculty for the course, we have been able to evaluate almost 150 videos in just a few days, allowing us to identify the best candidates to bring in for in-depth, in-person interviews.

4.1.2 In-Person Interviews. The live teaching demonstration plays the largest role in determining which finalists we hire. A number of factors allow candidates to perform to their best ability and make objective evaluation across different students easier. We choose a fixed topic for all students and let them know in advance, so that they can prepare ahead of time. We let students know that the purpose of the demonstration is not to test their knowledge, but to see how they would teach and interact with real students. Along these lines, the faculty pretending to be students should be realistic; they should not ask overly tricky questions, act obstinately, or exaggerate confusions. A useful strategy is to borrow common misconceptions and/or questions we see in class.

4.2 Gender Balance/Imbalance Throughout Our Hiring Process

4.2.1 Initial Applications and Teaching-Demonstration Videos. Although women only account for 16.5% of our initial applicant pool, they make up 37% of the finalists selected for in-person interviews. That is, women who apply are three times more likely to be selected for an interview as men who apply.

A primary justification for this is that women tend to submit higher quality teaching-demonstration videos. A closer look at the distribution of video scores shows 75% of women score higher than 3.5, which is roughly our cutoff for acceptable videos that merit additional review by a second faculty member, whereas only 50% of men score above this threshold. That is, it appears a much higher percent of women who apply submit high-quality teaching demonstrations. It also appears that women perform better in our evaluation of additional criteria for students who score similarly on their teaching demos. The criteria include previous teaching experience, thoughtful responses to free-form questions, and whether any faculty personally recommend the candidate.

While we do not have evidence to explain why women tend to submit higher-quality teaching demonstrations, a possible factor is that self-selection processes by which students decide whether or not to apply may be different for men and women. For example, studies have found that among computer science students, women tend to have lower confidence in their computing abilities than men (e.g. [2, 10]). It may be this influences more severe self-selection of women than for men. Another possibility is that some of the challenges faced by women are formative in ways that strengthen their application. For example, women faced with lower confidence or stereotype threat may work especially hard to produce a high-quality teaching demo. Whatever the self-selection criteria may be, they do not appear to include GPA or grade in the course, as we have found that there is no statistically significant difference between the GPA or grades of men and women who apply.

4.2.2 In-Person Interviews. Though the pool of applicants selected for in-person interviews is just 37% women, women make up 56% of the staff we actually hire. Women we interview are twice as likely to be hired as men we interview. The primary reason for this is that they score higher than men in three of the four categories we use to evaluate their in-person teaching demonstrations. Anecdotally, women also seem to provide more thoughtful answers in the question/answer portion of the interview, but we do not have concrete data to evaluate the effects of this.

As with teaching-demonstration videos, we do not have evidence to explain why women do better in in-person interviews. It is striking that even after a first-step filtering process of evaluating videos, we find that women selected through that filter still do better than men who pass through the same filter. This suggests that both steps of the process are important in identifying the best candidates and achieving a gender-balanced yet effective staff.

4.2.3 Challenges. A significant challenge in our process is in getting women to apply in the first place. Despite the course being about 25% women, only an average of 16.5% of applicants in each term were women. Anecdotally, we have found that it requires more effort to convince promising women to apply than men. Furthermore, only 4% of women candidates applied more than once, while 16% of men candidates applied at least twice. This suggests that we should reach out to candidates who didn't quite make the cut to encourage them to apply again.

Cutoff	Count	% Women	% Men
GPA 4.0	29	24.1%	75.9%
GPA 3.9	102	17.6%	82.4%
GPA 3.8	146	17.1%	82.9%
GPA 3.7	199	16.6%	83.4%
GPA 3.6	247	17.0%	83.0%
CS2 Grade A+	62	17.7%	82.3%
CS2 Grade A	204	14.2%	85.8%
CS2 Grade A-	303	16.5%	83.5%
CS2 Grade B+	347	17.0%	83.0%

Table 4: Gender breakdown of applicants who meet various GPA or grade cutoff thresholds.

