Modeling Improvement for Underrepresented Minorities in Online STEM Education

Nigel Bosch
University of Illinois at Urbana-Champaign
Champaign, IL, USA
pnb@illinois.edu

Eddie Huang
University of Illinois at Urbana-Champaign
Urbana, IL, USA
ezhuang2@illinois.edu

Lawrence Angrave
University of Illinois at Urbana-Champaign
Urbana, IL, USA
angrave@illinois.edu

Michelle Perry
University of Illinois at Urbana-Champaign
Champaign, IL, USA
mperry@illinois.edu

ABSTRACT

Previous research has shown that students from underrepresented minority groups tend to receive lower grades in online classes than their peers, especially in science-focused courses. We propose that there may also be benefits to online courses for these students (e.g., opportunities for peer discussions where minority status is less salient), though little is currently known about these potential benefits. We present a new perspective on learning outcomes by measuring improvement, rather than grades alone. In learning management system data from seven semesters of an online introductory science course, we found that students from underrepresented minority racial groups were indeed less likely to receive high grades, and scored lower on exams; however, their exam scores improved throughout the semester a similar amount compared to their peers. We also compared improvement to students' behaviors, including exam submission times and forum usage, finding that these behaviors were related to improvement. Finally, we also briefly discuss implications of these findings for reducing inequalities in education, and the possibilities for underrepresented minority students in online STEM education in particular.

CCS CONCEPTS

• **Applied computing → E-learning**; *Learning management systems*;

KEYWORDS

Online education, STEM education, underrepresented groups

ACM Reference Format:
Nigel Bosch, Eddie Huang, Lawrence Angrave, and Michelle Perry. 2019. Modeling Improvement for Underrepresented Minorities in Online STEM Education. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3320435.3320463

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3320463

1 INTRODUCTION

Students from underrepresented groups are disadvantaged in many education contexts [3, 15, 24, 26, 39, 40]. They are, broadly speaking, especially less likely to succeed in online educational environments [24, 39]. However, the online environment also has potential to alleviate some of the disadvantages these students face in traditional classroom contexts due to stereotype threat [5]. For example, pressures for these students to blend in, stay quiet, and a lack of opportunities for building group camaraderie may all be diminished in the online environment where appearance (e.g., gender, race) is less obvious during social interactions.

In this paper we investigate student success in an online learning space, comparing students from underrepresented groups to their peers in terms of online learning behaviors mined from log files. We examine multiple outcomes of success including course grade, exam scores, and improvement in exam scores throughout the course. We thus uncover aspects of what likely helps these students succeed, leading to recommendations for interventions and course design changes intended to promote the educational success of students from underrepresented groups.

Previous research has largely focused on outcomes such as course completion, dropout, or participation (see [12] for a recent review). Conversely, Kizilcec et al. [27] examined students behavioral trajectories during massive open online courses (MOOCs). In this paper we also consider trajectories, but focus on exam grade trajectory (i.e., improvement throughout a course) as a key outcome measure for students, in a credit-bearing college course. Based on previous work such as [39], we also expect to see differences in behavior and outcomes for students from demographic groups that are traditionally underrepresented in higher education. In particular, we focus on an introductory STEM (Science, Technology, Engineering, and Mathematics) course and consider student demographic characteristics including gender, race, and first-generation college student status (students whose parents did not attend college – a proxy for socioeconomic status).

The results in this paper demonstrate that student demographics are related to success in an online STEM course, especially for students from underrepresented minority racial groups (URM students). Our findings thus corroborate previous research in this respect, but we also make novel contributions. We deepen the understanding of what distinguishes URM students from their peers (and what makes them the same) in the online STEM education space by examining multiple definitions of successful outcomes while triangulating outcomes, student demographics, and behaviors. We demonstrate a simple method for analyzing student improvement within a course, which can easily be applied to future research. Overall, we find that URM students in this STEM course get lower course grades and exam scores, despite no notable differences in online behaviors. However, URM students improved throughout the course similarly to their peers.

We situate our results in the context of related research (described below), detail our data collection and analytic methods, and break down the nuanced relationships between demographics, outcomes, and behaviors.

2 RELATED WORK

Many related research projects show the connections between students' online behaviors and outcomes (see review in [12]), and demographic correlates of outcomes [3, 24, 26, 39–42] – the full extent of which is beyond the scope of this paper. In this section we thus focus on (i) research that highlights the importance of underrepresented status in education (broadly speaking), (ii) demographic differences in online STEM courses specifically, and (ii) research that provides detailed descriptions of behavior analysis in online courses.

2.1 Importance of students' background characteristics

Researchers have explored many different aspects of how students' background characteristics have related to educational experiences and outcomes. For example, Ogbu's anthropological work focused on – among other things – differentiating types of racial and ethnic minority statuses and their impacts on education [35]. In particular, he drew a distinction between *voluntary* minority groups (those who predominately immigrate to pursue better outcomes for themselves and their descendants) and *involuntary* minority groups (those who were historically displaced or forced to immigrate). Following Ogbu, the URM students included in our study were from involuntary minority groups, who tend to receive fewer educational opportunities and who experience less pressure to succeed from teachers, parents, and community members [31, 35].

There is also a large body of research on the importance of having students having a sense of belonging at their educational institutions [16, 18, 21, 23, 36, 37]. For example, Johnson et al. [23] found that URM students have less of a sense of belonging than their peers in first-year university studies. Similarly, in secondary school, URM students are underrepresented in advanced placement classes and report experiencing a less inviting academic culture compared to their peers [33].

