Grading 600+ students: A Case Study on Peer and Self Grading

Maurício Aniche
Delft University of Technology
Delft, The Netherlands
m.f.aniche@tudelft.nl

Frank Mulder
Delft University of Technology
Delft, The Netherlands
f.mulder@tudelft.nl

Felienne Hermans
Leiden Institute of Advanced Computer Science
Leiden, The Netherlands
f.f.j.hermans@liacs.leidenuniv.nl

***Abstract*—Grading large classes has become a challenging and expensive task for many universities. The Delft University of Technology (TU Delft), located in the Netherlands, has observed a large increase in student numbers over the past few years. Given the large growth of the student population, grading all the submissions results in high costs.**

We made use of self and peer grading in the 2018–2019 edition of our software testing course. Students worked in teams of two, and self and peer graded three assignments in our course. We ended up with 906 self and peer graded submissions, which we compared to 248 submissions that were graded by our TAs. In this paper, we report on the differences we observed between self, peer, and TA grading.

Our findings show that: (i) self grades tend to be 8–10% higher than peer grades on average, (ii) peer grades seem to be a good approximator of TA grades; in cases where self and peer grade differ significantly, the TA grade seems to lie in between, and (iii) the gender and the nationality of the student do not seem to affect self and peer grading.

***Index Terms*—computer science education, software engineering education, software testing education, peer grading.**

I. Introduction

Grading large classes has become a challenging and expensive task for many universities. The Delft University of Technology (TU Delft), located in the Netherlands, has observed a large increase in student numbers over the past few years [1, 2, 3]. Our Computer Science Bachelor's programme went from ≈350 students in 2017–2018 to ≈900 students in 2018–2019. In practice, this means that lecturers need to adjust not only their teaching methods, but also their assessment methods, while keeping their budgets at somewhat reasonable levels.

In the case of *Software Testing and Quality (CSE1110)*, a software testing course that runs in the 4th quarter of the 1st year of the Computer Science Bachelor's programme, students often worked on three graded lab assignments (which together accounted for 20% of their final grade) and a final exam (the remaining 80%). The students' submissions for the three lab assignments used to be graded by our team of teaching assistants (TAs) throughout the quarter.

Given the large growth of the student population, grading all the submissions became impractical. In a simple calculation, even with students working in teams of two, the number of submissions to be graded in the course could go up to 450×3 = 1,350 submissions. Given that we have empirically observed that TAs take one hour per submission on average, this would lead to a total of 1,350 hours of grading. Assuming we pay an average of 25 euros per TA hour, the total cost of grading would exceed 30,000 euros.

From the educational point of view, we did not want to remove the lab work (as we believe it is valuable for better learning) nor did we want to stop grading it (as we believe most students would not work on an ungraded assignment). As a way to keep the lab work but reduce its grading costs, we opted for experimenting with self and peer grading. In other words, students grade themselves and their peers.

Research has shown various benefits of self and peer assessment and grading. As examples, Adams and King [4] argue that self-assessment is a valuable teaching and learning aid; Walser [5] shows that self-assessment provides students with the opportunity to reflect on the course and their performance, and to help them monitor their own progress. El-Koumy [6] even argues that self-assessment is a fundamental aspect in constructivist education.

However, self and peer assessment do not come without challenges. Boud and Falchikov's literature review [7], containing research papers from the 1930s up to the 1980s, shows that students' self-assessments tend to be higher than the teachers' assessments. The same phenomenon is also observed in more recent studies, e.g., [8, 9, 10]. Falchikov and Boud [11] also observed that grades tend to be more similar to the real ones in more advanced courses (and less in introductory courses, like in our case). According to Liu and Carless [12], the reliability of the assessment (i.e., how much one can trust it) is indeed an important reason why teachers do not make use of self and peer assessment techniques more often.

We made use of self and peer grading in the 2018–2019 edition of our software testing course. Students worked in teams of two, and self and peer graded three assignments.[1] We ended up with 906 self and peer graded submissions, which we compared to 248 submissions that were graded by our TAs. In this paper, we report on the differences we observed among self, peer, and TA grading.

[1] Given that students work in teams of two in our course, whenever we use the term *self grading*, we refer to the team as a single entity, grading their own work. The same applies to the term *peer grading*, which refers to one team grading the submission of another team of two students.

We summarise our findings as follows: (i) self grades tend to be 8–10% higher than peer grades on average; however, around 25% of the teams give themselves a self grade lower than their peers; a perfect match between self and peer grades rarely happens. (ii) peer grades seem to be a good approximator of TA grades; in cases where self and peer grades diverge significantly, the TA grade appears to lie in between. (iii) the gender and the nationality of the student do not seem to affect self and peer grading.

II. Research Methodology

The goal of this paper is to **understand how comparable peer, self, and TA grades are** and whether **self and peer grading are affected by socio-demographic characteristics of the students**. To that aim, we propose the following research questions:

RQ$_1$. How do peer and self grades compare to each other?
RQ$_2$. How do TA grades compare to self and peer grades?
RQ$_3$. Do socio-demographic characteristics influence how students self or peer grade?

In Figure 1, we summarise the design of our study. To answer these RQs, we collected data as follows: 1) teams submitted their assignments at specified deadlines, 2) we sent teams the solutions of the exercises, 3) teams graded their own work, following our rubrics, 4) teams graded the work of an anonymous team, following the same rubrics, 5) TAs graded teams where self and peer grades were divergent. Teams performed this procedure three times (once per assignment).

In the following sections, we describe how teams and assignments work (Section II-A), the process that teams followed to self and peer grade (Section II-B), the process that TAs followed to grade a subset of the submissions (Section II-C) and, finally, the data collection and analysis methods we use to answer the research questions (Section II-D).

A. Teams and Assignments

Due to the large number of students participating in our course, we grouped them in teams of two members each. Students were free to choose with whom to form a team. Our rules stated that partners are equally responsible for the assignments. In Table I, we show the number of teams we study. As expected, teams composed of two Dutch students represented the vast majority. We also observe an imbalance between the number of male-only teams and those that include a female student.

