An Examination of Unofficial Course Reviews in a Graduate Program at Scale

Bobbie L. Eicher
bobbie.eicher@cc.gatech.edu
Georgia Institute of Technology
Atlanta, USA

David A. Joyner
david.joyner@gatech.edu
Georgia Institute of Technology
Atlanta, USA

ABSTRACT

Past research on the ways that students evaluate their courses has focused largely on how those evaluations relate to the specific course instructor. This research examines a set of data from a public site where students unofficially rate the courses in a very large online graduate program operating at scale. We examine the relationship between the unofficial scores students give to their classes with data on enrollment trends over time and the assessment strategies used within the courses themselves to examine additional factors that shape the ratings students choose, as well as how they use those ratings to choose what courses to take in the future. We find several different notable relationships: reviews in this context are largely impervious to the extreme response bias prevalent on other review sites; review content does not appear to significantly influence enrollment trends; more difficult classes tend to receive more favorable ratings overall, although individual students do not rate difficult classes more favorably; and project-based classes are perceived by students to be less difficult.

CCS CONCEPTS

• **Applied computing** → **Collaborative learning**; **Distance learning**; **E-learning**.

KEYWORDS

course reviews, adult education, online education, education at scale, student evaluations of teaching

ACM Reference Format:
Bobbie L. Eicher and David A. Joyner. 2022. An Examination of Unofficial Course Reviews in a Graduate Program at Scale. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3491140.3528330

1 INTRODUCTION

Offering education at increasingly larger scales has made it far easier to collect detailed information about how students interact with their courses over time; examples of this include clickstream data mining [8, 13, 18], user modeling [7], and predictive modeling [5, 14, 22]. This new capability—though it does pose risks, as misused and misgathered data can be toxic [19]—has provided a great opportunity for improving our understanding of how learning happens and what it takes to succeed.

An as-yet somewhat underappreciated element of this, however, is that students can now also take advantage of scale to improve their own information-gathering about their educational opportunities. One prominent class of such initiatives is the emergence of independent sites for students to review courses and instructors; large, cross-institution sites have sprung up like RateMyProfessor.com and ClassCentral.com, but there are also instances where students create their own dedicated platforms solely for their university or program. Such sites gather millions of reviews for courses offered across hundreds of institutions, giving a valuable large-scale resource for investigating elements that contribute to learner satisfaction. Further research is needed to understand how students work together to amass this useful information and how they respond to that information once they have it. This is especially important because there are schools actively discouraging the use of these sites without a complete understanding of how students are using them or what value they may offer [4], which may prove harmful to some students if there is real utility to the reviews.

This research focuses on one course review site built and maintained by the students in a very large online graduate program in computer science. This particular site is both non-commercial and unofficial, run by neither a defined independent entity nor the university itself, and as such the students have control over what information they collect in each review and how it is then made available. In this paper we examine the infrastructure that they have put into place, the data they have collected, connections between their data and other data sources about the program, and ways in which access to this information may be shaping their behavior and educational experience.

1.1 RQ1: Extreme Response Bias

A very common view among educators is that informal reviews cannot be trusted because they will only be filled out by students who go out of their way to do so, and these students are likely to be those who feel unusually strongly rather than a representative sample of the students who have studied with a particular instructor [3, 17, 20, 21]. Is our pool of reviews dominated by expressions of strong emotion?

1.2 RQ2: Course Registration Decisions

Is there any indication that students respond to the reviews by being more or less likely to enroll in specific classes? For example,

L@S '22, June 01–03, 2022, New York City, NY, USA

ACM ISBN 978-1-4503-9158-0/22/06…$15.00
https://doi.org/10.1145/3491140.3528330

if the reviews report that a specific course is unusually difficult, does that affect whether students take it in the future?

1.3 RQ3: Difficulty and Favorability

Numerous other studies have pursued the idea that students may reward easy courses with more favorable reviews, with varying conclusions about whether a correlation exists [2, 3, 6, 13, 15–17]. Does the students in this program reward easier classes in this way?

1.4 RQ4: Difficulty, Favorability, and Assessment

Do the kind of assessments used in a class (e.g., project-focused courses compared to those that emphasize exams) have a significant impact on how students review the courses?