Previous research also shows that non-traditional student characteristics (e.g., being a single parent, working full time) increase the likelihood of enrolling in online sections of classes. Wladis et al. [40] studied over 25,000 students and found that these characteristics compounded on each other; students were more likely to enroll in online classes if they belonged to a larger number of these demographic groups. In this paper we consider first-generation college-student status, as one indicator of non-traditional college-student status, as it relates to STEM course success.

Given such previous research findings, we considered the relations between gender and course grades, and first-generation status and grades. However, URM status appeared to be the strongest individual risk factor for receiving a low grade (Figure 4, Table 1) – a relationship that has also been extensively studied in previous research.

Xu et al. [42] studied over 40,000 students in online college courses and found that Black students received significantly lower grades relative to their peers. Kaupp [24] examined an even larger database of students (4.5 million) in online and face-to-face classes, finding that Latina/o students received significantly lower grades than their peers in online classes. We were thus motivated to consider URM students, which included Black, Latina/o, Native American, and multiracial groups in our analyses. Based on these reported findings, we expected that URM students might receive lower grades.

2.2 Minority status in STEM courses

Previous research has demonstrated that gender – specifically, being male – predicts persistence in STEM courses and majors [17, 26], despite the promising career opportunities that a STEM degree provides [8, 29]. For example, Griffith et al. [19] studied longitudinal data across universities in the United States to compare women's persistence in STEM majors to that of their peers, finding that women were significantly less likely to remain in STEM majors than men. Similarly, Crues et al. [13] examined persistence in a large MOOC focused on computer science (a STEM topic) and found that women were less likely to persist in the course.

Stereotypes about the brilliance or giftedness required for success in a field can also create difficulties for URM students in STEM. Leslie et al. [30] found that people's beliefs about the giftedness required for success in various fields (versus dedication required) correlated significantly with the proportion of African Americans who obtained a Ph.D. in those fields ($r = -0.54$). Similarly, they found that fewer women obtained Ph.D.'s in fields that value brilliance over dedication ($r = -0.58$). Furthermore, Bian et al. [4] found that these biases emerge in children as young as 6 years old. STEM fields – which tend to stereotypically require brilliance – can thus be especially prone to excluding female and URM students.

2.3 Behavior analysis in online courses

A plethora of previous work has documented relationships between students' online course behaviors and their outcomes (e.g., [2, 11, 12, 20, 22, 28, 32, 34, 43]). Links between behavior and success are thus well-established. This paper focuses primarily on differences in behavior between URM students and their peers, which is a relatively less well analyzed aspect of behavior-outcome relationships. Some previous work has examined similar connections, however.

Bosch et al. [6] explored URM status in detail with data from a university-level online STEM course, but did not examine students' race with finer granularity than White and non-White, and did not consider nuanced definitions of success. Conversely, Guo et al. [20] examined four STEM MOOCs – which included over 140,000 students – but examined race only in terms of students' countries of origin.

In this paper, we build on previous research by defining success in multiple ways to discover how online STEM education at the university level is (or is not) serving URM students effectively. We focus both on exam-taking behaviors – because exam outcomes are germane to the measure of improvement we consider – and on online discussion forum usage – because forums have been shown to be important predictors of outcomes [10, 11].

3 METHOD

3.1 Learning management system data

The data we analyzed originated from student enrollments over seven sequential semesters, over a three year period, in an introductory STEM course, offered at a large Midwest land-grant university. Every student action (i.e., clickstream data) within the web-based course was time-stamped and archived by LON-CAPA, a Learning Management System (LMS). The LMS recorded a total of 2,418,509 events, which included exam submission times, scoring per question attempt, forum views, and forum posting (see Figure 1 for an illustration of a student's view of LON-CAPA).

We received university institutional review board approval before obtaining any data that is analyzed in this paper, and examined only anonymized data from semesters that were completed, per university policies. We also obtained consent from instructors involved with the course before obtaining data.

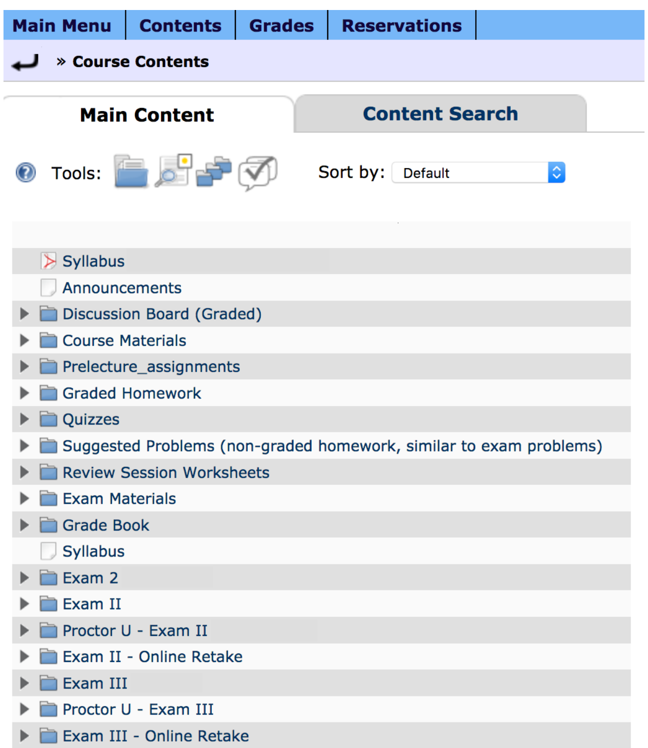

Figure 1: Screenshot of the main navigation page for a course in the LON-CAPA learning management system.

