Teaching Assistants as Assessors: An Experience Based Narrative

Faizan Ahmed, Nacir Bouali, and Marcus Gerhold
Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

Keywords: Teaching Assistants, Grading Consistency, Grading Variation, Reliability.

Abstract: This study explores the role of teaching assistants (TAs) as assessors in a university's computer science program. It examines the challenges and implications of TAs in grading, with a focus on their expertise and grading consistency. The paper analyzes grading experiences in various exam settings and investigates the impact on assessment quality. We adopt an empirical methodology and answer the research question by analyzing the data from two exams. The chosen exams have similar learning objectives but they differ in how TAs graded them, thus providing an opportunity to reflect on different grading styles. It concludes with recommendations for enhancing TA grading effectiveness, emphasizing the need for detailed rubrics, training, and monitoring to ensure fair and reliable assessment in higher education.

1 INTRODUCTION

The computer science program at our university has seen massive growth in recent years. The change posed a logistical challenge for the examiners and called for hiring more lecturers. To provide education at scale, support from teaching assistants (TAs) became essential. Teaching assistants are hired for all quarters [1], but a majority of them are hired to help with the first-year courses since those have a higher number of students, and also, the courses are not very advanced.

The population groups of teaching assistants are diverse in their education levels. The following is a categorization of teaching assistants: a) Undergraduate students (UTA): These are usually year two or year three students. In most cases, they have followed the same courses they are assisting. b) Graduate students (GTA): These are master's degree students. c) Ph.D. Candidate (Ph.D.): Pursuing their Ph.D. and may not have followed the same courses they are assisting in. In most universities, Ph.D. candidates are also considered graduate teaching assistants (Wald and Harland, 2020). However, their job contract at our university categorizes them as an employee. Therefore, we have grouped them separately.

[1] At our university, the academic year is organized into a quarter system, comprising four quarters each lasting 10-11 weeks. The complete curriculum is detailed on our website: https://www.utwente.nl/en/tcs/education-programme/tcscurriculum/

Teaching assistants have various roles depending on the context in which they are hired. Some of their roles include tutoring, assisting the teacher in the classroom, leading small student groups, preparing assignments, auditing assignment descriptions, grading assignments, exams and projects, and providing general administrative support. Kerry et al. have provided a typology of various roles for teaching assistants (Kerry, 2005).

TAs are helpful in scaling the program size. To maintain quality of education, their various roles must be carefully evaluated and intervened by providing necessary training and monitoring (Wald and Harland, 2020). Especially so when TAs are used to help with assessment. The use of teaching assistants as assessors in higher education can potentially impact the quality of assessment in a number of ways. On the plus side, teaching assistants might add new ideas to the assessment process and may have a fresher perspective on students' needs and aptitudes. With smaller class sizes and more time to spend on grading and evaluation, they might also be able to provide students feedback that is more personalized. However, employing teaching assistants as assessors could have certain disadvantages as well. As a result, the validity and impartiality of their evaluations may be impacted. Teaching assistants might not have the same amount of training or experience as full-time faculty members. Additionally, they can lack the education or assistance needed to accurately assess students' work and offer helpful feedback.

Ahmed, F., Bouali, N. and Gerhold, M.
Teaching Assistants as Assessors: An Experience Based Narrative.
DOI: 10.5220/0012624200003693
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 1*, pages 115-123
ISBN: 978-989-758-697-2; ISSN: 2184-5026
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

It is of paramount importance to discuss and investigate the impact of using teaching assistants for assessment. Therefore in this paper we share our experience with respect to the question "How does the deployment of teaching assistants as assessors impact the assessment quality in higher education?". We also elaborate on the questions a) Which factors, such as transparency, reliability, and validity, are most affected by the deployment of teaching assistants as assessors? b) What are good practices in deploying teaching assistants for assessments?

In this paper, we focus on detailing experiences and analysis of example cases to emphasize the need for a rigorous scientific investigation of the problem. Our focus is on two factors mainly: expertise and consistency. The methodology includes the analysis of grading data from exams, employing statistical tests such as two-sample t-tests to assess grading consistency between TAs, and visualising them via box plots. The methodology also encompasses a review of grading settings, including digital exam environments and grading parties, to understand how different setups affect grading outcomes.