Throughout our software testing course, teams worked on JPacman [13], an educational implementation of the PacMan game in Java that contains several opportunities for different testing techniques to be applied. All assignments included theoretical and reflexive questions about software testing (which teams answered in a written report), as well as technical questions (which teams answered by submitting source code). The lab work was worth 20% of the teams' final grades (the remaining 80% comes from a theoretical exam; teams were required to obtain at least 57.5% of the points of the lab work to pass the course).

TABLE I: The teams that participated in our study (N=332). All teams are studied in RQ$_1$ and RQ$_2$. The 48 non-identified teams are related to students we did not have access to their personal information, and thus, they are not studied in RQ$_3$.

	# of teams	**% of teams**
Overall participants (RQ$_1$, RQ$_2$)		
Total number of teams	332	100.0%
Teams per Nationality (RQ$_3$)		
Two Dutch students	116	34.9%
Two EU students	61	18.3%
One Dutch and one EU student	46	13.8%
One EU and one non-EU student	23	6.9%
Two non-EU students	20	6.0%
One Dutch and one non-EU student	18	5.4%
(Non-identified)	48	14.4%
Teams per Gender (RQ$_3$)		
Two male students	219	65.9%
A male and a female student	41	12.3%
Two female students	24	7.2%
(Non-identified)	48	14.4%

The lab work was composed of one warm-up assignment (which we also use as a way for teams to get used to the self and peer grading procedure) and three graded assignments:

- **Assignment 0 (warm-up, ungraded):** Clone the project from GitHub, configure the project in your IDE, write your first JUnit test, run coverage analysis.
- **Assignment 1:** Write a smoke test, functional black-box testing, boundary tests, reflect on test understandability and best practices. Assignment 1 was composed of 15 exercises, and rubrics contained a total of 99 points.
- **Assignment 2:** White-box testing, mock objects, calculate code coverage and apply structural testing, use decision tables for complex scenarios, reflect on how to reduce test complexity and how to avoid flaky tests. Assignment 2 was composed of 18 exercises, and rubrics contained a total of 92 points.
- **Assignment 3:** Apply state-based testing, test reusability, refactor and reflect on test smells. Assignment 3 was composed of 23 exercises, and rubrics contained a total of 62 points.

The reader can find details of our lab assignments and rubrics in the online appendix [14]. We also refer the reader to another publication of ours that describes the educational methodology we follow in the course [13]. In a nutshell, our course is composed of 14–16 lectures, divided over 8–9 weeks, that mix theory (i.e., the different testing techniques, how they work, scientific evidence) and practice (i.e., where students see snippets of code and how to test it using state-of-the-art testing tools). In addition, students have two lab sessions a week, where they solve the lab assignments we explain above. The lectures cover the same topics of which the knowledge is required in the lab work.

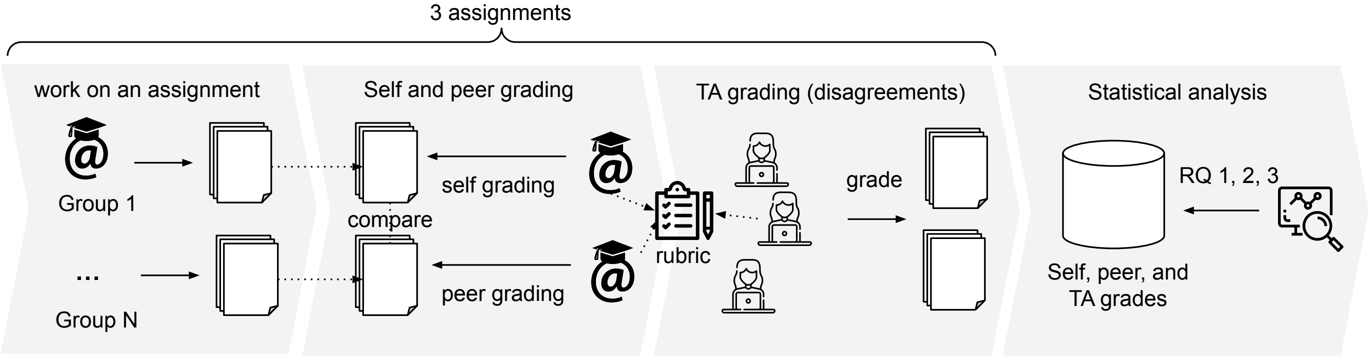


Fig. 1: The procedure we follow in this study.

In this particular edition of the course, we gave teams eight working days to submit assignment 0, eleven working days for assignment 1, twelve working days for assignment 2, and twelve working days for assignment 3. An assignment started as soon as the previous one had been submitted.

B. The self and peer grade procedure

Right after the due date of each of their assignments, we provided the teams with the solutions and extensive explanations of all the exercises (available in our online appendix [14]). We then required teams to grade themselves and to grade an anonymous team. We also instructed teams that both students should work in pairs during the self and peer grade activities.

The teams made use of a set of rubrics that we devised in advance. These rubrics are sets of questions with closed answers to each of the questions of the assignment. For the questions that required source code, our rubric stated what one should expect from it. For example, for a given question about state-based testing, our solution contained a list of six tests that one should derive for that state machine. Our rubrics for the question was: *[0] No implemented test cases, [1] At least 3 tests implemented correctly, [2] At least 5 tests implemented correctly, [3] 6 or more tests implemented correctly.* In case of open or reflexive questions, our rubrics contained a list of points we would expect students to discuss about that question. For example, in a question about why a test suite can become slower over time and what testers can do about it, our rubrics contained *[0] invalid reason or no explanation, [2] one example (e.g., file or database access) with no generalisation of the problem, [4] a generalisation of the problem + a solution for the problem.*

We used our own in-house solution for the review process. Each team had its own login credentials for the tool. For each submission, the peer review tool showed two tasks. The first task was about grading their own work, using the rubric that we previously discussed. Rubrics were always listed from the smallest possible grade (often a zero, meaning that the team did not submit that exercise) to the highest possible grade (meaning that the team's submission was perfect). The numbers between brackets indicated the number of points that the answer gives. We show the number of points per assignment in Table II. Teams were instructed to compare the solutions we provided with their provided answer. We also offered an open text field at the end of the review, where teams could explain some of their choices.