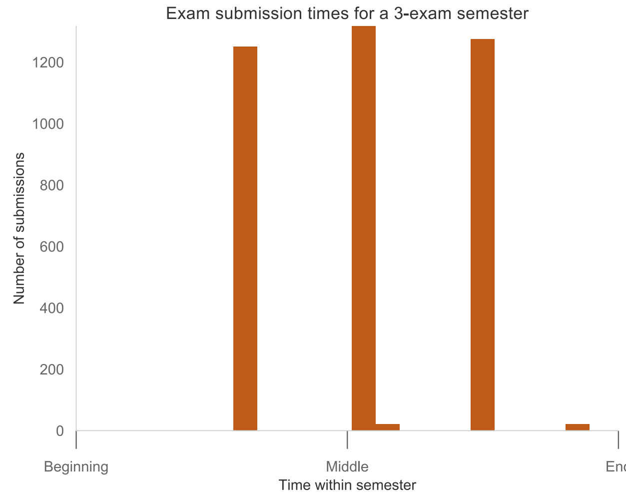

Figure 2: Example of a semester in which there were three clear exam submission times. Improvement was measured from the first exam to the second two.

The course varied in some respects across semesters. In five of the seven semesters, teaching assistants were available to answer questions (e.g., in the discussion forums), while in the two other semesters there were no teaching assistants. The presence of teaching assistants may have altered how students interacted in the discussion forums, although data from many more semesters will be needed to determine if that is the case. Additionally, two semesters were summer sessions, in which the pace of the course was accelerated. In some semesters there were three exams (e.g., Figure 2), while in others there were only two (e.g., Figure 3). In semesters with two exams, we measured improvement in exam scores from the first to last exam, while in semesters with three exams we measured improvement from the first exam to the mean score of answers in both of the subsequent two exams.

In all semesters, forum participation was counted as part of students' grades (5%). Each week, students were required to post questions, worked solutions to homework problems, or answers to other students' questions. Making weekly postings a course requirement may have resulted in increased participation (and different quality of participation) relative to online courses with optional forum participation; however, some students did not participate regularly, despite this requirement.

3.2 Student demographic data

We obtained student demographic, enrollment-by-semester, and grade data from the university's data warehouse. Prior to our analysis, a separate university data provider anonymized the LMS and demographic data. Anonymization included (i) replacing original student identifiers with a unique anonymous code, and (ii) aggregating unnecessarily detailed grade and demographic information.

We also obtained standardized test scores as a measure of academic preparation. Specifically, we obtained scores from the ACT test [1], a standardized test frequently taken by secondary school students as part of applications to universities in the United States.

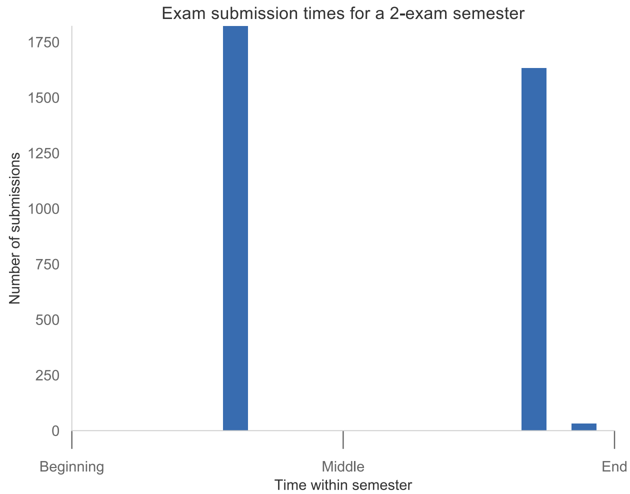

Figure 3: Example of a semester in which there were only two exams.

In this paper, we examine the ACT composite score, which aggregates individual English, mathematics, reading, and science component scores. Due to privacy concerns that precise ACT scores might identify individual students, the university data provider grouped scores into three levels selected based on aggregate university-wide test scores; group 1 included scores < 28, group 2 included scores in the range 28 − 32, and group 3 consisted of scores > 32 (out of a maximum of 36).

Our focus is on all student enrollments where the student made a significant semester-long time investment and appeared to have an intent to complete the course and earn a passing grade. With this in mind, we included all students with enrollments ($n = 506$) that led to the recording of an end-of-course grade outcome – i.e., one of {A, B, C, D, W(ithdrawal), or F(fail)} – and included an ACT score. To avoid privacy issues introduced by analyzing small groups of students, we obtained grades that were divided into four groups: A, B, C, and other (i.e., D, W, or F). Students who registered for the course and either dropped the course before the drop-deadline or never interacted with the LMS were not considered.

We obtained three binary demographic attributes for each student: Underrepresented-in-STEM racial/ethnic minority ("URM"), underrepresented-in-STEM gender ("F"), and first-generation college student status ("1"). Underrepresented minority status were students who self-identified as Hispanic, Black, Native American, or multiracial. Gender was aggregated into two categories, with the female group including all gender identities underrepresented in STEM (i.e., not "male").