2 RESEARCH CONTEXT

This research focuses on an asynchronous online Master of Science in Computer Science degree at Georgia Tech, a major research institution within the United States. As of spring 2022, the program has over 12,000 active students and it largely (though not exclusively) targets and is populated by working professionals [9]. While all program requirements are identical to the version of the program offered on campus, students are restricted to part-time enrollment. This means that no student may take more than 3 courses per semester, and most students take only one course at a time in a given semester.

The program-specific review data has been collected by the students themselves through a website called OMSCentral [1, 12], that remains entirely under student control and does not use an institutional login. This means that students can create accounts and write reviews anonymously, but also means that there is no verification that someone writing a review has actually taken a course, or that a reviewer has not submitted multiple reviews under different names for the same class. The information collected has expanded slightly over time, but as of spring 2022 the site collects 6 pieces of information with each review:

- Course: The course the review is evaluating.
- Term: The semester in which the student took the course.
- Difficulty: The reviewer's assessment of the course's difficulty on a scale of 1 (Very Easy) to 5 (Very Hard)
- Rating: The student's overall perception of the course on a scale of 1 (Strongly Disliked) to 5 (Strongly Liked), also referred to as Favorability in this paper to distinguish it more clearly from the difficulty rating
- Workload: The student's estimate of their average amount of time spent per week
- Comments: Free text commentary and review of the course where students can provide more detail or share whatever additional information they like (though the maintainers of the site do delete any content they deem to be inappropriate)

3 DATA COLLECTION

3.1 OMSCentral

The student maintainers of the site agreed to provide a spreadsheet covering all reviews, but including only the course, term, difficulty rating, favorability rating, and workload. This eliminated any data about the specific accounts posting or the written comments, ensuring that there would be no possibility of data being included that could link specific students to the reviews they had written.

Table 1: Assessment categories and descriptions as provided to instructors in the survey. The terms in quotation marks refer to how assessments were described in the syllabus.

Category	Description
Exam	anything referred to as "exam" or "test"
Homework	anything referred to as "homeworks", "written critiques", "programming assignments", or "problem sets"
Lab	anything referred to as "lab"
Participation	anything referred to as "class discussion", "forum posts", or "peer review"
Project	anything referred to as "projects", "mini-projects", "project proposals", or "project presentations"
Quiz	anything referred to as "quiz"
Other	anything that did not fall into one of the categories above

3.2 Course Assessment Data

For comparison with the course reviews, we also used data on the assessments used in each course and how they are weighted from previous work examining the syllabi and assessment practices in the program [10, 11]. The relevant portion of that research involved extracting information about assessments used and how they are weighted from the syllabus of most courses in the program (for the spring 2021 offerings). These were then clustered into categories, with a focus on categorizing the assessments according to the language the professors used to describe them. A brief summary of the categories is shown in Table 1.

3.3 Enrollment Data

For course-by-course enrollment history, total enrollment numbers for each course offered in the program and for each semester of the program's existence were pulled from the official university registration system. This includes only the total enrollments taken for each class since the program's inception. It also includes all students who were enrolled at the end of the registration period, and therefore does not exclude those students who eventually withdraw rather than finish a course.

To compensate for the fact that the program was rapidly growing over time and some classes were offered more irregularly than others, our calculations involving enrollment include only those courses that have been offered every semester since spring of 2017, and normalize the values by looking at the share of enrollment each course had rather than the absolute numbers. Therefore, change from one semester to another would be how the enrollment within that pool of courses changes relative to each other. This calculation would treat a course as having become smaller if it grew more slowly than other courses in the pool, even if the actual number of students in the class increased.

4 RESULTS AND DISCUSSION

4.1 RQ1: Extreme Response Bias

For this question, we focus on the favorability rating the students assigned to their review to look for indications that they are dominated by people with especially strong opinions, whether positive or negative. The breakdown of how many times each possible score was seen is shown in Figure 1. If extreme responses were driving the reviews, we would expect to see a lot of the responsees falling under either "Strongly Liked" or "Strongly Disliked". We do see a tendency for the sentiment to be positive, but there are fewer people selecting "Strongly Liked" than "Liked" and the difference between "Neutral" and "Strongly Disliked" is less than 1%. Based on this, we conclude that extreme response bias does not appear to be an issue in this dataset.