As soon as the team was done with the self review, the system then randomly assigned a group to be graded by that team, using the same rubrics. Teams did not know which group they were grading (we explicitly asked teams not to include any information that could identify themselves in the reports and source code). The tool then enabled the team to download the report and source code submitted by their peers. We gave teams five working days to deliver both peer and self grades for the warm-up assignment 0, six working days for assignment 1, seven working days for assignment 2, and four working days for assignment 3. The varying number of days between the different assignments happened due to the overall schedule of the course. We made sure that the teams had at least one monitored lab session before the deadline, and we avoided exam days and holidays.

Note that we chose this order (first the self grade and then the peer grade) on purpose. Our rationale was that, with self grading coming first, teams would be able to learn from their mistakes first, and then be better prepared to grade others. We also note that, once submitted, teams could not update their self or peer grades.

As explained above, self and peer grading were compulsory activities of our course. We warned students that not delivering peer reviews could imply failing the course. In the end, only a few self and peer reviews were missing and no students/teams were punished. We therefore discarded these data points when we compared the differences between self and peer grades. Also, the warm-up in assignment 0 was used as a way to educate students and make them familiar with self and peer grading; therefore, we only analyse the data from assignments 1–3.

C. The TA grading procedure

For each assignment (1–3), as soon as teams delivered their self and peer grades, our team of TAs graded a subset of the assignment submissions. While we acknowledge that grading

TABLE II: The number of assignments that were submitted by teams, graded by TAs, and reviewed by TAs in the three different strata. Confidence interval (CI) at a 95% confidence level.

Assignment	**Number of exercises**	**Total number of points**	**# of team submissions**	**# of submissions graded by self, peers, and TAs**	**CI**	**# of teams (stratum 1)**	**# of teams (stratum 2)**	**# of teams (stratum 3)**
Assignment 1	15	99	329	62 (18.84%)	11.22	26 (7.90%)	7 (2.12%)	29 (8.81%)
Assignment 2	18	92	324	81 (25.00%)	9.44	33 (10.18%)	5 (1.54%)	43 (13.27%)
Assignment 3	23	62	317	105 (33.12%)	7.83	28 (8.83%)	50 (15.77%)	27 (8.51%)
SUM	56	253	970	248	5.37	87 (8.96%)	62 (6.39%)	99 (10.20%)

them all would give us more data to analyse, we had a team of 17 TAs, each available for 4 to 6 hours a week. Given that we had a total of 332 teams participating in the lab work, and that we had 3 different graded assignments (so a total of approximately 332×3 = 996 submissions to grade), and that our experience shows that each submission takes 1 hour to grade, we did not have enough budget to grade them all.

We performed *stratified sampling*. In the following, we describe the three different *strata*, which were derived based on our experience as lecturers of this course:

1) **High difference between self and peer grade**. We selected all groups where the difference between the self grade and the peer grade was higher than 25% of the maximum number of points for the assignment.
2) **Both self and peer grades are high**. We selected all groups where both the peer and the self grades had a grade higher than 90% of the maximum number of points for the assignment.
3) **Random teams**. For each assignment, we randomly selected a number of groups that did not fit in any of these categories. More specifically, teams in this category had a peer and self grade that had a difference of at most 25% of the points, and grades were not higher than 90%.

The 25% and 90% thresholds as well as the number of random groups per assignment were chosen arbitrarily, after some experimentation. Our goal was to limit the number of submissions to be graded, while making sure that the ones we selected were the most relevant ones. As discussed before, grading all submissions (and thereby acquiring more data points) would require a substantially larger TA budget. Another reason to consider is that we planned to use these grades as part of the final grade of our course, so therefore TAs also had to grade the teams that did not deliver either the peer or the self grade (so for them there was more work involved than only the submissions in the three strata). We discuss the impact of this decision and how future work should tackle it later in Section VI.

We note that all our TAs are former students of this course, have worked on the same assignments by themselves in their years, and passed the course with high grades. In other words, our TAs are quite familiar with the course material and assignments.

In Table II we show the number of submissions our TAs graded per assignment, as well as the confidence interval (at a confidence level of 95%) such a sampling strategy brings to this study.

D. Data collection and analysis

To answer RQ_1, we explored the differences between self and peer grades, using violin plots and statistical tests. The data points (i.e., grades) are grouped by teams and assignments. More formally, given an assignment n, and a team m, $S_{n,m}$ represents the self grade that team m assigned to itself in assignment n, and $P_{n,m}$ represents the peer grade that an anonymous team assigned to team m in assignment n.

We performed an individual pairwise comparison between the grades. More formally, for the two grading techniques (self grading, peer grading), and an assignment n, we compared the difference of the paired data points $self - peer$. A positive number indicates that the self grade was higher than the peer grade. Similarly, a negative number indicates that the peer grade was higher then the self grade. A difference of zero means that both strategies ended up with the same grade.

We answer RQ_2 in a similar fashion, with the addition of the TA grades that exist for teams that were part of the three strata. We made use of violin plots and statistical tests to measure the differences between peer, self, and TA grades. We grouped the analysis per stratum and assignment.

To answer RQ_3, we compared teams according to their composition in terms of nationality and gender. For this, we used the official information that was provided by students when they enrolled at the university. Some students (around 14% of the teams) however, belonged to a different cohort, and we did not have access to these data for them. These teams were excluded from the analysis of this RQ.[2]

Given that teams always consisted of two members, we grouped teams into the following groups:

1) **Gender:**[3] [male, male], [female, female], [male, female].
2) **Nationality:** [dutch,dutch], [european, european], [non-european, non-european], [dutch, european], [dutch, non-european], [european, non-european].

More formally, given the set of gender-specific teams (i.e., teams composed of male-male, female-female, or male-female

[2] The statistical analysis was conducted after the course was finished. Once we realised the university did not have such information, we also had no means to contact the students again.

[3] We acknowledge that the separation of gender in between male and female only is not inclusive. This is, however, how our university currently stores this data. For a more inclusive society, we hope to change it in the future.

students) or nationality-specific teams (e.g., teams composed of Dutch-Dutch or European-European students), and the two grading techniques (self and self-to-peer grading), we perform (non-paired) comparisons between two sets of teams, in terms of their self and self-to-peer grades, for all the assignments. Note that, in RQ$_3$, we intend to measure whether personal characteristics affect how teams grade *themselves* and *their peers*. Therefore, we define a *self-to-peer* grade ($SP_{n,m}$) representing the grade that team m *gave* to their anonymous peer (and not the grade they received from their peers) in assignment n.