3.3 Summary of course outcomes

Table 1 shows the fraction of A or B grades awarded for each demographic group. For Table 1 we combined A and B outcomes to succinctly show the proportion of students who achieved an above-average outcome. Of the three groups reported, the lowest fraction of A/B grades awarded was to the URM group. There were 24 A/B grades assigned out of 83 (28.9%) URM students, compared to 206 of 423 (48.7%) for their peers. This difference was statistically significant ($p < 0.001$) and forms the impetus for the next section of this paper, where we examine student behavior and performance within the course.

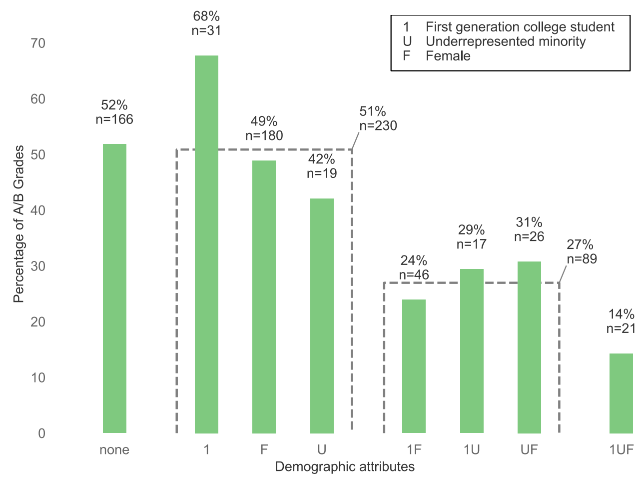

Figure 4: Percentage of students who received a high grade (an A or B) in the introductory STEM course that we discuss in this paper[1]. Students who were members of traditionally underrepresented minority racial groups tended to receive lower grades; furthermore, being a member of multiple underrepresented groups was related to even lower probability of receiving a high grade.

Table 1: A/B grade outcomes for underrepresented minority, gender and first-generation college students. There were 230 students who received A or B grades, out of all 506 students.

Demographic	n	A/B grades
URM	83	24 (28.9%)
not URM	423	206 (48.7%)
First generation	115	40 (34.8%)
not First generation	391	190 (48.6%)
Female	273	110 (40.3%)
Male	233	120 (51.5%)
All students	506	230 (45.5%)

Students may belong to more than one underrepresented demographic group, which may exacerbate challenges faced by the student. Figure 4 summarizes the fraction of A or B grades for each of the 8 sub-groups individually (none, 1, F, U, 1F, 1U, UF, and 1UF)[1].

[1]Bars in Figure 4 are independent to illustrate the importance of having a single versus having multiple underrepresented group statuses. Table 1 presents a complementary perspective where rows are not necessarily independent; for example, a female URM student is in only the *UF* section of Figure 4, but in both the *URM* and *Female* rows of Table 1.

The dashed lines illustrate the mean fraction of A/B grades for students from one underrepresented group (1, F or U) and students from two groups (1F, 1U, UF). One simple (but incomplete) interpretation of these aggregate results is that members of multiple underrepresented groups are especially at risk, because the fraction of A/B grades drops precipitously (roughly halves) when contrasting members of 0 or 1 underrepresented groups to members of 2 or 3 underrepresented groups. Thus "multiple underrepresented demographic groups means at-risk" could be a memorable – though *prima facie* – general rule for advisors and teaching assistants.

However, within the apparently unaffected aggregate group belonging to one underrepresented group (1, F or U) with a normative average of 51%, the diminished grades of the small ($n = 19$) "U" group were averaged out by the better grades of traditional-first-generation male students in the "1" group. Interventions or support *only* for students who were members of multiple underrepresented groups would miss this. Instead, a more appropriate guide is that all groups with URM status (U, 1U, UF, 1UF) and first-generation female students (1F), have an elevated risk of a lower grade. This corresponds to the last 5 right-hand groups of Figure 4.

Multiple-group effects are the focus of ongoing research that is beyond the scope of this paper. Although the sample we consider contains several hundred students, it is too small to permit in-depth analyses of students who are members of multiple underrepresented groups. Thus, we focus our more detailed analyses on the single group with the most notable disparity in grade outcomes: URM students.

3.4 Analysis procedure

3.4.1 Measuring improvement. We measured improvement in the course by dividing each semester into first and second halves and calculating mean exam question score within each half. We calculated the beginning of each semester by selecting students who achieved high grades and finding the median time of their first action in the LMS. We utilized the same method with students' last actions to calculate the end of each semester. Then, we defined the midpoint of each semester as halfway between the beginning and end.

Each exam question could be answered incorrectly (0 points), partially correctly (0.5 points) or correctly (1 point). The LMS allowed students to make multiple attempts, with only the best answer counting toward students' grades. Thus, we also considered only their best answer from all attempts per question. We then calculated mean exam question score as the mean from among their best answers for each half of the course. We defined improvement as *mean exam score (second half) − mean exam score (first half)*.

We further restricted the dataset to students with improvement scores – i.e., students who had exam scores in both halves of the semester – and those with ACT scores, resulting in a dataset of 279 students for all subsequent analyses. This dataset includes notably fewer students, predominately excluding those who withdrew (and thus submitted no more exam answers) or failed because they also did not submit exam scores. In future work we discuss potential methods to include some of the students not considered further in this paper, but such analyses are beyond the scope of this paper.

3.4.2 Extracting key student behaviors. We focused on a select set of student behaviors that are related to course outcomes (see related work). In particular, we calculated (i) the number of forum posts students made, (ii) the number of times students viewed forums, and (iii) mean exam submission time before due date. We calculated each of these behavioral features per half of the course to correspond to the course halves in our measure of improvement.

