Scaling Up Data Science Course Projects: A Case Study

Bhavya Bhavya, Jinfeng Xiao, and Chengxiang Zhai

University of Illinois at Urbana-Champaign
{bhavya2, jxiao13, czhai}@illinois.edu

ABSTRACT

Large-scale, online Data Science (DS) courses and degree programs are becoming increasingly common due to the global rise in popularity and demand for data scientists. Although project-based learning is integral to gaining hands-on experience in DS education, providing fair, timely, and high-quality feedback on varied projects for a large number of diverse students is challenging. To address those challenges in scaling up the assessment of DS group projects, we integrated multiple techniques, such as rapid feedback, peer grading, graders as meta-reviewers, etc. We present a case study of deploying those strategies for group projects in a large online DS course titled Text Information Systems offered in Fall, 2020. We synthesize our findings from analyzing student and grader survey responses, and share useful lessons and future work.

Author Keywords

scalable assessment; data science education; course projects

CCS Concepts

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

1. INTRODUCTION

The rising popularity and demand for data scientists have led to an increase in the number of online Data Science (DS) courses and degree programs offered by many universities. These courses and degree programs are taken by a diverse group of learners globally with different backgrounds and proficiency levels, including working professionals in multiple industries. This creates more opportunities for collaboration among students having diverse academic backgrounds and skills; one such opportunity is collaborative group projects.

Project-based learning is crucial for providing hands-on experience in DS education [9]. However, it is challenging to grade all the individualized projects fairly with timely feedback, especially at scale. Several strategies have been proposed for project assessment at scale, including peer-reviews [1, 12], graders as meta-reviewers [5], rapid feedback [7], etc. Since each of those strategies addresses different assessment challenges, it is desirable to combine them in practice. But, their combined effectiveness remains unknown, especially for assessment of large-scale DS course projects.

L@S '21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06.
https://doi.org/10.1145/3430895.3460168

We combined and adapted several such strategies for project assessment in a DS course on text information systems with a diverse audience of ≈ 400 students in Fall, 2020. The course was offered at a large public university in the U.S. To evaluate the effectiveness of the course project design, we synthesized our observations from student peer reviews and conducted surveys with students and graduate teaching assistants (TAs). In this paper, we share our design of the course project, findings and lessons learned from our experience.

2. RELATED WORK

Many innovative approaches, tools and platforms have been developed for assessing large-scale practical DS assignments. CLaDS [4] is a cloud-based virtual platform that allows a large number of learners to work on autograded leaderboard-style competitions based on large data sets efficiently. Nevertheless, such autograding platforms do not support open-ended, personalized group projects.

Peer assessment and grading are commonly used strategies for proving qualitative and quantitative feedback on group projects and assignments [1, 12, 10]. Since peer-reviews can be perceived as unreliable, there have been efforts to improve their accuracy and reliability [6, 5]. Joyner et al. [5] investigated a two-tier approach to grading short, individual assignments where graders act as meta-reviewers and grade assignments in the context of peer reviews. They found that this approach improves students' perceptions of grader feedback without decreasing grader efficiency. Rapid feedback [7] is another useful peer-review methodology that aims to provide continuous feedback on in-progress work. These strategies have not been explored for programming-based group projects, especially in DS courses. Moreover, they have not been combined and evaluated in practice.

3. COURSE BACKGROUND

Team-based projects can help students gain practical experience in DS courses at scale. For flexibility, it is also desirable to encourage students to work on a variety of project topics of their choice. We aim to investigate how to assess such varied data science projects at scale in the context of a large DS class.

The course project and intervention strategies were deployed in a semester-long course on text mining at a large public university located in the U.S. . In Fall 2020, the course was fully online. Students include 1) regular undergraduates and graduates in majors such as Computer Science; 2) DS degree program graduate students who are typically working professionals in the Information Technology industry. A total of 378 students took the course, 66.9% of whom were DS degree program graduate students hereafter referred to as *DSG*, 25.3% were undergraduates, and 7.6% were other graduate students.

Although the students were diverse also in other ways, e.g., geographical regions, in this work, we limit our discussion on **diversity** based on their academic levels and degree programs. Eight graduate TAs and one course instructor supported the students.

Project topics: We released multiple project topics varied by the core course concepts (e.g. text classification) and the types of project (i.e. research, system enhancement, competitions, and free topics). *Research* projects aim to reproduce existing research papers relevant to the course. *System enhancement* projects aim to enhance existing relevant open-source software, systems, toolkits, etc. *Competitions* are leaderboard-based DS contests. Students could choose from the above released topics or propose *free topics* relevant to the course.