In all the RQs, we applied statistical tests and effect sizes to understand whether the differences are statistically significant. And while we tested the normality of all the distributions we compared using D'Agostino's [15] and Pearson's [16] tests, we decided to always use non-parametric tests. The reason was that, although some distributions indeed presented normal characteristics, they all have a significant number of outliers, which are known to affect the power of any parametric test. We therefore use the Wilcoxon signed-rank test [17] to compare paired samples, the Mann-Whitney rank test [18] for unpaired samples, and Cliff's Delta to measure the effect size of the differences. We used *scipy*'s implementations of all the tests, except for Cliff's Delta where we used an open-source implementation[4] because *scipy* does not offer one natively.

We set our alpha-level to 0.05. In addition, although we performed several tests in a row, we opt to not apply any type of correction in the alpha-level (e.g., Bonferroni correction). Besides its contradictory use [19, 20], we are interested in each individual comparison, and not in inferring the overall behaviour of all the compared data sets. Nevertheless, we provide all p-values in our online appendix [14] for readers who prefer to interpret the results with a stricter alpha-level.

III. Results

In the following sections, we present and discuss the results of the three research questions we posed. Additional graphs can be found in our online appendix [14].

A. *RQ$_1$: How do peer and self grades compare to each other?*

In Figure 2, we show the pairwise differences between self and peer grades. A positive number in the violin plot indicates that the self grade is higher than the peer grade. Similarly, a negative number indicates that the self grade is smaller than the peer grade. A difference of zero means that self and peer assessments resulted in the same grade.

Observation 1: Self grades are, on average, 8–10% higher than peer grades. In the three assignments, teams often attributed higher grades to themselves than the ones given by their peers. We observe a mean difference of 8.34±12 points higher in the first assignment (around 8% of the total points of the assignment), 8.8±12 points higher in the second assignment (around 10%), and 5±8.5 points higher in the third assignment (around 8%). This is also confirmed by the Wilcoxon and Cliff's Delta tests. We observe statistically significant differences in the medians of all assignments, with a Cliff's Delta of 0.41 (1st assignment), 0.38 (2nd assignment), and 0.35 (3rd assignment), under an alpha-level of 0.05.

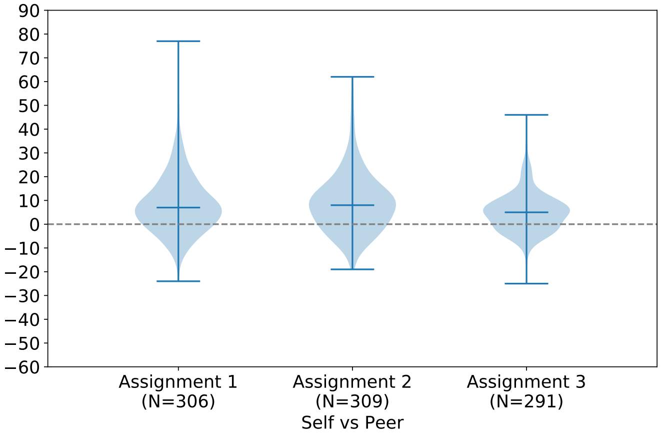


Fig. 2: Violin plots representing the difference, in points, between self and peer grades (*self* − *peer*), per assignment. Horizontal lines represent the maximum, the median, and the minimum values, respectively. The dashed line represents the point where self and peer grades are exactly the same.

Observation 2: Around a quarter of the self grades are smaller than their peer grades. 24% of the teams (or more precisely, 22.2% of the teams in assignment 1, 22.3% in assignment 2, and 26.1% in assignment 3) gave themselves a grade lower than what their peers gave them.

Observation 3: A negligible number of self grades match the peer grades exactly. Only 11 teams in assignment 1 (3.5%), 8 teams in assignment 2 (2.5%), and 16 teams in assignment 3 (5.4%) had the exact same self and peer grade.

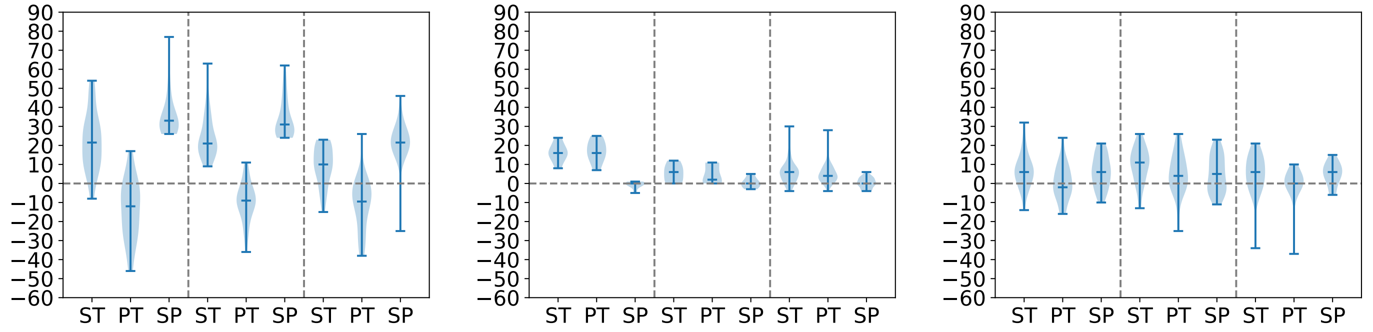
RQ$_1$ findings: Self grades tend to be 8–10% higher than peer grades. Around 25% of the teams give themselves a self grade lower than their peers. Precise matches between the self and peer grade rarely happen.

B. *RQ$_2$: How do TA grades compare to self and peer grades?*

In Figure 3, we show the differences between the self, peer, and TA grades, per assignment, per stratum.