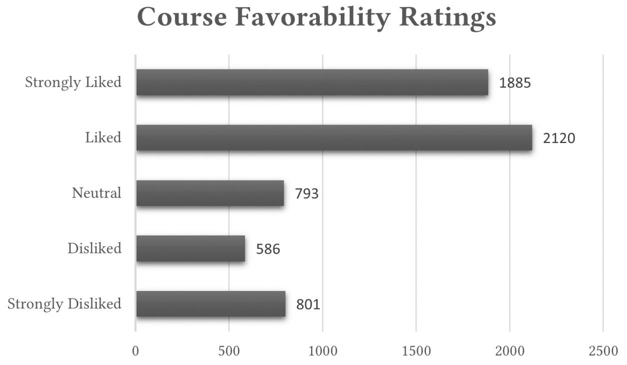

Figure 1: Number of times each favorability rating score was selected for a review.

Some simple statistical information on difficulty, rating, and workload is available in Table 2. The mean and skew are shown for both the reviews taken individually, and aggregated by course (limited to courses with at least 50 reviews to minimize the impact of courses with very few reviews). These likewise show no indication of reviews being dominated by the most extreme scores.

Table 2: Summary statistics for all reviews individually, and also aggregated by course (including only courses with greater than 50 reviews).

	Individual		Aggregated	
	Mean	Skew	Mean	Skew
Difficulty	3.12	-0.06	3.27	0.19
Rating	3.56	-0.70	3.63	-0.61
Workload	14.30	2.89	15.05	0.88

Table 3: Correlations between the average student review values across all semesters for each category among themselves and also with the weightings of different assessment types in the same courses.

	Difficulty	Rating	Workload
Difficulty	1.00		
Rating	0.39^{*}	1.00	
Workload	0.90^{***}	0.35	1.00
Quizzes	0.02	0.00	-0.01
Projects	-0.48^{**}	-0.12	-0.42^{*}
Homeworks	0.23	0.01	0.41^{*}
Exams	0.30	0.15	0.08
Participation	-0.15	-0.02	-0.11

Note: * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4.2 RQ2: Course Registration Decisions

For the sake of comparability, our look at course registration focuses only on those courses that were offered every semester after Spring 2017. Based on these calculations, the correlation between enrollment change each semester and difficulty scores had a mean of 0.03, for rating the correlation was -0.01, and for workload it was 0.02. None of these values were statistically significant. There is no indication then, based on these calculations, of a correlation between the ratings or workload and changes in enrollment in the semester immediately following.

We have other data that tells us a large proportion of the students do look at reviews and report finding them valuable [12], so this is not explained by a lack of awareness. It may be the case that aggregate registration changes very little because students are driven mainly by specific program requirements and do not have enough options available to meet requirements to change their choices. It is also possible that there is an effect that are calculations do not show (for example, because registration for one semester happens before the previous semester ends, student's may not be seeing the reviews from the most recent semester early enough for that to be the set influencing their decisions). We intend to do future work exploring the topic of how students use reviews, such as whether students may change their registration decisions in response to reviews, but in ways that either are not visible when the student body's behavior is viewed in aggregate or which could not be captured with the specific comparisons we attempted.

4.3 Difficulty and Favorability

In an effort to minimize the effect of courses with very few reviews that may be heavily influenced by one or two students with strong opinions, we restricted this calculation to courses with at least 50 reviews available. The correlations between the different ratings when all the reviews for each course are aggregated together are shown in Table 3. In the aggregate for each course, there is a statistically significant correlation of 0.39 between how favorably a student rates a course and how difficult the same student says it is. When we did the same calculation for the individual reviews of the same pool of courses without aggregating them by course, however, there was no statistically significant correlation. It appears that the students as a group are somewhat more likely to give favorable ratings to courses that are generally seen as more difficult, but that this relationship only exists for the average view of each course and not for the views students express individually.

What we can say with confidence is that there is no indication of the specific relationship that has been widely hypothesized (that easier courses would be reviewed more favorably). We have some reason to believe there is a relationship that goes in the other direction, but it is a possibility that will require further exploration, first to confirm the result with additional data and then to seek possible explanations.

4.4 RQ4: Difficulty, Favorability, and Assessment

Finally, we examined whether the assessments used in a course appear to affect student perceptions of the course. The results are available in Table 3. Unsurprisingly, there is a very strong relationship between difficulty and overall workload reported. This relationship is represented more visually in Figure 2.