Paper Organization. In Section 2, we look at the related work on TAs' deployment as graders. In Section 3 we provide the details of the exam chosen and substantiate our choices. In Section 4 we briefly discuss the digital exam environment we used and the corresponding grading setup, while Section 5 provides an analysis to evaluate the consistency of the grading and also the impact of a TA's subject expertise on the grades. Finally, we elaborate on our experience and give some recommendations on improving the TAs' quality and consistency of the grading in Section 6. Concluding remarks are given in Section 7.

2 BACKGROUND

Although teaching assistants have been used for a very long time as graders in higher education (see e.g. Svinicki (1989)), their role is not very well discussed in the literature. Rather, more emphasis is given to the categorization of their various roles, evaluating their suitability for a TAship or to the TA hiring practices and design of TA training (Liggett, 1986). Another aspect that received interest is the effectiveness of a TA as a teacher, a demonstrator, and a tutor, as well as the use of the TAs to prepare and audit educational material (Minnes et al., 2018). In most cases, various experiences or research works can be found for the use of the TAs to evaluate laboratory exercises or assignments (see, e.g., Alvarado et al. (2017); Pickering and Kolks (1976)).

A systematic literature review on using teaching assistants in computer science was provided by Mirza et al. (2019). They have covered the literature on the different roles of teaching assistants and other related topics. It is clear from their paper that only a few researchers have reported on the use of TAs as graders.

Although there is a shortage of literature on this topic from computer science education, in other education programs, researchers have shared their experiences. For example, Liggett (1986) have reported on evaluating the reliability of grading from the teaching assistants for mechanical engineering. They computed the reliability by using various grading settings. Their primary purpose was to compute the effectiveness of the TA training on the grading reliability. Similarly, Marshman et al. (2018) have reported using TAs to grade introductory physics courses. They reported on the importance of rubrics to assess the students' work. They also emphasized on the need for training of teaching assistants.

Suitability of TAs as graders is critically discussed by Hogan and Norcross (2012) (see also Wald and Harland (2020)) from a domain independent perspective. They have categorized assessments into two types when it comes to deploying teaching assistants as graders: 1. assessing factual information that requires recall, and 2. assessing items that require interpretative judgments. Instead of providing a clear guideline they explain the advantages of using graduate teaching assistants for assessing the latter.

It is a common practice to use undergraduate TAs to grade various assessments. They are primarily used for assessing assignments. Dickson (2011) questions the ethics of using UTAs for grading. They proceed to describe their own experience with using UTAs in grading qualitative assignments. Dickson argues that since the undergraduate students have gone through the exercises more recently than the experienced professor, they can provide much more helpful feedback (Dickson, 2011).

Alvarado et al. (2017) also reported on the use of TAs for grading the assignments in the context of micro-classes. In a related study the authors reported on the usage of undergraduate TAs for student-facing activities while graduate TAs helped with grading (Minnes et al., 2018). Also, van Dam (2018) reflected on using undergraduate TAs for grading an introductory computer science course. The TAs were also used to provide feedback on design choices using a detailed grading rubric.

Maintaining consistency in large courses where multiple graders are involved is a challenging task.

The differences can be attributed to various reasons. Kates et al. (2022) have attributed the inconsistency to interpersonal comparability.

In the context of TA training and mentoring, Lanziner et al. (2017) have reported on the experience of TAs assuming a different role. According to the study, TAs find their roles as graders more challenging and appreciate the training and guidance to perform their job as assessors. Similarly, Riese and Kann (2020) presented TAs experience in their different roles in computer science education. They also emphasize the need for clear and concise grading criteria to help TAs grade students' work. Doe et al. (2013) argue that providing a rubric is insufficient for consistent and effective grading. They also question the accuracy, consistency, and effectiveness of grading from faculty members in the Psychology discipline.

3 THE EXAM GRADING EXPERIENCE

We have selected two exams that are part of the first-year courses. The selected course is rather subjective and consists mostly of open questions. We selected this course since grading other courses' exams with objective answers is deemed easy (Wald and Harland, 2020). Most questions required some level of interpretive judgment. The course was designed for first-year computer science (BCS) and Business Information Technology (BIT) students. The course was split in 2020 due to administrative reasons. The learning objectives are also adapted to cater better to the needs of the respective study program. While the BCS program (see Subsection 3.1) created a new examination format, the BIT program kept the same exam (see Subsection 3.2) structure.