Observation 4: In cases where self and peer grades differ significantly, TA grades seem to be in between the self and peer grades. We observe in Figure 3a, stratum 1 (25% difference between self and peer grades), that grades given by the TAs are somewhat in between the peer and self grades. We see that the median difference between self and TA grades are 22±15.4 (note the positive number), while the difference between peer and TA grades are -13±15.8 (note the negative number). The difference in all the means are statistically significant at an alpha level of 0.05. In other words,

[4] https://github.com/neilernst/cliffsDelta.



(a) Stratum 1: 25% difference between self and peer grades. 1st assignment N=26, 2nd assignment N=33, 3rd assignment N=28.

(b) Stratum 2: Self and peer grades higher than 90%. 1st assignment N=7, 2nd assignment N=5, 3rd assignment N=50.

(c) Stratum 3: Random teams. 1st assignment N=29, 2nd assignment N=43, 3rd assignment N=27.

Fig. 3: The differences, in points (Y axis), between self vs TA (ST), peer vs TA (PT), and self vs peer (SP) grades, for assignments 1, 2, and 3, in the three different strata. Assignments are separated by dashed vertical lines.

while teams give themselves a much higher grade, peers also tend to be much stricter, and give their peers lower grades when compared to TA grades.

Observation 5: The differences between TA grades and peer grades do not seem significantly high in assignments that received high self and peer grades. We observe in Figure 3b, stratum 2 (self and peer grades higher than 90%), that, apart from assignment 1, TAs seem to agree with the high grades given by the team (self) and by the peer. In both assignments, TAs gave around 5±5 less points than peers and the team; we also do observe a statistically significant difference in the medians at an alpha level of 0.05, which should be interpreted with a grain of salt, given that the number of data points in this stratum is low. In assignment 1, we also observe a more considerable difference between TA grades and other grades. We have no clear explanation for why that happens. We conjecture this might be due to the teams still getting used to the project itself and the self-grading process.

Observation 6: Peer and TA grades were closer in the random group. We observe in Figure 3c, stratum 3 (randomly selected teams), that the average difference between peer and TA grades is 0.5±9.4 points in assignment 1, 4.9±10.9 points in assignment 2, and 0±8.6 points in assignment 3. In contrast, the difference between self and TA grades are 6.6±9.2 points in assignment 1, 9.7±9.2 points in assignment 2, and 6±10.9 points in assignment 3.

RQ$_2$ findings: Peer grades seem to be a good approximator of TA grades. In cases where self and peer grade diverge significantly, the TA grade appears to lie in between.

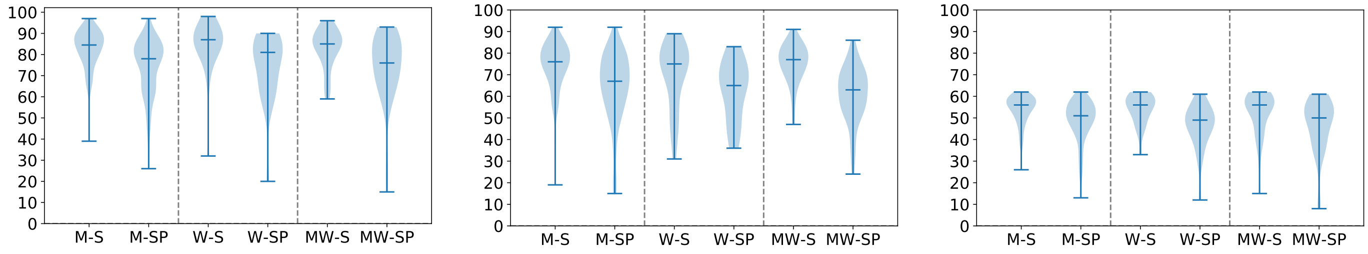
C. RQ$_3$: Do socio-demographic characteristics influence how students self or peer grade?

In Figure 4, we show the self- and self-to-peer grades from groups composed of only men, only women, and men and women. In Figure 5, we show the self- and peer-to-self grades given by the groups, according to their nationalities, in assignment 1. The plots for the other assignments are available in our online appendix [14].

Observation 7: Gender seems not to affect the self and self-to-peer grade distributions. We did not find any statistically significant differences when comparing the (non-paired) distribution of the self grades between male only, female only, and male and female teams in the three assignments (i.e., Mann-Whitney's $p > 0.05$ in all the pairwise comparisons). Effect sizes ranged from negligible to small. We observed the same results when comparing the distribution of self-to-peer grades; again, no statistically significant differences, and small to negligible effect sizes.

Observation 8: Teams gave themselves a higher self grade, regardless of the gender. Similar to the previous results, we observe teams giving themselves a grade that is significantly higher than the peer grades in the three assignments, also when broken down by gender. More formally, we obtained Mann-Whitney's $p < 0.05$ and medium to large Cliff's Delta values, when comparing male only, female only, male and female teams to the overall distribution of TA grades, for all the three assignments. By visually inspecting the plots, we do not see any group that would give themselves a much higher grade than other groups.

Observation 9: Nationality seems not to affect self and self-to-peer grade distributions. We only observed negligible differences between the teams (i.e., Mann Whitney's $p > 0.05$ and Cliff's Delta negligible or small), regardless of their composition. The largest difference we observe happens in Assignment 2, where Dutch-only teams have a self grade median of 73 points ($[Q25 = 65, Q75 = 79]$), whereas European-only teams have a median of 79 points ($[Q25 = 74, Q75 = 83]$), a six points difference. We also do not observe any particular pattern when visually inspecting the plots (see appendix [14]).



(a) Assignment 1

(b) Assignment 2

(c) Assignment 3

Fig. 4: The self-grades (S) and self-to-peer (SP) grades between groups composed of only men (M), only women (W), and men and women (MW). Assignments are separated by dashed vertical lines. Y axis represents the number of points.

RQ_3 findings: Gender and nationality do not seem to affect the way teams perform self and peer grading.

IV. Lessons Learned

Our study brought us interesting insights on the usefulness of self and peer grading, as well as the lessons learned on how to make use of it in practice.

Self and peer grading seem to be a good alternative for grading lab assignments. Our experience shows that using self and peer grading *can serve* as a suitable replacement for (paid) TAs.