3.4.3 Testing relationships between URM status, behaviors, and improvement. We measured the strength of correlations for all analyses via Spearman's *rho*, which is appropriate given the non-normal distributions of variables in our data (e.g., ordinal count-based behavioral features described above). For correlations including URM status, we coded URM students as 1 and their peers as 0. Thus, for example, a positive correlation between URM status and a behavior would indicate that URM students exhibited that behavior more than their peers. In analyses where we controlled for a third variable when testing the correlation between two other variables (e.g., controlling for the relationship between URM status and ACT score), we computed semi-partial Spearman correlations calculated with the `ppcor` [25] package in `R` [38].

4 RESULTS

We examined success measures including grades, exam scores, and improvement. As reported above (Section 3), analysis of URM, gender, and first-generation college student demographics showed that URM status was the largest risk factor apparent in our data (Figure 4). Thus, we conducted remaining analyses focused on URM students.

4.1 Exam score comparisons

Mean exam question scores in the first half of the course were 35.2% (SD = 12.5%) for URM students and 39.9% (SD = 13.2%) for their peers, which was a significant difference (4.7% difference, $p < 0.05$) – although this effect was not large and was not significant after Bonferroni correction [14]. Differences were even larger in the second half of the course: URM students scored 41.6% (SD = 13.2%) on average, versus 48.8% (SD = 13.2%) for their peers (7.2% difference, $p < 0.005$). However, after controlling for the effect of academic preparation (ACT score) on exam scores via semi-partial Spearman correlation tests, these differences between URM students and their peers were no longer notable ($p = 0.397$ for the first half of the course, $p = 0.144$ for the second half).

In both halves of the course, mean exam scores were below typical failing threshold (e.g., 60%) for both URM students and their peers. However, students were also graded on other aspects of the course, such as participation in discussion forums. Additionally, over half of students received a low grade in the course (Table 1), so low exam scores are unsurprising.

4.2 Improvement in exam scores

Despite low exam scores, URM students improved throughout the course (mean improvement = 6.4%; SD = 12.9%). Their peers may have improved slightly more (M = 8.9%; SD = 13.3%), though this difference was not significant ($p = 0.353$). We also conducted analyses accounting for the influence of ACT score on improvement, but there were still no statistical differences between URM students and their peers in terms of improvement ($p = 0.780$).

Thus, results indicated a similar amount of improvement for URM students compared to their peers.

4.3 URM student behaviors in the online learning environment

Given their lower exam scores and grades, we expected that URM students might also exhibit different patterns of behavior potentially related to lower grades – such as procrastinating on exams or engaging less with discussion forums. However, that was not the case. In the first half of the course, URM students submitted exams at roughly the same time as their peers ($p = 0.661$), made a similar number of discussion forum posts ($M = 8.43$ versus $M = 8.13$ for their peers; $p = 0.399$), and accessed discussion forums a similar number of times ($M = 76.5$ versus $M = 72.9$ for their peers; $p = 0.297$).

Results from the second half of the course largely replicated these findings. In the second half, exam submission times differed little ($p = 0.666$), as did the number of discussion forum posts ($M = 4.80$ for URM students versus $M = 4.43$ for their peers; $p = 0.098$) and the number of discussion forum accesses ($M = 41.3$ for URM students versus $M = 42.2$ for their peers; $p = 0.562$).

4.4 Relationships between improvement and behaviors

So far, results in this paper have shown that URM students exhibited behaviors similar to those of their peers, and that they improved similarly as well. These findings raise the question of whether or not their similar improvement could have been the product of their similar behaviors. Indeed, we found that exam submission time and number of discussion posts were significantly related to improvement (Table 2).

The direction of these correlations may be somewhat surprising for both halves of the course. Submitting exam answers earlier was related to lower improvement ($rho = -0.231$ [first half], $rho = -0.218$ [second half]), as was posting more frequently in discussion forums ($rho = -0.293$ [first half], $rho = -0.208$ [second half]).

Controlling for ACT score makes little difference for these effects (e.g., $rho = -0.223, p < 0.001$ for the correlation between exam improvement and answer submission time). However, another possible explanation for these correlations is that students who improve the most were those who lack online course-taking habits we might expect to be beneficial (such as interacting on the forums). We conducted a follow-up analysis in the next section to analyze this possibility.

4.5 Relationships between course grades and behaviors

We found that behaviors we expected to be beneficial (submitting exam answers earlier, posting to the discussion forums, and viewing the forums) were positively related to course grades during both halves of the semester (Table 3). These results are in contrast to behavior correlations with improvement (Table 2), suggesting that students who are high performing (and thus have less room to improve) already exhibit these behaviors, while students with more room to improve benefit from them. A semi-partial Spearman correlation between course grade and improvement, controlling for the relationship between ACT score and improvement, showed no relationship between course grade and improvement ($rho = -0.042$, $p = 0.490$), indicating a complex relationship between grades, improvement, and behaviors.

Table 2: Spearman correlations between behaviors and exam score improvement for both halves of the course. Negative correlations indicate less improvement given a greater amount of the given behavior.