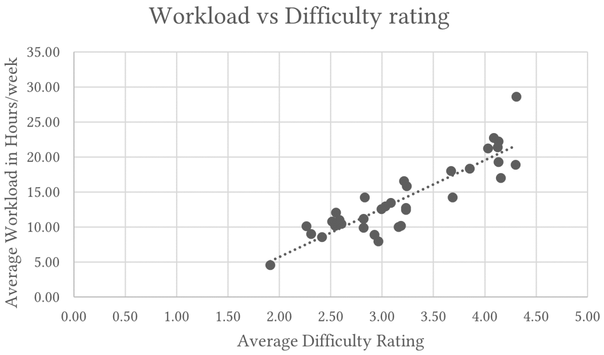

Figure 2: Relationship between average difficulty rating and average workload reported on reviews for classes with at least 50 reviews, aggregated by course.

More interestingly, courses that use projects appear to be regarded as both easier and less time-consuming to a statistically significant degree. This may be due to the specific population participating in the program. Course projects are often identical to those used in courses on campus, which are intended to be used by a population of graduate students who have mainly come to graduate school directly out of their undergraduate degree. However, this particular program is mainly working adults who often have years of experience as professional programmers [9], and so this could explain why the complexity of the programming assignments may not be very challenging to them as a group. It is also possible that the amount of time many students have been away from school has left them feeling out of practice at skills like preparing for exams and writing papers, so these things feel much more difficult to them relative to programming assignments.

Homework assignments appear to have a statistically significant impact on increasing the amount of time spent on classes without a similar effect on the difficult, so it may be that these are typically regarded as more time-consuming than challenging. None of the other assessments appear to have any effects, though it is worth noting that Exams fell only slightly short of the $p < 0.05$ standard to be considered statistically significant. Since the standard in this work for an "exam" was only that it was referred to as an exam in the syllabus of the course, it may be worthwhile in future work to look at the types of exams used in the program more closely, and examine whether there may be some sub-types that do have an effect.

5 FUTURE WORK

There are two major directions we hope to explore in future work: examining how the students use reviews in making decisions about their academic careers, and analyzing the text students submit along with their review scores which we omitted from this initial study.

5.1 Student Interaction With Reviews

We have found some indication in this preliminary work that there are genuinely interesting trends and correlations within the review data, but we do not yet have a detailed sense of how students are using the data and whether its use improves or harms their academic experience overall. Some early data about this is available [12], but we particularly need to do a deeper exploration in light of the lack of any obvious effect of reviews on registration patterns.

For example, while students widely use the site and report valuing the information, there is no data on whether there is an actual benefit to its use. We are particularly interested in exploring whether students make any clear changes to their behavior in response to reading reviews. We have seen no indication of an effect on registration at the level of the courses as a whole, but it may still be that individual students are altering their behavior in interesting ways (such as using workload information to choose classes based on how demanding they expect the other aspects of their life to be that term). Such information could help future students to make more informed decisions in their use of this kind of tool. It could also lead to a better understanding of what kind of additional information about each course it would be valuable to make available officially.

5.2 Text Analysis

A quick look through the length of the most recent 150 reviews, checking only their length, showed that they averaged 330 words. Given that there are 6,894 reviews total at this time, a significant amount of student time has gone into this content. This volume of text offers a tremendous opportunity to gain more insight into the kinds of information students feel they should offer, and how this differs among subject areas and individual classes.

ACKNOWLEDGMENTS

The authors would like to express their great appreciation to Mehmet Bajin (mehmet.bajin@gatech.edu) for providing an anonymized version of the review data from the OMSCentral site, and also more broadly for the work he has done as the creator and primary maintainer of this free resource.

REFERENCES

[1] Mehmet Baijin. 2022. *OMSCentral*. OMSCentral. https://omscentral.com

[2] April Bleske-Rechek and Amber Fritsch. 2011. Student Consensus on RateMyProfessors Com. *Practical Assessment, Research, and Evaluation* 16 (2011), 12 pages. https://doi.org/10.7275/GKF7-CE80

[3] April Bleske-Rechek and Kelsey Michels. 2019. RateMyProfessors Com: Testing Assumptions about Student Use and Misuse. *Practical Assessment, Research, and Evaluation* 15, 1 (Nov. 2019), 12 pages. https://doi.org/10.7275/ax6d-qa78

[4] Abby Boyd and Bridget G. Trogden. 2021. Why Rate My Professors and Other Sites Do Not Benefit Students' Registration Decisions.