We observe in RQ_2 that, for the randomly selected teams we analysed, peer grades were not that different from the TA grades (a 5±5 points difference).

It is indeed true that the number of data points in the random stratum (stratum 3), when compared to the overall population of submissions, brings a confidence interval of ≈9 under a confidence level of 95%. In other words, the difference can be up to 9% larger or smaller than what we present above. Nevertheless, we argue that, for lab assignments that represent a smaller fraction of the overall final grade, such as the one studied in this paper, a 15% difference in the final grade might not be that important (e.g., for an assignment that is worth 2.0 points in the final grade, 15% means only 0.3 points).

The reductions in TA costs might be worth the small difference in the grade. As a simple calculation, we estimate TAs to take 1 hour to grade a submission, at an average hourly cost of 25 euros. The (329+324+317 =) 970 submissions would cost a total of 24,250 euros in TA hours. More conservative teachers might consider paying TAs to grade submissons where the self and peer grade diverge. From Table II, we observe that the number of submissions where self and peer grade diverge for more than 25% (stratum 1) ranges from 8 to 10%. That would mean a 90% reduction in costs when compared to grading all submissons. A more conservative threshold, such as 15, would require TAs to grade 25% of all the submissions.

Self grading as a way to explore the rubric. In this course edition, we asked students to perform self and peer grading, as we were interested in understanding their differences. While our results show that peer grading seems to be more effective than self grading, we conjecture that the process of self grading first, and then peer grading later, caused students to better understand the exercise, the rubric and the expected answer, and therefore should be done as a way to increase the quality of the peer grading. Interesting future work would be to investigate whether peer grades are more precise when students also perform self grading.

Students' perceptions. We did not conduct any survey at the end of our course to collect the students' perceptions on the self and peer grading (which we suggest as a great addition for researchers that intend to replicate this work). The informal perception we collected by interacting with the students throughout the course was positive. Our overall impression was that students understood the seriousness of the task (at the beginning of the course, we reminded them of the academic honour agreement they have with the university).

We noticed a significant increase in their workload. Teams took an average of 4 hours for self- and peer-grading. The four assignments (warm-up + 3 graded assignments) implied an increase of 16 hours in their (already high) workload. We recommend lecturers to decide whether this is feasible within their courses.

Some teams had questions about our rubrics and how to grade some exercises. We expected such questions given that this was the first time we opened our rubrics for the students. We offered teams the support of TAs (i.e., students could ask questions about the grading). We also offered the advice that, whenever in doubt between two points in the rubric, go for the one with the highest grade (i.e., do not punish yourself or your colleague in case of doubt).

Again, all these insights are based on our perceptions as teachers, and lack scientific validity. Future work should invest in collection instruments to measure the perception of students at each step of the process.

Handling disagreements. Before starting the course, we conjectured that we would need to handle disagreements between groups that give themselves a high self-grade while their peers give them a lower grade. We allowed students to show their disagreements by contacting the lecturers and the TAs. Unexpectedly, the number of complaints we received was negligible. We conjecture this was due to the clear

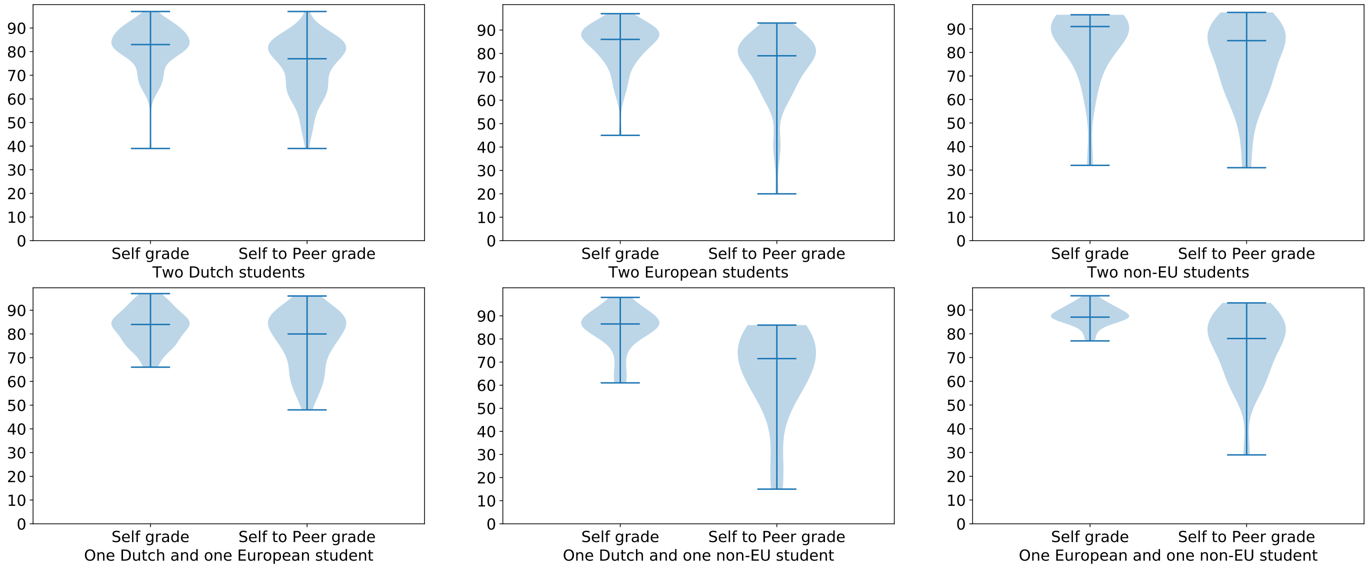


Fig. 5: The self- and peer-to-self grades given by groups, according to their nationalities, in assignment 1. The Y axis represents the number of points. The plots for the other assignments are available in our online appendix [14].

rules we stated at the beginning. We explained to them that disagreements would be solved by a teaching assistant (i.e., the TA would grade the submission). In practice, our solution was: whenever a TA graded a submission, we used the TA grade as the final one; otherwise, we picked the highest grade between the peer and the group. Given that the largest differences were reviewed by the TAs, we do not think that picking the highest grade inflated the overall grades significantly. Nevertheless, choosing the threshold that determines whether a submission gets reviewed by a TA or not, is an important design decision, and has a financial impact.