Measure	Course half	*rho*	*p*
Mean seconds until exam due	1	−0.231	< 0.001
	2	−0.218	< 0.001
Number of discussion posts	1	−0.293	< 0.001
	2	−0.208	< 0.001
Number of discussion accesses	1	−0.105	0.081
	2	−0.108	0.072

Tables 2 and 3 revealed relatively consistent effects across both halves of the semester. We thus also examined the changes in behavior over time to determine whether these similarities might be driven by consistent behaviors across halves of the course. Table 4 shows that behaviors were indeed consistent – Spearman correlations ranged from $rho = 0.541$ to 0.920, indicating that student behaviors remained relatively consistent over time.

Table 3: Spearman correlations between behaviors and course grades scores per course half (first or second half). Positive correlations indicate more of the given behavior was associated with a better grade.

Course behavior	Course half	*rho*	*p*
Mean seconds until exam due	1	−0.020	0.742
	2	0.002	0.971
Number of discussion posts	1	0.178	< 0.005
	2	0.152	< 0.05
Number of discussion accesses	1	0.189	< 0.005
	2	0.186	< 0.005

5 DISCUSSION AND CONCLUSIONS

We were interested not only in the grades that URM students received in an online STEM course, but also in how much they improved compared to their peers. Surprisingly, our results showed that URM students and their peers improved a similar amount, despite receiving lower grades (which was expected from previous research). Moreover, URM students' behaviors in the course were

Table 4: Spearman correlations between behaviors across first and second halves of the course.

Course Behavior	*rho*	*p*
Mean seconds until exam due	0.920	< 0.001
Number of discussion posts	0.541	< 0.001
Number of discussion accesses	0.684	< 0.001

also quite similar to their peers' behaviors. These findings lead to multiple possible conclusions with implications for designing effective and fair online courses.

5.1 Prior knowledge

It is possible that URM students in the course we analyzed entered the course with lower prior knowledge, on average. If this is the case, they might show similar behaviors and improvement – indicating that the course effectively promoted learning – but still finish with lower knowledge (and grades). In future work we will collect measures of prior knowledge to test this possibility. If this is indeed the case, it may indicate that the LMS and course we examined is effective for URM students, but that these students would benefit from additional preparation before beginning the course. Such preparation could be included as a new prerequisite course or as additional video lectures for the beginning of the course that have been specifically designed to address exam questions URM students typically miss.

5.2 Online course-taking abilities

In a similar vein, URM students might be less familiar with effective online course-taking habits due to inequalities in education that they experienced (e.g., fewer computer labs in high school). We will collect data regarding students' experiences with taking online classes to explore this possibility in future work. If prior online class experience proves to be a strong predictor of success, URM students might be well-served by tutorials given at the beginning of courses to teach students to leverage online education resources effectively.

5.3 Implications for inclusive online STEM courses

In the seven semesters of the STEM course we explored in this paper, we found that URM students exhibited similar behaviors and improved a similar amount relative to their peers. These findings are encouraging for the possibility that online STEM courses may alleviate some of the stereotypical pressures of being a URM student in a face-to-face class. However, this is not always the case with online classes [39, 40]; thus, further investigation is needed to determine which aspects of online courses promote equal participation and exam score improvement (e.g., required versus optional discussion forums).

Other researchers have suggested methods for developing more inclusive courses that may also be applied to the online space. For example, Hurtado & Carter [21] found that discussion of course materials outside of class improved students' sense of belonging. In the online environment, this could translate to required discussion of questions or other topics in forums.

Chestnut et al. [7] suggest strategies such as promoting a growth mindset among students (versus a fixed mindset about the potential to improve their abilities) [9], which might require changes to instructional content in online courses to incorporate computer-administered mindset interventions (e.g., [44]). The recorded nature of online discussion forums also permits possible interventions based on the language of students' posts – for example, if fixed mindset characteristics are detectable from posts.

Notably, we also found that ACT score – a measure of academic preparation – predicted better exam grades, and explained the difference in exam scores between URM students and their peers. This finding aligns with previous research on URM student differences in primary and secondary education [4, 9, 18], and highlights the importance of creating inclusive educational environments for students of all ages.

5.4 Limitations and future work

This study had a few limitations, which we plan to address in future work. First, improvement could only be measured for students who participated throughout the duration of the course, since those who withdrew or stopped participating early in the course did not have later exam scores that could be measured. In future work, we will also consider an alternative definition of improvement that is measured over the trajectory of each student's participation – whether that consists of three weeks or three months. However, such a definition of improvement has its own drawbacks, since comparing students who withdraw early and those who complete the course may not be fair comparisons.

Second, our analyses were limited to a single STEM course, although our data included seven semesters (offerings) of the course. In future work, we plan to collect data from other courses (within STEM topic domains) and repeat the analyses in this paper, to discover which patterns of improvement and behaviors may be course-specific and which generalize to other STEM disciplines. Including additional data will also enable more sophisticated analysis methods, such as machine learning models trained with fine-grained records of a wide range of student actions in the LMSs. In turn, these models may yield new insights or opportunities for beneficial interventions.

Third, we focused our analyses on URM students, although preliminary data suggest that students who are members of multiple groups traditionally underrepresented in STEM (e.g., first-generation college students who are also female) may merit thorough investigation to determine why they receive lower grades (Figure 4). Such intersectional groups naturally restrict the dataset further, however. Thus, further data collection is needed to obtain a sufficiently large sample for thorough analysis.

5.5 Conclusion

Our results demonstrate one path for creating a future with fairer STEM education through online courses. We found that URM students behaved similarly and improved similarly compared to their peers, lending credence to the hypothesis that online educational environments can offer URM students an environment with apparent reduced stereotype threat where they feel free to participate and thus reap the corresponding benefits.