4. PROJECT DESIGN FOR ASSESSMENT AT SCALE

In this section, we describe how we address two major research questions (RQs) in assessment of data science projects, which have not been well addressed in the existing work: **RQ1:** How to optimally provide high quality and timely scaffolding and feedback at scale? **RQ2:** How to ensure that different types of projects are graded fairly under the same set of rules?

4.1 Timely and high-quality scaffolding and feedback

For **RQ1**, we implemented two solutions: Peer assessment with TAs as meta-reviewers, and multiple checkpoints.

4.1.1 Peer assessment with TAs as meta-reviewers

Peer assessment has been used successfully in prior literature to assess group projects [12, 1] but it may not always be reliable. So, instead of completely relying on peer grading, we provided peer reviews and grades as additional contexts to the TAs during grading. A similar approach was used in prior literature for grading short written assignments [5]. In that case study, students found the feedback given by graders in the context of peer reviews to be better than the feedback given without peer-reviews. Although the study did not observe the overall improvement (or effects) in the grading time compared to grading without peer-reviews, we think it could reduce the time in our case of grading projects, which are generally more time-consuming than grading short written assignments. We chose Microsoft Conference Management Toolkit (CMT) [1] for managing projects because it has several useful features, particularly managing both reviewers and meta-reviewers.

Reviewer and Meta-reviewer assignment: To make the best use of TAs' time and knowledge, each TA was responsible for a group of projects (assigning peer-reviewers, grading, and general guidance) based on their interests. To ensure familiarity with project topics, TAs manually assigned each student to review 1-2 projects that are similar to their own projects. Each project was assigned at least 2 reviewers to get multiple perspectives and for robustness.

4.1.2 Multiple checkpoints

To continuously guide students and provide rapid feedback [6], we divided the project into three stages: 1) initial stage where students were asked to form groups, choose their topics and submit a short proposal on the feasibility, significance, and novelty of the topic and their plan to complete it, 2) midpoint stage to submit a short document on their progress updates, and 3) the final stage to submit completed projects including documented source code and results.

Qualitative and quantitative feedback were provided at all stages of the project, either by TAs or peers. Only the TAs provided qualitative feedback on the initial stage submissions as we anticipated it would be hardest for peers to provide feedback/suggestions on the proposed topics and plans given their limited experience. For the later stages, the peers provided both qualitative and quantitative feedback, which were made available to the TAs for grading. TAs assigned only quantitative feedback on the other two stages for efficiency. The same set of peer-reviewers assigned to a given project reviewed submissions at both the midpoint and final stages so that they would not have to re-familiarize themselves with new projects and could also provide progressive feedback.

4.2 Fairness in grading

To address **RQ2** especially under peer-evaluation settings, we clearly defined the expectations and objective rubrics.

4.2.1 Set expectations and objective rubrics

We made the expected deliverables and grading rubrics clear and consistent across different topic options. All teams were required to submit a proposal at the initial stage, a progress report at the midpoint stage, and three pieces of deliverables at the final stage: well documented source code, main results, and a demo video of the finished project. The peer reviewers give scores by filling out a peer-grading form with a few multiple-choice questions that were consistent across all topic choices and had little ambiguity in how many points to give. In addition to quantitative scores, there were also open-ended questions in the peer-grading form for providing more detailed qualitative feedback in free-form text, which helped the students learn and the TAs grade. The peer-grading forms were released at early stages of the course, so that they not only facilitated the peer review process at the end, but also made it clear at the beginning to all students about what should be done for good grades.

The main quantitative grading rubric components were 1) reproducibility, i.e., whether the reviewers could successfully reproduce the reported results with the submitted codes and documentation, and 2) completion, i.e., how many project requirements are satisfied. This would ensure the grade is assigned objectively. Since source code setup process can be challenging (e.g. due to differences in the programming environments, etc.), we encouraged the peer reviewers and authors to meet online for live demos and discussions.

5. EVALUATION METHODS

We analyzed data from the following sources to evaluate various design strategies that we used.

Web-based surveys: All three roles involved in the peer-review process (authors, peer-reviewers, and TAs) were surveyed about their experience at the end of the course. The surveys asked multiple-choice and qualitative questions about time consumption, experience with the peer-review tool

[1] https://cmt3.research.microsoft.com/

(CMT), and feedback on the overall process. The first author of this paper performed open coding on the responses to qualitative questions.

149 students participated in the author experience survey, including 36 undergraduates, 102 DSGs, and 11 other graduate students, and 154 participated in the Reviewer experience survey. Mostly the same set of students participated in both those surveys. Five TAs participated in the TA experience survey.