We hire a number of TAs to help us grading assignments and the written exams for these courses. At our institute, Ph.D. candidates are generally considered tutors/lecturers. Predominantly, our Graduate Teaching Assistants (GTAs) in computer science are alumni, mainly due to limited interest from other students who are hesitant to undertake TA roles without prior experience with specific assignments. Consequently, our TA pool is largely composed of Undergraduate Teaching Assistants (UTAs), who form the majority of our grading TAs.

3.1 Exam A

The exam was written in the context of a software system design course for students following a bachelor's degree in Computer Science (BCS). The learning objectives of the course are two-fold: 1. Students are asked to specify existing small software systems in the Unified Modelling Language (UML) before they 2. define a new software system in UML. Additionally, 3. students are asked to identify and explain common phases in software engineering, and 4. evaluate the code base of software systems by means of software metrics and software smells.

In the exam students are provided small UML models that describe a certain context, e.g., patients booking an appointment with hospital doctors by using an IT-system, or lecture room allocation in universities via a scheduling system. These diagrams contain both semantic and syntactic mistakes, and students are asked to point them out and correct them. Their proficiency in the phases of software engineering and software metrics is assessed by open questions, e.g., a question could be "What are the consequences of missing the *requirements elicitation* phase?". Generally, it is challenging to give an all-encompassing answer-key that covers all possible answers. A challenge of equal proportion is to communicate the wide range of possible answers to the ca. 20 grading TAs. Below we summarize the exams of the last three years to better illustrate our experiences in grading with TAs since the redesign of the course in 2020:

2020-2021. The exam had a total of 13 question comprising 100 points total. Two questions asked students to spot syntactical and semantical issues with provided UML diagrams. Additionally, one question presented them a scenario accompanied by three UML diagrams. Students were asked to check consistency between each component and with the scenario overall. The remainder were open questions.

2021-2022. The exam had a total of 12 questions comprising 60 points total. The reason for the vast decrease in the number of points was twofold: 1. to give a better indication to students how verbose their answer should be in open questions, and 2. to give stricter indication to TAs when and when not to give points. This was a conscious choice because most TAs found *some* merit to the answer of open questions. Decreasing the total points was intended reduce that effect. In addition, we provided stronger numerical guidelines in the tasks, e.g., "Name at least three advantages [...]", or "[...] briefly describe in two to three sentences." Three questions provided UML models to spot syntax and semantics mistakes. The remainder were open questions.

2022-2023. The exam had a total of 11 questions that comprise 60 points total. There were three questions that asked for UML syntax mistakes to be corrected, while the remaining eight were open questions. For the first time we tested the students' proficiency in UML syntax drawing by using a drawing tool provided by our university's e-assessment platform. The tool was a generic drawing tool, not specified in UML syntax and students had the chance to test this tool in an ungraded *mock exam*. Hence, while providing means to assess UML proficiency, this question introduced a new level of subjectivity, e.g., by comparing hand-drawn diamonds for aggregation/association relations in class diagrams to hand-drawn arrow-heads for generalisation relations.

In this paper, we have only included the statistical analysis of the most recent exam.

3.2 Exam B

The exams were given in the context of a software design course for students pursuing a Business and Information Technology (BIT) bachelor degree. The course covers two main axes, one on low-level design using Unified Modelling Language (UML) as a notation, and another on software maintenance and metrics.

2020-2021. The students were given a case study and were asked to provide the activity diagrams, the use case model, the class diagram and the state machine, as well as answering a question on software complexity. The exam was then graded by 9 teaching assistants. Each two were responsible for grading one diagram and one TA was assigned the complexity part.

2021-2022. The number of diagrams was reduced to three instead of four, to allow the students more time to draw properly on a computer. The exam was then graded by 4 TAs, each handling one diagram. A decision that was made to make the grading consistent across each question.

2022-2023. The exam has seen a shift in its structure, as we decided to reduce the number of diagrams the students draw to one, the class diagram. The remaining questions were shifted to providing student with faulty designs and asking them to fix them. Every question is then assigned to one TA for grading.

The side effect for the grading is that the answers to the questions are now fixed, so the teaching assistants have a better grading key with which they can compare the answers. This change was actually felt during the review session, as most regrading requests were concerned with the class diagram question, in which the students had to design a class diagram from scratch. Very few students have requested a regrading of the other questions.