Socioeconomic disadvantages, racism, prejudice, bias, and cultural stereotypes are problems minority students frequently face [15]. Our results likely reflect the varying educational, cultural, societal and socioeconomic challenges, support, and opportunities that students from different backgrounds experience prior to enrolling in – and while enrolled in – this online STEM course. Indeed, our results show no statistical differences in either improvement or exam scores for URM students, relative to their peers, after controlling for prior academic preparation. Our results suggest that URM students work in the same ways, which yield the same sorts of results, as their non-URM peers in this online STEM course.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305A180211 to the Board of Trustees of the University of Illinois. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

[1] Jeff Allen and Jim Sconing. 2005. *Using ACT assessment scores to set benchmarks for college readiness.* Technical Report 3. ACT Inc, Iowa City, IA. https://eric.ed.gov/?id=ED489766

[2] Ryan S. Baker, David Lindrum, Mary Jane Lindrum, and David Perkowski. 2015. Analyzing early at-risk factors in higher education e-learning courses. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. International Educational Data Mining Society, 150–155.

[3] David N. Beede, Tiffany A. Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E. Doms. 2011. *Women in STEM: A gender gap to innovation.* SSRN Scholarly Paper ID 1964782. Social Science Research Network, Rochester, NY. https://dx.doi.org/10.2139/ssrn.1964782

[4] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* 355, 6323 (Jan. 2017), 389–391. https://doi.org/10.1126/science.aah6524

[5] Lin Bian, Sarah-Jane Leslie, Mary C. Murphy, and Andrei Cimpian. 2018. Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of Experimental Social Psychology* 76 (May 2018), 404–420. https://doi.org/10.1016/j.jesp.2017.11.006

[6] Nigel Bosch, R. Wes Crues, Genevieve M. Henricks, Michelle Perry, Lawrence Angrave, Najmuddin Shaik, Suma Bhat, and Carolyn J. Anderson. 2018. Modeling key differences in underrepresented students' interactions with an online STEM course. In *Proceedings of TechMindSociety '18.* ACM, New York, NY, 6:1–6:6. https://doi.org/10.1145/3183654.3183681

[7] Eleanor K. Chestnut, Ryan F. Lei, Sarah-Jane Leslie, and Andrei Cimpian. 2018. The myth that only brilliant people are good at math and its implications for diversity. *Education Sciences* 8, 2 (June 2018), 65. https://doi.org/10.3390/educsci8020065

[8] Daryl E. Chubin, Gary S. May, and Eleanor L. Babco. 2005. Diversifying the engineering workforce. *Journal of Engineering Education* 94, 1 (Jan. 2005), 73–86. https://doi.org/10.1002/j.2168-9830.2005.tb00830.x

[9] Susana Claro, David Paunesku, and Carol S. Dweck. 2016. Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences* 113, 31 (Aug. 2016), 8664–8668. https://doi.org/10.1073/pnas.1608207113

[10] Scott Crossley, Mihai Dascalu, Danielle S. McNamara, Ryan S. Baker, and Stefan Trausan-Matu. 2017. Predicting success in massive open online courses (MOOCs) using cohesion network analysis. In *12th International Conference on Computer Supported Collaborative Learning (CSCL 2017)*, B. K. Smith, M. Borge, E. Mercier, and K. Y. Lim (Eds.). International Society of the Learning Sciences, Philadelphia, PA, 103–110.

[11] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16).* ACM, New York, NY, USA, 6–14.

[12] R. Wes Crues, Nigel Bosch, Michelle Perry, Lawrence Angrave, Najmuddin Shaik, and Suma Bhat. 2018. Refocusing the lens on engagement in MOOCs. In *Proceedings of the 5th ACM Conference on Learning@Scale (L@S 2018)*, Rose Luckin, Scott Klemmer, and Kenneth R. Koedinger (Eds.). ACM, New York, NY, 11:1–11:10. https://doi.org/10.1145/3231644.3231658

[13] R. Wes Crues, Genevieve M. Henricks, Michelle Perry, Suma Bhat, Carolyn J. Anderson, Najmuddin Shaik, and Lawrence Angrave. 2018. How do gender, learning goals, and forum participation predict persistence in a computer science MOOC? *ACM Transactions on Computing Education* 18, 4 (2018), 18:1–18:14.

[14] Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (March 1961), 52–64. https://doi.org/10.1080/01621459.1961.10482090

[15] Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. In *Handbook of social psychology* (4 ed.), Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (Eds.). Vol. 2. McGraw-Hill, Boston, 357–411.

[16] Tierra M. Freeman, Lynley H. Anderman, and Jane M. Jensen. 2007. Sense of belonging in college freshmen at the classroom and campus levels. *The Journal of Experimental Education* 75, 3 (April 2007), 203–220. https://doi.org/10.3200/JEXE.75.3.203-220

[17] Jennifer L. Glass, Sharon Sassler, Yael Levitte, and Katherine M. Michelmore. 2013. What's so special about STEM? A comparison of women's retention in STEM and professional occupations. *Social Forces* 92, 2 (Dec. 2013), 723–756. https://doi.org/10.1093/sf/sot092

[18] Carol Goodenow. 1993. The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools* 30, 1 (1993), 79–90. https://doi.org/10.1002/1520-6807(199301)30:1<79::AID-PITS2310300113>3.0.CO;2-X

[19] Amanda L. Griffith. 2010. Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review* 29, 6 (Dec. 2010), 911–922. https://doi.org/10.1016/j.econedurev.2010.06.010

[20] Philip J. Guo and Katharina Reinecke. 2014. Demographic differences in how students navigate through MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S '14).* ACM, New York, NY, 21–30. https://doi.org/10.1145/2556325.2566247

[21] Sylvia Hurtado and Deborah Faye Carter. 1997. Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education* 70, 4 (1997), 324–345. https://doi.org/10.2307/2673270

[22] Suhang Jiang, Mark Warschauer, Adrienne E. Williams, Diane O'Dowd, and Katerina Schenke. 2014. Predicting MOOC performance with week 1 behavior. In *Proceedings of the 7th International Conference on Educational Data Mining.* Educational Data Mining Society, 273–275.