Peer reviews: We downloaded the 506 peer-reviews (339 by DSGs, 125 by undergraduate students and 42 by other graduate students) submitted for the 204 group projects at both the midpoint and completion stages.

6. FINDINGS

We now discuss our findings from the data related to the time and effort taken, fairness, scaffolding and peer-review feedback to students using our implemented solutions.

6.1 TA and peer-reviewer effort

On average, TAs took much less time to evaluate each project in the final stage compared to the students (around 10 min vs. 1-2 hr). This time difference could partly be attributed to the TAs being more experienced. However, because all the five TAs reported they found peer-reviews to be helpful (ratings >=4 on a 5-point Likert scale), the peer-reviews could have helped reduce TAs' grading time perhaps by suggesting how deeply the TAs should examine a project (e.g., TAs could trust the reported results if all reviewers give high reproducibility scores, and may need to test the codes only if inconsistency between the code and result is raised by a reviewer or if reviewers disagree). A more thorough investigation is needed in future to verify this.

Evaluating reproducibility was also challenging, tiring and time-consuming especially when the authors submitted insufficient documentation, or the peer reviewers did not have appropriate computational resources. As one student reported, *"... One project assigned to me [...] had minimal instructions or made assumptions about the reviewer's OS. This made it difficult to test..."*. This issue can be alleviated to some extent by enforcing constraints, e.g. operating systems, programming languages.

TAs also had to spend extra time on manually assigning peer-reviewers due to lack of support for automatic assignment based on project similarity in CMT. Three out of the five TAs reported taking 1-2 hr. for this step, 2 reported 30-60 min.

6.2 Pedagogical benefits

Scaffolding via rapid feedback: Students reported that receiving timely peer-reviews (mainly qualitative feedback), especially at the midpoint stage, were useful for project improvement. For example, one peer-reviewer gave detailed suggestions on some methodologies to try: *"... you may be able to seed/start (as a prior distribution?) your 'is a directory' classifier ..."*, which the author of the project found to be very helpful. Even in cases where the peer-reviews did not provide any suggestions for improvement, students liked having another set of eyes to validate their work.

Obviously, the quality of the feedback varies and may depend on many factors like peer-reviewers' assessment skills, motivation, etc. To get an idea of whether the academic level and degree program has any impact on the feedback quality, we compared the reviews submitted by the three groups of students (DSG, undergraduates and other graduates). Since length of the peer-reviews can give an indication of their quality (i.e., longer reviews typically provide more useful feedback), we used it as a proxy for quality while comparing the reviews. In future, we plan to perform a more thorough analysis of review quality [3]. Based on Wilcoxon Rank Sum Test [8], we found that peer-reviews submitted at the mid-point stage by DSG students ($M = 22.58, SD = 24.55, n = 339$) were significantly longer than those submitted by other students ($M = 18.01, SD = 22.65, n = 167$), $z = 2.86, p = .004$. Similar results were obtained for the final stage $p < .0001$. Since the DSG students are working professionals, they might be more experienced in reviewing and thus gave better feedback.

Surprisingly, some students preferred getting feedback from students instead of TAs. They reported it created a low-stress environment to receive early feedback by someone with similar levels of expertise who would also be able to empathize with them, e.g., *"I like having my peers appreciate the effort I put in as TAs understand the subject area a lot more and so the complexities might seem simple to them"*.

Collaboration and social interactions: The peer-reviews sometimes initiated collaboration across groups. For example, a peer-reviewer shared their own project to potentially solve a challenge faced by the author, *"...my project is to deliver a portable environment for running [toolkit] using [containers], which might [..] assist your work..."*.

Moreover, a few students also appreciated the opportunity to interact with their classmates during the peer-reviews, *"... Being an online student, I got the feeling that I was in the classroom especially when looking at others presenting ..."*.

6.3 Fairness

From survey results on fairness of the scores assigned by peer-reviewers, 145 students reported they felt scores were Fair, 4 said Too Low and 1 said Too High. This suggests that most students found their scores to be fair.

To check for consistency among the scores across different project types, we performed Wilcoxon Rank Sum tests on pairs of TA-assigned final scores of the 4 project types (i.e. competitions, systems, research, and free topics) and found insufficient evidence of the scores being different. This suggests that the rubrics were also helpful for the TAs to grade various project types objectively. As reported by one TA, *"...each track [...] had clear submission and grading instructions so as not to leave anything to the grader's judgement"*.