Are these results valid for assignments that represent a major part of the final grade? Students knew in advance that the lab work (for which self and peer grading was done) would be worth 20% of the final grade. It is hard to conjecture whether their behaviour would have been the same if they were asked to self and peer grade something like a final exam, where the responsibility (and the impact of getting a higher or lower grade) is much higher. We argue that this is an important thing to research in the future, as we often spend many TA hours on grading exams.

Are these results valid for other computer science disciplines rather than software testing? This entire observational study was conducted within a software testing course. We do not particularly see any specific characteristic of the contents of our course and/or the way our assignments are designed that would make our results not generalisable to other computer science courses. However, as we later discuss in the Threats to Validity section, the development of good and clear rubrics that support the students in the self and peer grading task is of utmost importance, and our results may be highly influenced by them. Future work should investigate best ways of devising rubrics for self and peer grading. We nevertheless invite lecturers from other courses to experiment with self and peer grading, and we also invite computer science researchers to replicate our study in different courses.

V. Related Work

In this section, we discuss some of our decisions when performing the case study and how it compares to related work that took a different decision.

Regarding encouragement, our students knew they would lose points in case they did not perform their assessment well. Bloxham and West [21] took an opposite approach: authors awarded students with an extra 25% on their assignment marks for the quality of the peer review. We only measured the quality of the reviews that were delivered by the students in a non-systematic way (e.g., random checks throughout the course), but we did not observe anything significant. Nevertheless, we believe that rewarding students positively instead of punishing them is a good idea, and will consider this in a future edition of the course.

Regarding rewards for good assessments, we asked students to first perform their self assessment and later perform the peer review. We conjectured that students should first get familiar with the rubrics using their own solutions, so that they can better perform peer review. Boud [22], on the other hand, suggests the other way around: students first do peer review, and later assess themselves. The author conjectures that seeing a different answer first might enable the student to better assess him/herself. This is indeed an interesting thought that we should evaluate in future editions of the course.

Regarding training, King [4] argues that students first need to develop skills in self-assessment for it to become useful. We did not give our students any formal training on self and peer assessment, besides discussing some best practices during the initial lecture. However, we provided students with a highly detailed rubric that they could follow, in an similar approach to Sadler and Good [10], where authors actually observed a high correlation between (middle school) students' grades and the teacher's grade. We nevertheless believe that, once students are more used to self assessment and better trained for it, results will only get better.

In our research, we also explored whether socio-demographic factors (i.e., country of origin and gender) affect the self- or peer-grading process. Our motivation comes from recent research showing the significant gender gap that exists in higher education. For instance, male and female students may have different success paths [23], female students who want to choose CS later in life have a higher self-efficacy [24], and only excellent female students choose computer science [25]. In addition, international students also face their own challenges [26, 27]. Our results do not show any differences in the grades, regardless of demographics and gender. We argue this increases the reliability of such grading methods even more. Nevertheless, we argue that many more replications are needed before these findings can be generalisable.

Finally, we conclude that self and peer grading seems to be a feasible alternative to grading by lecturers and TAs. Other lecturers came to similar conclusions. Freeman and Parks [9] also concluded that peer grading may be accurate enough for low-risk assessments (in their case, in introductory biology courses). According to the authors, peer grading can help relieve the burden on teaching staff. Sadler and Good [10] offer a similar conclusion to ours: self grading and peer grading appear to be reasonable aids to saving the instructor time. And while Liu and Carless [12] state that reliability is one of the main reasons for resistance, the same authors also believe that it is already well-recognised that students are reasonably reliable assessors.

VI. Threats to Validity

In this section, we discuss the threats to the validity of our study and how we mitigate them.

A. Internal and construct validity

To answer RQ$_2$, we make use of the subset of deliverables that TAs graded. The random group (stratum 3) contained 8–11% of the assignments in that group. This brings us a confidence interval of around 9, under a confidence level of 95%. While a higher number of data points would give us more insights, it was simply not feasible, cost-wise. We nevertheless present all the obtained statistical results as well as violin plots for visual inspection in our paper and our online appendix [14].

The rubrics themselves and how they were presented to the students/teams may have affected the way teams performed the self and peer grading. The rubrics were devised by the lecturers and reviewed by our team of teaching assistants. While these rubrics have been used by our teaching assistants in previous years, this was the first time students used them. As we discuss in our lessons learned (Section IV), a few points of the rubrics were found to be confusing by the teams. As a way to mitigate the impact of rubrics in the grades, we discarded the grades for two exercises that the majority of teams found confusing. We also told teams to always go for the highest grade whenever in doubt between two items of the rubric (both during the self and the peer grading). While we have no reason to believe that the rubrics we used could have drastically affected our results, we argue that the design of the rubrics plays an important role in the reliability of the grades, and we suggest lecturers to design them carefully.

Another point to consider is that the teams performed the self grading before the peer grading. As we discussed in the Related Work section, it is common to see teachers asking their students to first perform peer grading and then do the self grading. Future work should investigate whether the order in which we give the task to students affects the final outcome.

Finally, while we asked teams to avoid adding their names in the assignments (as to make peer grading fully anonymous), we did not fully ensure this was the case. Our TAs did not identify any non-anonymous submissions among the ones they graded themselves. Therefore, we argue that our results would not change drastically in case a few submissions were not properly anonymised.

B. External validity

In this paper, we studied a cohort of more than 600 students, divided into 332 teams of two, at Delft University of Technology, and used data from three software testing lab assignments. As we discussed before, while we conjecture that these results could apply to other courses and institutions with characteristics similar to ours (i.e., large classes, diverse students, lab work that represents a fraction of the final grade), future work should replicate these results.

We also note that the self and peer grades were conducted by the students in teams of two. Our results may not generalise to students working alone. We leave the exploration of whether students would have the same behaviour if working alone for future work.

This study was conducted in the context of a lab assignment that was worth "only" 20% of the final grade. More work is required before we can generalise our findings to more important assignments, e.g., self and peer grading of midterm and final exams.

VII. Conclusion

The steady increase of student numbers in many computer science programs requires lecturers to propose innovative and cost-effective grading solutions. In particular, courses that incorporate extensive lab assignments, like ours, now reach the limits of what is possible with a solution where lecturers and TAs do all the grading.