[23] Dawn R. Johnson, Matthew Soldner, Jeannie Brown Leonard, Patty Alvarez, Karen Kurotsuchi Inkelas, Heather T. Rowan-Kenyon, and Susan D. Longerbeam. 2007. Examining sense of belonging among first-year undergraduates from different racial/ethnic groups. *Journal of College Student Development* 48, 5 (Sept. 2007), 525–542. https://doi.org/10.1353/csd.2007.0054

[24] Ray Kaupp. 2012. Online penalty: The impact of online instruction on the Latino-white achievement gap. *Journal of Applied Research in the Community College* 19, 2 (April 2012), 3–11.

[25] Seongho Kim. 2015. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods* 22, 6 (Nov. 2015), 665–674. https://doi.org/10.5351/CSAM.2015.22.6.665

[26] René F. Kizilcec and Sherif Halawa. 2015. Attrition and achievement gaps in online learning. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale.* ACM, New York, NY, 57–66. https://doi.org/10.1145/2724660.2724680

[27] René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK 2013).* ACM, New York, NY, 170–179. https://doi.org/10.1145/2460296.2460330

[28] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs.* Association for Computational Linguistics, Stroudsburg, PA, 60–65.

[29] David Langdon, George McKittrick, David Beede, Beethika Khan, and Mark Doms. 2011. *STEM: Good jobs now and for the future.* US Department of Commerce, Washington, DC, USA.

[30] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347, 6219 (Jan. 2015), 262–265. https://doi.org/10.1126/science.1261375

[31] Mikael Luciak. 2004. Minority status and schooling—John U. Ogbu's theory and the schooling of ethnic minorities in Europe. *Intercultural Education* 15, 4 (Dec. 2004), 359–368. https://doi.org/10.1080/1467598042000313395

[32] Leah P. Macfadyen and Shane Dawson. 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education* 54, 2 (Feb. 2010), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

[33] Jason McDonald. 2014. *The role of invitational theory on minority student enrollment in advanced placement courses.* PhD Thesis. University of Central Florida. https://stars.library.ucf.edu/etd/4767

[34] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. 2003. Predicting student performance: An application of data mining methods with an educational Web-based system. In *33rd Annual Frontiers in Education Conference (FIE 2003)*. IEEE, Piscataway, NJ, 13–18. https://doi.org/10.1109/FIE.2003.1263284

[35] John U. Ogbu and Herbert D. Simons. 1998. Voluntary and involuntary minorities: A cultural-ecological theory of school performance with some implications for education. *Anthropology & Education Quarterly* 29, 2 (1998), 155–188. https://doi.org/10.1525/aeq.1998.29.2.155

[36] Patrick O'Keeffe. 2013. A sense of belonging: Improving student retention. *College Student Journal* 47, 4 (Dec. 2013), 605–613.

[37] Terrell L. Strayhorn. 2012. *College students' sense of belonging: A key to educational success for all students* (1st ed.). Routledge, New York, NY.

[38] R Core Team. 2013. R: A language and environment for statistical computing. (2013).

[39] Claire Wladis, Katherine M. Conway, and Alyse C. Hachey. 2015. The online STEM classroom—Who succeeds? An exploration of the impact of ethnicity, gender, and non-traditional student characteristics in the community college context. *Community College Review* 43, 2 (April 2015), 142–164. https://doi.org/10.1177/0091552115571729

[40] Claire Wladis, Alyse C. Hachey, and Katherine Conway. 2015. Which STEM majors enroll in online courses, and why should we care? The impact of ethnicity, gender, and non-traditional student characteristics. *Computers & Education* 87 (Sept. 2015), 285–308. https://doi.org/10.1016/j.compedu.2015.06.010

[41] Di Xu and Shanna Smith Jaggars. 2011. The effectiveness of distance education across Virginia's community colleges: Evidence from introductory college-level math and English courses. *Educational Evaluation and Policy Analysis* 33, 3 (Sept. 2011), 360–377. https://doi.org/10.3102/0162373711413814

[42] Di Xu and Shanna S. Jaggars. 2014. Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. *The Journal of Higher Education* 85, 5 (Sept. 2014), 633–659. https://doi.org/10.1080/00221546.2014.11777343

[43] Cheng Ye and Gautam Biswas. 2014. Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics* 1, 3 (2014), 169–172.

[44] David S. Yeager, Carissa Romero, Dave Paunesku, Christopher S. Hulleman, Barbara Schneider, Cintia Hinojosa, Hae Yeon Lee, Joseph O'Brien, Kate Flint, Alice Roberts, Jill Trott, Daniel Greene, Gregory M. Walton, and Carol S. Dweck. 2016. Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology* 108, 3 (April 2016), 374–391. https://doi.org/10.1037/edu0000098