4 EXAM TOOL AND GRADING SETTINGS

The students took each exam using a digital environment. The digital environment provides functionalities to ask multiple-choice and open-ended questions. It also provides functionalities to ask short questions (like fill in the blanks) that can be automatically graded. For both programs the questions of the exams are open-ended. In addition to its own (non-specific) drawing tool, the digital environment also allows for *external* drawing tools which, more or less, enforce the UML meta-model, and allow the students to draft their UML diagrams with ease.

It also provides the functionality to grade the students work simultaneously. Course teachers can embed grading rubrics within the questions, and students' names are anonymous to reduce bias. The digital environment also provides statistics related to the graded exams, for example, pass rate, average grade per question, per-question exam analysis, and more. Through a log, grade changes and changers are traceable. However, this information is neither downloadable nor used to compute statistics. We have manually copied this information for analysis in the next section.

The number of students in BCS and BIT programs differ significantly. Therefore, more TAs are hired to grade the BCS exam than the BIT exam. These TAs also helped during the lab sessions and have direct contact with students.

4.1 Grading Party

The BCS exam used a grading party. Grading TAs and teachers gathered in a room. The session starts with an explanation of the exam and the grading rubric. The teacher also explains the possible variations in the answers. The grading work is then divided among the TAs based on their preferences. Due to a large number of students, at least two TAs are assigned to grade a question. For questions requiring more time to grade, more TAs were assigned. The following process was adopted for the grading party:

1. TAs attend a grading session with lecturing staff

present. They are presented the answer key and can ask questions.

2. After the grading session, a faculty staff member closely inspects a random sample of student exams. Since the TA grading is done horizontally (i.e., per question), then sampling student exams vertically (i.e., per student) provides a wide-ranging overview, and
3. An exam review is arranged in which students can flag the potential misgrading that happened during the grading session.

4.2 Grading Individually

The BIT exam uses the other approach. The teacher embedded the rubric in the digital environment and provided an extended explanation separately. The TAs choose the part they like to grade, and the grading work is then assigned. In this case, the TAs work individually, and the teacher monitors the grading progress remotely. Since the number of students in this course is small, typically only one teaching assistant grades each question. However, for certain questions, multiple TAs are involved in the grading.

5 AN ANALYSIS OF EXAM GRADES

Our analysis of the exam grades is focused on capturing variations in the grades given by a TA. We focus on capturing inconsistencies in grades assigned by different TAs. Also, we investigate the differences in the grading patterns between UTA and GTA. Some TAs have been working for the program for a longer time. They were UTAs, and are now GTAs. We reflect on their learning curve to grade exams in the discussion section. Due to the similarity of courses, some TAs worked for both programs (BCS and BIT).

5.1 Exam A

The BCS exam was graded by 22 TAs and three teachers. Out of 22 TAs 15 were UTA while 7 were GTA. After that, the questions were divided among TAs based on their preferences. At least two TAs graded each question (to emphasize, the set of students each TA was grading was different). Some questions had more tasks and required more verbose answers than others. Consequently, more TAs are needed to grade them to guarantee that the grading party can be finished in a feasible time. As soon as a TA is done

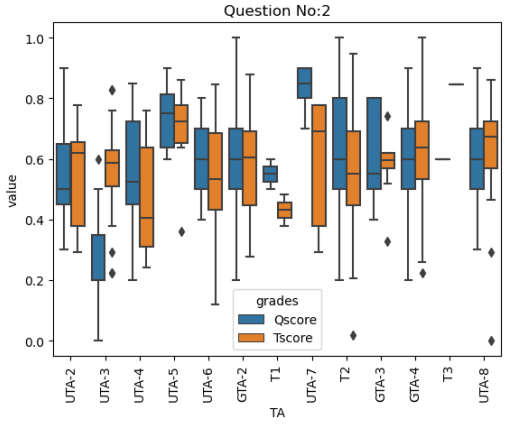

Figure 1: Plot of score obtained by students for a single question (blue box plot) versus their total score obtained in the exam (orange box plot). Both scores are scaled between 0 and 1.

grading a question and still willing to do more grading, a new question is assigned to them.

Since we wanted to capture variation in grading, we adopted the box plot. Box plots are handy for visualizing variations. We created one box plot for each TA and combined a box plot for the same questions in one figure (cf. Figure 1). TA names are removed to maintain privacy. For the questions that are graded by two TAs, we have also applied a two-sample t-test to find out whether the difference in mean grading is statistically significant or not. Note that the test is only suitable for pairwise comparison. Therefore, we have applied it only to questions that, at most, two teaching assistants grade.