7. DISCUSSION

From the case study, we found that the strategies leveraged for addressing the challenges of DS group projects assessment at scale were mostly useful but their implementation could be further improved. The analysis also helped us identify new opportunities mainly from the large scale and diversity of such courses. Below, we discuss the implications of our case study.

Need for better peer-assessment tools for DS group projects: From 6.1, reproducibility-centered assessment with many project options could make assessment harder. So, it could be useful to add uniformity constraints (e.g., programming languages) while still providing flexibility in project topics to achieve both individualized learning and unified assessment. Automatic reviewer assignment based on matching project topics should also be supported to save time.

Multi-stage and Multi-tier assessment: From 6.2, qualitative peer-assessments at multiple project stages can be useful to provide timely suggestions and validation/appreciation of student projects at scale. But students sometimes lack the expertise and motivation to assess the projects, especially in case of performing subjective evaluation (e.g., evaluating the quality of a project proposal), where TAs can offer more assistance. Thus, both students and TAs can offer different tiers of assessment support. Moreover, the more experienced/motivated students in large courses (e.g., working professionals) could act as another tier between TAs and the less experienced students for providing for providing high quality feedback and guiding more inexperienced students.

Multi-tier student collaboration via peer review: From 6.2, it is possible to develop more sub-communities or tiers of collaboration outside of the individual groups via peer review. For example, groups working on similar project topics could help each other understand their topics better or offer suggestions. Students could also collaboratively develop complementary projects which collectively contribute towards a large project, similar to synergistic modular assignments proposed in [2]. Peer-reviewing could also encourage more interaction with classmates in online courses, where students are known to feel isolated [11].

Clear and objective rubrics: It is crucial to provide clear and objective rubrics that ensure fair and consistent quantitative scores across project topics and graders. Our analysis in 6.3 shows that it is feasible to design such rubrics.

Limitations: We acknowledge that many of the findings come from self-reported survey responses, which may not always be reliable. Moreover, since this is a case study, the findings may not generalize to all courses.

8. CONCLUSION

In this paper, we discussed our experience with deploying strategies (e.g., peer-reviewing with TAs as meta-reviewers, rapid feedback) for addressing two challenges with assessing data science group projects at scale: fairness, and timely and high-quality feedback and scaffolding. By analyzing student /TA survey responses and peer reviews, we found that those strategies are both feasible and useful for scaling up group project assessment. In the future, we would like to develop better infrastructure for supporting the various mechanisms we used. Further, it would be useful to investigate the benefits of leveraging the diversity and scale in such courses by developing sub-communities that can collaborate with each other in multiple ways, e.g., for multi-tiered peer assessment.

REFERENCES

[1] Gabriel Badea and Elvira Popescu. 2019. Using LearnEval Peer Assessment Platform in Project-Based Learning Settings: A First Experience Report. In *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 1–5.

[2] Bhavya, Assma Boughoula, Aaron Green, and ChengXiang Zhai. 2020. Collective development of large scale data science products via modularized assignments: An experience report. In *Proceedings of the 51st ACM SIGCSE*. 1200–1206.

[3] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI*. 1–13.

[4] Chase Geigle, Ismini Lourentzou, Hari Sundaram, and ChengXiang Zhai. 2018. CLaDS: a cloud-based virtual lab for the delivery of scalable hands-on assignments for practical data science education. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. 176–181.

[5] David A Joyner, Wade Ashby, Liam Irish, Yeeling Lam, Jacob Langston, Isabel Lupiani, Mike Lustig, Paige Pettoruto, Dana Sheahen, Angela Smiley, and others. 2016. Graders as meta-reviewers: Simultaneously scaling and improving expert evaluation for large online classrooms. In *Proceedings of the Third (2016) ACM L@S*. 399–408.

[6] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. *ACM TOCHI* 20, 6 (2013), 1–31.

[7] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM L@S*. 75–84.

[8] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[9] Bina Ramamurthy. 2016. A practical and sustainable model for learning and teaching data science. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. 169–174.

[10] Manikandan Ravikiran. 2020. Systematic Review of Approaches to Improve Peer Assessment at Scale. *arXiv preprint arXiv:2001.10617* (2020).

[11] Alfred P Rovai and Mervyn J Wighting. 2005. Feelings of alienation and community among higher education students in a virtual classroom. *The Internet and higher education* 8, 2 (2005), 97–110.

[12] Maya Usher and Miri Barak. 2018. Peer assessment in a project-based engineering course: comparing between on-campus and online learning environments. *Assessment & Evaluation in Higher Education* 43, 5 (2018), 745–759.