[5] Christopher Brooks, Craig Thompson, and Stephanie Teasley. 2015. Who You Are or What You Do: Comparing the Predictive Power of Demographics vs. Activity Patterns in Massive Open Online Courses (MOOCs). In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*. Association for Computing Machinery, New York, NY, USA, 245–248. https://doi.org/10.1145/2724660.2728668

[6] John A. Centra. 2003. Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education* 44, 5 (Oct. 2003), 495–518. https://doi.org/10.1023/A:1025492407752

[7] Ronny Cook, Judy Kay, and Bob Kummerfeld. 2015. MOOClm: User Modelling for MOOCs. In *User Modeling, Adaptation and Personalization (Lecture Notes in Computer Science)*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). Springer International Publishing, Cham, 80–91. https://doi.org/10.1007/978-3-319-20267-9_7

[8] 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)*. Association for Computing Machinery, New York, NY, USA, 6–14. https://doi.org/10.1145/2883851.2883931

[9] Alex Duncan, Bobbie Eicher, and David A. Joyner. 2020. Enrollment Motivations in an Online Graduate CS Program: Trends & Gender- and Age-Based Differences. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. Association for Computing Machinery, New York, NY, USA, 1241–1247. https://doi.org/10.1145/3328778.3366848

[10] Bobbie Lynn Eicher and David Joyner. 2021. Components of Assessments and Grading At Scale. In *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event, Germany) *(L@S '21)*. Association for Computing Machinery, New York, NY, USA, 303–306. https://doi.org/10.1145/3430895.3460165

[11] Bobbie Lynn Eicher and David Joyner. 2021. Toward Reshaping the Syllabus for Education at Scale. In *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event, Germany) *(L@S '21)*. Association for Computing Machinery, New York, NY, USA, 355–358. https://doi.org/10.1145/3430895.3460987

[12] Bobbie Lynn Eicher and David Joyner. 2022. Student Use of Course Reviews at Scale. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22)*. Association for Computing Machinery, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491140.3528332

[13] James Felton, Peter T. Koper, John B. Mitchell, and Michael Stinson. 2006. *Attractiveness, Easiness, and Other Issues: Student Evaluations of Professors on RateMyProfessors.Com.* SSRN Scholarly Paper ID 918283. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.918283

[14] Josh Gardner, Christopher Brooks, Juan Miguel Andres, and Ryan Baker. 2018. Replicating MOOC Predictive Models at Scale. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale (L@S '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3231644.3231656

[15] Herbert W. Marsh. 1987. Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Educational Research* 11, 3 (Jan. 1987), 253–388. https://doi.org/10.1016/0883-0355(87)90001-2

[16] Herbert W. Marsh and Lawrence A. Roche. 1997. Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *American Psychologist* 52, 11 (1997), 1187–1197. https://doi.org/10.1037/0003-066X.52.11.1187

[17] James Otto, Douglas A. Sanford, and Douglas N. Ross. 2008. Does Ratemyprofessor.Com Really Rate My Professor? *Assessment & Evaluation in Higher Education* 33, 4 (Aug. 2008), 355–368. https://doi.org/10.1080/02602930701293405

[18] Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. 2017. Detecting Changes in Student Behavior from Clickstream Data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/3027385.3027430

[19] Justin Reich. 2020. Two Stances, Three Genres, and Four Intractable Dilemmas for the Future of Learning at Scale. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery, New York, NY, USA, 3–13. https://doi.org/10.1145/3386527.3405929

[20] Thomas Timmerman. 2008. On the Validity of RateMyProfessors.Com. *Journal of Education for Business* 84, 1 (Sept. 2008), 55–61. https://doi.org/10.3200/JOEB.84.1.55-61

[21] Dimitrios Tsekouras. 2017. The Effect of Rating Scale Design on Extreme Response Tendency in Consumer Product Ratings. *International Journal of Electronic Commerce* 21, 2 (April 2017), 270–296. https://doi.org/10.1080/10864415.2016.1234290

[22] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen, and Dustin Tingley. 2017. MOOC Dropout Prediction: How to Measure Accuracy?. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. Association for Computing Machinery, New York, NY, USA, 161–164. https://doi.org/10.1145/3051457.3053974