The results of the tests show that the difference in mean grades is not statistically significant. We conclude that the TAs graded questions consistently. However, when a question is graded by more than two TAs, the variation is higher. Since we cannot judge significance using a t-test, we opted to compare it with the total grade. The implicit assumption is that if, on average, a TAs is giving a lower (or higher grade respectively) while the total grade has an opposite trend, then the TA might be grading harshly (or generously respectively). We emphasize that before drawing any conclusions, a sample must be checked by the teachers to verify the claim.

To illustrate, we provide one example. Figure 1 shows a box graph for one question since the question has only 6 points while the total score is 10. To compare both graphs, we scaled both between 0 and 1. As can be seen from Figure 1 UTA4 has assigned lower grade for this question, while for the same student population, their overall score is much higher. Thus, this grade section requires further investigation either by a teacher or another TA. Another interest-

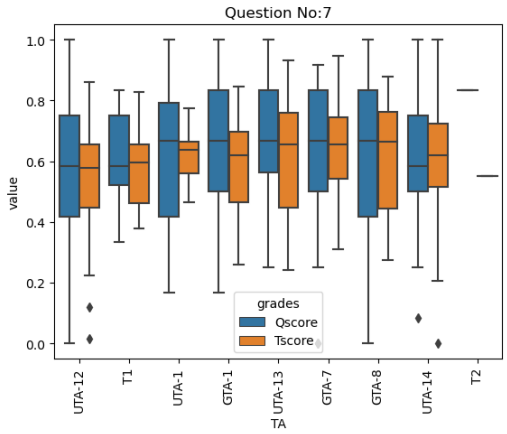

Figure 2: The plot shows grades obtained by students for question 7 (blue box plot) and the total score obtained by a student (orange box plot). Both scores are scaled between 0 and 1.

ing observation is for T3, which represents a teacher with only one data point. It might be the case that this question was a borderline case (i.e. the difference for a student to pass or fail the exam), and the teacher decided to review and regrade.

We also like to investigate the claim that GTAs are better in grading exams since they have higher qualification. We were not able to identify any significant difference for the questions shown in Figure 1. Figure 2 shows grading for another question. It is again an open-ended subjective question. Here, too, no significance difference is observed between GTA and UTA.

5.2 Exam B

The teacher has a more personal approach to hiring TAs and relies on a small team of experienced and proven teaching assistants. Since the number of students lies between 55-120 for this course, the grading work is significantly smaller than for the BCS course. We analyze the exams for three years to capture the grading practice for a TA.

For most of these exams, there is no significant difference in grading from different TAs. Figure 4 shows a box plot for a question that appeared in the 2021 exam. The question is related to the activity diagram and is accompanied by a detailed rubric that converts an open-ended grading task into a binary grading (see Figure 3). Besides the clear rubric, there is a difference in grading from different TAs. TA18 gives lower grades, while TA 7 has higher variation in their grading.

Throughout our exams, we have tried to assign specific teaching assistants to be involved in grading over the years. Through such practice, we have noticed that a senior TA (one who has been grading the same type of exam more than once) is more likely to handle answers which are not covered by the provided rubric. Senior TAs are, however, more likely to make mistakes while grading open questions, as their judgement more than often falls short into capturing the myriad of forms that the mistakes can take. A different problem is created by junior TAs who usually stick to the provided rubric, as these feel more constrained to follow the grading key to the letter, subsequently failing to capture the partially correct answers.

6 DISCUSSION & RECOMMENDATIONS

To guarantee fair yet feasible assessment, TAs have become indispensable. Hence, consistency of grading within the TA corpus is of crucial importance.

We propose that TA grading consistency by means of statistical analysis becomes a default in this multi-step process. Item analysis of exam questions is common practice. This holds true for easily-assessed multiple-choice questions, as well as open questions. Analysis involves critical characteristics such as Kronbach's alpha for multiple assessors grading one item, the R_{ir} value relating the difficulty of individual questions to the overall exam, or the average score that participants got per question commonly indicated by p'.