	GPA		CS2 Grade	
	Correl.	P-Value	Correl.	P-Value
Video	0.0620	0.218	0.0796	0.114
Clarity	0.0431	0.678	0.0747	0.472
Technical	0.107	0.303	0.129	0.214
Whiteboard	-0.0329	0.752	-0.00180	0.986
Responsiveness	-0.00439	0.966	0.0985	0.342
Course Evals	-0.0806	0.523	0.0566	0.654

Table 5: Correlation of GPA and grade in course to application-video score, scores for the four categories from in-person interviews, and course evaluations of TAs.

4.3 What-if? Hiring Based on GPA or Grade

We considered what our staff would look like if we hired solely based on overall GPA or grade in the course. Table 4 shows the gender breakdown of applicants under various cutoffs for GPA or grade. Regardless of the cutoff used, the ratio would not differ substantially from the 18% of first-time applicants who are women. In comparison, our interview process results in 2-3 times as many women as hiring based on GPA or grade would.

We also examined how GPA and grade correlate to scores for candidates' application videos. The "Video" row in Table 5 shows small positive correlations, but neither is statistically significant.

For the candidates we interviewed, we further compared how their GPA and grade in the course correlated with their scores for the four categories we evaluated during the interview. Table 5 shows all correlations are small and not statistically significant.

Finally, we considered the correlation between course evaluations for our staff and their GPA at time of hiring or grade in the course. Table 5 illustrates that neither correlation is statistically significant. This suggests that neither is a useful criterion for hiring staff in our course. We believe that a major reason for this is that most applicants have high enough GPA and grades to make both insignificant factors in the applicant's teaching effectiveness.

4.4 Limitations

A significant limitation to our study is that it only captures students who decide to go through the application process. The initial teaching-demonstration video is a barrier to entry, and we only evaluated candidates who put in the effort to submit a video.

It is not clear whether our application and interview process will translate to upper-level CS courses. While we have had success hiring excellent staff for CS2, it may be the criteria we use play a different role in upper-level courses. Additional criteria like further experience with specific subject matter may also be significant. Upper-level courses also tend to be smaller and faculty may know students personally, which gives additional information in evaluating candidates but may also make objectivity more difficult.

Another limitation is that our initial review of teaching videos and evaluation of in-person interviews could be influenced by implicit bias. We aim to counteract this by having multiple faculty review videos and each in-person interview, which gives multiple perspectives on each candidate's performance. We also use a multi-part rubric for our in-person interviews, which allows more objective evaluation of individual criteria. Finally, course-evaluation data show no significant difference in evaluation scores by gender, providing some evidence that our interview process is not biased.

5 CONCLUSIONS

In this study, we examined how an application process that incorporates teaching-demonstration videos and in-person interviews results in a gender-balanced course staff in a CS2 at a large, public institution. Despite the initial application pool being only 16.5% women, we ended up hiring a set of TAs that was 56% women, without making hiring decisions based on gender. Our analysis indicates women perform better in both the teaching videos they submit and in-person teaching demonstrations. We also showed that the resulting course staff was rated as highly effective by students, that there was no statistically significant difference in performance between the women and men on staff, and that the diversity produced by our hiring process did not come at the cost of reduced teaching effectiveness compared to previous terms.

We also observed that in our pool of applicants, GPA and grade in the course had no significant correlation with any of the evaluation metrics for application videos, in-person interviews, or course evaluations. Furthermore, there was no significant difference between women and men with respect to GPA or grade. This indicates that hiring solely based on GPA or grade would result in a gender ratio similar to that of the applicant pool without producing a more effective course staff.

Based on our analysis, we take several lessons that are likely to be applicable to similar courses at other institutions. The two-phase application process of teaching videos and in-person interviews scales to a large applicant pool while still producing an effective course staff. Evaluating applicants numerically on well-defined categories allows the evaluation effort to be split across multiple people and enables retrospective analysis such as this work. Combined with using two evaluators for each candidate, numerical scores may also reduce the likelihood of implicit bias. Finally, we believe that data collection is important to identify challenges in hiring a balanced staff, such as the underrepresentation of women in our applicant pool and the lower fraction of women than men who apply more than once.

REFERENCES

[1] Christine Alvarado, Gustavo Umbelino, and Mia Minnes. 2018. The persistent effect of pre-college computing experience on college CS course grades. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM, 876–881.