As one of the possible solutions to this problem, we have experimented with self and peer grading in the 2018–2019 edition of the software testing course we deliver in the 1st year of the Computer Science Bachelor's programme at the Delft University of Technology. After finishing an assignment, teams would grade themselves, and grade a random team, using a closed-ended rubric. Large differences between self and peer grades were solved by the teaching assistants.

Our results show that:

1) Self grades tend to be 8–10% higher than peer grades. Around 25% of the teams give themselves a self grade lower than their peers. Precise matches between the self and peer grade rarely happen.
2) Peer grades seem to be a good approximator of TA grades. In cases where self and peer grade diverge significantly, the TA grade appears to lie in between.
3) Gender and nationality do not seem to affect the way teams perform self and peer grading.

We argue that, while our results show that teams tend to inflate their grades by $\approx 10\%$, at least when compared to their peers and TAs, for assignments that are worth a small part of the final grade (e.g., lab work), self and peer grading seems to be a good way to reduce the effort and costs of grading.

We therefore suggest lecturers to explore the possibility of self and peer grading for parts of their courses. We hope that this will not only reduce teaching costs, but also free up some of the lecturers' time, which can then be spent on course improvements.

ACKNOWLEDGEMENT

We thank the hard effort of all our teaching assistants. This course (and this paper) would not be possible without them.

The icons we use in our methodology figure are freely provided by Freepik (https://www.flaticon.com/authors/freepik) and Smartline (https://www.flaticon.com/authors/smartline).

REFERENCES

[1] B. Belleman, "Number of university students exceeds record-breaking 300,000," https://www.delta.tudelft.nl/article/number-university-students-exceeds-record-breaking-300000, 2019, last accessed in July, 2020.

[2] S. Bonger, "Computer sciences remains popular," https://www.delta.tudelft.nl/article/computer-sciences-remains-popular, 2019, last accessed in July, 2020.

[3] C. van Uffelen, "Numerus fixus for computer science engineering," https://www.delta.tudelft.nl/article/numerus-fixus-computer-science-engineering, 2018, last accessed in July, 2020.

[4] C. Adams and K. King, "Towards a framework for student self-assessment," *Innovations in Education and Training International*, vol. 32, no. 4, pp. 336–343, 1995.

[5] T. M. Walser, "An action research study of student self-assessment in higher education," *Innovative Higher Education*, vol. 34, no. 5, p. 299, 2009.

[6] A. S. El-Koumy, "Student self-assessment in higher education: Alone or plus?" *Education Resources Information Center (ERIC), USA*, 2010.

[10] P. M. Sadler and E. Good, "The impact of self-and peer-grading on student learning," *Educational assessment*, vol. 11, no. 1, pp. 1–31, 2006.

[7] D. Boud and N. Falchikov, "Quantitative studies of student self-assessment in higher education: A critical analysis of findings," *Higher education*, vol. 18, no. 5, pp. 529–549, 1989.

[8] T. Vogelsang and L. Ruppertz, "On the validity of peer grading and a cloud teaching assistant system," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 2015, pp. 41–50.

[9] S. Freeman and J. W. Parks, "How accurate is peer grading?" *CBE—Life Sciences Education*, vol. 9, no. 4, pp. 482–488, 2010.

[11] N. Falchikov and D. Boud, "Student self-assessment in higher education: A meta-analysis," *Review of educational research*, vol. 59, no. 4, pp. 395–430, 1989.

[12] N.-F. Liu and D. Carless, "Peer feedback: the learning element of peer assessment," *Teaching in Higher education*, vol. 11, no. 3, pp. 279–290, 2006.

[13] M. Aniche, F. Hermans, and A. van Deursen, "Pragmatic software testing education," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 414–420.

[14] M. Aniche, F. Mulder, and F. Hermans, "Grading 600+ students: A case study on peer and self grading (appendix)," http://doi.org/10.5281/zenodo.4475243.

[15] R. B. d'Agostino, "An omnibus test of normality for moderate and large size samples," *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971.

[16] E. S. Pearson, R. B. D'AGOSTINO, and K. O. Bowman, "Tests for departure from normality: Comparison of powers," *Biometrika*, vol. 64, no. 2, pp. 231–246, 1977.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[18] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[19] T. V. Perneger, "What's wrong with bonferroni adjustments," *BMJ: British Medical Journal*, vol. 316, no. 7139, 1998.

[20] S. Nakagawa, "A farewell to bonferroni: the problems of low statistical power and publication bias," *Behavioral Ecology*, vol. 15, no. 6, 2004.

[21] S. Bloxham and A. West, "Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment," *Assessment & Evaluation in Higher Education*, vol. 29, no. 6, pp. 721–733, 2004.

[22] D. Boud, "The role of self-assessment in student grading," *Assessment in Higher Education*, vol. 14, no. 1, pp. 20–30, 1989.

[23] R. Kolster and F. Kaiser, "Study success in higher education: male versus female students," Center for Higher Education Policy Studies (CHEPS), Netherlands, WorkingPaper 2015-07, 2015.

[24] E. Aivaloglou and F. Hermans, "Early programming education and career orientation: The effects of gender, self-efficacy, motivation and stereotypes," in *SIGCSE '19*. United States: Association for Computing Machinery (ACM), Feb. 2019, pp. 679–685, the 50th ACM Technical Symposium on Computer Science Education, SIGCSE 2019 ; Conference date: 27-02-2019 Through 02-03-2019. [Online]. Available: https://sigcse2019.sigcse.org/info/cfp.html

[25] C. Wilcox and A. Lionelle, "Quantifying the benefits of prior programming experience in an introductory computer science course," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018, pp. 80–85.

[26] C. J. Perry, "Comparing international and american students' challenges: A literature review," *Journal of International Students 2016 Vol 6 Issue 3*, vol. 6, no. 3, pp. 712–721, 2012.

[27] C. R. Glass and C. M. Westmont, "Comparative effects of belongingness on the academic success and cross-cultural interactions of domestic and international students," *International Journal of Intercultural Relations*, vol. 38, pp. 106–119, 2014.