Our university is utilizing a digital examination system for most of its exams. This system collects a vast amount of data, offering insights into grades and the grading process. Although it provides various exam-related statistics, it currently lacks the capability to generate statistics about the graders. A recent update allows examiners to review the grading history for specific exam items, including who graded them and any changes made, but this review is limited to an individual question and student basis. For a comprehensive analysis, such as calculating inter and intra-rater variation in grading, exporting this log would be beneficial. To enhance grading consistency, several additional features are suggested for the digital examination system. These include:

- Tracking the percentage of exams graded by each teacher to understand individual grading loads.
- Counting the number of personnel involved in grading each question to ensure adequate coverage and diversity.
- Implementing visualization tools, such as bar

Correction criterion	Points
0.5 points for each correct swimlane: hotel staff, customer, company contact, system clock. (Some students split the hotel staff onto receptionist and remaining staff, this should count as okay as long as the activities are distributed correctly).	2 points
For the correct use of one initial node and 3 final nodes. (0.5 points each)	2 points
activity: book online. (customer)	1 point
activity: ask at the reception. (customer)	1 point
activity: cancel online. (customer)	1 point
activity: call hotel. (customer)	1 point
activity: cancel reservation. (hotel staff)	1 point
activity: register customer (hotel staff)	1 point
activity: check in customer (hotel staff)	1 point
activity: register service. (hotel staff)	1 point
activity: check out. (customer)	1 point
activity: check out customer. (hotel staff)	1 point
activity: print bill. (hotel staff)	1 point
activity: pay bill. (customer)	1 point
activity: pay company bill. (company contact)	1 point
activity: undo booking. (clock)	1 point
activity: charge one night to credit card. (clock)	1 point
activity: send bill. (clock)	1 point
activity: check in (customer)	1 point
Correct use of the decision nodes.	2 points
Correct use of fork/join.	2 points
Correct use of merge nodes.	1 point
Total points:	*26 points*

Figure 3: Grading rubric for grading activity diagram.

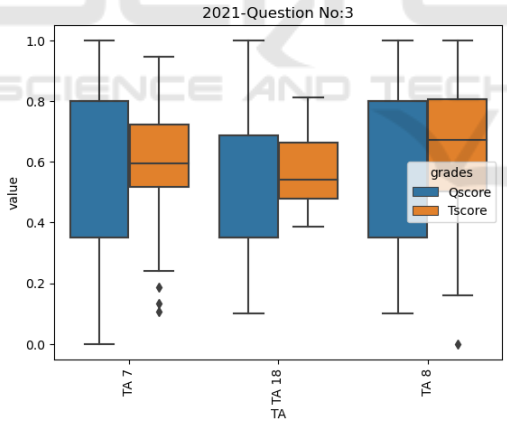

Figure 4: The plot shows grades obtained by students for the activity diagram question (blue box plot) and the total score obtained by a student (orange box plot). Both scores are scaled between 0 and 1.

graphs, for analyzing grading variation and identifying outliers.

- Calculating statistical significance parameters to evaluate variation in grading across different graders and identify patterns of inconsistency.
- Monitoring the time spent grading each question to pinpoint potentially problematic questions that require an inordinate amount of grading time.
- Analyzing trends in grading over time to detect deviations or inconsistencies in grading patterns.
- Establishing a feedback mechanism for teaching assistants to report uncertainties or ambiguities encountered during grading, facilitating clarification and consistency.

TA evaluation should be standardized with similar scrutiny to prevent outliers. This step facilitates the ease in which commonly re-occurring themes can be checked: 1. Do UTAs grade too strictly whether that be due to their own insecurities, or perception that the questions were graded equally harshly when they took the course? 2. Do GTAs grade too leniently because they find *some* merit in each answer? The analysis shown in our work should become a mainstay in e-assessment tools. It can support lecturing staff twofold. First, live monitoring of these statistics lets them address issues on-the-fly. Secondly, *a posterior* analysis aids in the sampling step for manual inspection. Lastly, it helps in coaching and fostering assessment proficiency in TAs, since it enables lecturing staff to address harshness/leniency. This expedites the growth of a network of TAs proficient in fair and fast grading, since they work in pairs and can learn from each other. However, we do emphasize that this should not be the only safeguard for quality, but rather a supplement in a multi-faceted workflow.

We believe that using TAs in grading can only be *reliable* by forcing the students to participate in the review. This allows the misgraded students to flag their cases and allows the teachers to regrade the concerned copies ensuring a fair examination to everyone.