[2] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. 2003. Gender differences in computer science students. *ACM SIGCSE Bulletin* 35, 1 (2003), 49–53.

[3] Tracy Camp. 2002. The incredible shrinking pipeline. *ACM SIGCSE Bulletin* 34, 2 (2002), 129–134.

[4] Sapna Cheryan, Victoria C Plaut, Paul G Davies, and Claude M Steele. 2009. Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology* 97, 6 (2009), 1045.

[5] Adrienne Decker, Phil Ventura, and Christopher Egert. 2006. Through the looking glass: reflections on using undergraduate teaching assistants in CS1. *ACM SIGCSE Bulletin* 38, 1 (2006), 46–50.

[6] Benjamin J Drury, John Oliver Siy, and Sapna Cheryan. 2011. When do female role models benefit women? The importance of differentiating recruitment from retention in STEM. *Psychological Inquiry* 22, 4 (2011), 265–269.

[7] National Center for Women and Information Technology. 2018. By the Numbers. Retrieved August 8, 2018 from www.ncwit.org/bythenumbers

[8] Jeffrey Forbes, David J Malan, Heather Pon-Barry, Stuart Reges, and Mehran Sahami. 2017. Scaling introductory courses using undergraduate teaching assistants. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, 657–658.

[9] Susan M Haller and Timothy V Fossum. 1998. Retaining women in CS with accessible role models. In *ACM SIGCSE Bulletin*, Vol. 30. ACM, 73–76.

[10] Sandra Katz, David Allbritton, John Aronis, Christine Wilson, and Mary Lou Soffa. 2006. Gender, achievement, and persistence in an undergraduate computer science program. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 37, 4 (2006), 42–57.

[11] Amruth N Kumar. 2012. A study of stereotype threat in computer science. In *Proceedings of the 17th ACM annual conference on Innovation and technology in computer science education*. ACM, 273–278.

[12] LARC 2018. Learning Analytics Data Architecture, University of Michigan. https://enrollment.umich.edu/data-research/learning-analytics-data-architecture-larc

[13] Dan Leyzberg, Jérémie Lumbroso, and Christopher Moretti. 2017. Nailing the TA interview: Using a rubric to hire teaching assistants. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 128–133.

[14] Allison Master, Sapna Cheryan, and Andrew N Meltzoff. 2016. Computing whether she belongs: Stereotypes undermine girls' interest and sense of belonging in computer science. *Journal of Educational Psychology* 108, 3 (2016), 424.

[15] Katie Redmond, Sarah Evans, and Mehran Sahami. 2013. A large-scale quantitative study of women in computer science at Stanford University. In *Proceeding of the 44th ACM technical symposium on Computer science education*. ACM, 439–444.

[16] Stuart Reges. 2003. Using undergraduates as teaching assistants at a state university. In *ACM SIGCSE Bulletin*, Vol. 35. ACM, 103–107.

[17] Eric Roberts, John Lilly, and Bryan Rollins. 1995. Using undergraduates as teaching assistants in introductory programming courses: An update on the Stanford experience. *ACM SIGCSE Bulletin* 27, 1 (1995), 48–52.

[18] Eric S Roberts, Marina Kassianidou, and Lilly Irani. 2002. Encouraging women in computer science. *ACM SIGCSE Bulletin* 34, 2 (2002), 84–88.

[19] Olgun Sadik. 2015. Encouraging women to become CS teachers. In *Proceedings of the Third Conference on GenderIT*. ACM, 57–61.

[20] Jenessa R Shapiro and Amy M Williams. 2012. The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles* 66, 3-4 (2012), 175–183.

[21] Lily Shashaani. 1994. Gender-differences in computer experience and its influence on computer attitudes. *Journal of Educational Computing Research* 11, 4 (1994), 347–367.

[22] Steven J Spencer, Claude M Steele, and Diane M Quinn. 1999. Stereotype threat and women's math performance. *Journal of experimental social psychology* 35, 1 (1999), 4–28.

[23] Claude M Steele. 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American psychologist* 52, 6 (1997), 613.

[24] Jane G Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A McManus. 2011. STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of personality and social psychology* 100, 2 (2011), 255.