Based on our experience we recommend the following practices for deploying teaching assistants to grade exams:

- If teaching assistants are used as assessors, it is important that they are clear about their expectations and that the assessment process is transparent to students. This could be achieved through providing clear grading rubrics, giving feedback on assignments, and being available to answer questions about the assessment process.
- For grading exams with many students, it is advisable to arrange a grading party where active discussion among TAs is supported and appreciated. It is essential that the teaching staff also actively participates and grades a portion of the exams. It ensures a higher grading accuracy, provides a better foundation for guidance to TAs and actively promotes a deeper understanding of assessment practices.
- To increase the consistency in grading, assign one question per TA. The downside could be that the TA grades either generously or too harshly – But at the very least does so consistently. The risk could be mitigated by teaching staff by looking at the average grade per question. Extreme cases are easily identifiable.
- The need for close monitoring cannot be emphasized enough. The teaching staff must look for clues to intervene and adjust the grading practice. They must grade the borderline cases. Furthermore, they must be creative in creating visualizations to capture variations in the grading and identify anomalies.
- Work towards developing a team of TAs that help with grading. The team must consist of experienced TAs (who have also helped grade the same course in previous years) and junior TAs. We emphasize that including junior TAs is essential for continuity. Pay extra attention to junior TAs and be vigilant about senior TAs. A well-trained UTA can grade more consistently and reliably than an untrained GTA. Therefore, groom TAs so that they can perform better.
- Motivate students to participate in the review and explain the grading process to them for increased transparency. We recommend that the teaching staff must conduct the review and review must be as accessible as possible for students. If mistakes are spotted during review, audit all exams graded by the TA for whom a mistake was spotted.

7 CONCLUSIONS

In the introduction, we posed the question, "How does the deployment of teaching assistants as assessors impact the assessment quality in higher education?". The paper was not meant to answer the question but rather provide an experience-based narrative. Our experience indicates that using TAs for grading is a delicate process that leads to low grading quality that can be significantly improved by close monitoring and intervention from the teaching staff. The experiences and analysis described in this paper do not provide a conclusive answer; instead, they emphasize the need for a more carefully designed scientific study to identify impacting factors and corresponding mitigation strategies.

REFERENCES

Alvarado, C., Minnes, M., and Porter, L. (2017). Microclasses: A structure for improving student experience in large classes. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 21–26.

Dickson, P. E. (2011). Using undergraduate teaching assistants in a small college environment. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, pages 75–80.

Doe, S. R., Gingerich, K. J., and Richards, T. L. (2013). An evaluation of grading and instructional feedback skills of graduate teaching assistants in introductory psychology. *Teaching of Psychology*, 40(4):274–280.

Hogan, T. P. and Norcross, J. C. (2012). Undergraduates as teaching assistants. *Effective college and university teaching: Strategies and tactics for the new professoriate*, page 197.

Kates, S., Paulsen, T., Yntiso, S., and Tucker, J. A. (2022). Bridging the grade gap: Reducing assessment bias in a multi-grader class. *Political Analysis*, page 1–9.

Kerry, T. (2005). Towards a typology for conceptualizing the roles of teaching assistants. *Educational Review*, 57(3):373–384.

Lanziner, N., Smith, H., and Waller, D. (2017). Reflections from teaching assistants in combined learning assistant and course grader roles. *Proceedings of the Canadian Engineering Education Association (CEEA)*.

Liggett, S. L. (1986). Learning to grade papers.

Marshman, E., Sayer, R., Henderson, C., Yerushalmi, E., and Singh, C. (2018). The challenges of changing teaching assistants' grading practices: Requiring students to show evidence of understanding. *Canadian Journal of Physics*, 96(4):420–437.

Minnes, M., Alvarado, C., and Porter, L. (2018). Lightweight techniques to support students in large classes. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 122–127.

Mirza, D., Conrad, P. T., Lloyd, C., Matni, Z., and Gatin, A. (2019). Undergraduate teaching assistants in computer science: a systematic literature review. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*, pages 31–40.

Pickering, M. and Kolks, G. (1976). Can teaching assistants grade lab technique? *Journal of Chemical Education*, 53(5):313.

Riese, E. and Kann, V. (2020). Teaching assistants' experiences of tutoring and assessing in computer science education. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.

Svinicki, M. D. (1989). The development of TAs: Preparing for the future while enhancing the present. *New directions for teaching and learning*.

van Dam, A. (2018). Reflections on an introductory CS course, CS15, at brown university. *ACM Inroads*, 9(4):58–62.

Wald, N. and Harland, T. (2020). Rethinking the teaching roles and assessment responsibilities of student teaching assistants. *Journal of Further and Higher Education*, 44(1):43–53.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS