TA Marking Parties: Worth the Price of Pizza?

Evaluating TA Confidence and Efficacy in Group vs. Individual Marking Scenarios

Brian Harrington
Department of Computer and Mathematical Sciences
University of Toronto Scarborough
brian.harrington@utsc.utoronto.ca

Marzieh Ahmadzadeh
Department of Computer and Mathematical Sciences
University of Toronto Scarborough
marzieh.ahmadzadeh@utoronto.ca

Nick Cheng
Department of Computer and Mathematical Sciences
University of Toronto Scarborough
nick@utsc.utoronto.ca

Eric Heqi Wang
Department of Computer and Mathematical Sciences
University of Toronto Scarborough
ericheqi.wang@utoronto.ca

Vladimir Efimov
Department of Computer and Mathematical Sciences
University of Toronto Scarborough
vladimir.efimov@mail.utoronto.ca

ABSTRACT

Student teaching assistants marking examinations is a reality for many undergraduate computer science courses, and with the explosion in enrolments in CS programs, and the increase in class sizes, it is becoming ever more common. Many institutions employ a "marking party" model in which instructors and TAs gather together to mark exams collectively. This system is naturally more logistically difficult than simply dividing up the papers and allowing TAs to mark in their own time. However the reasoning is that the group nature of the marking party makes it easier for markers to ask for clarification and second opinions on tricky cases, and results in better, more consistent marking.

In this work we evaluate the marking party model by performing an experiment in which TAs are randomly assigned to either a marking party or solo marking when grading the final exam of a CS1 course. However, in addition to the student papers, each TA receives "fake" papers, constructed to lead to marking errors if TAs are not attentive or do not carefully follow the assigned marking rubric. We also evaluate the time that each method of marking takes, as well as survey the TAs as to their personal opinions on the two marking methods.

Our results show that the marking party not only allows TAs to mark faster, but produces more consistent marking, with fewer errors, and better intra and inter-marker reliability. There is clear evidence that organizing marking parties is likely worth the effort (and cost of providing lunch), as the benefits to students are significant, and the overhead of the logistics may be less than that of fixing marking errors and dealing with re-mark requests. And as an added bonus, the TAs seem to enjoy it.

CCS CONCEPTS

• **Social and professional topics → Computing education**; **Computer science education**; **CS1**;

KEYWORDS

Assessment, CS1, Grading, Teaching Assistants, Evaluation

ACM Reference Format:
Brian Harrington, Marzieh Ahmadzadeh, Nick Cheng, Eric Heqi Wang, and Vladimir Efimov. 2018. TA Marking Parties: Worth the Price of Pizza?: Evaluating TA Confidence and Efficacy in Group vs. Individual Marking Scenarios. In *ICER '18: 2018 International Computing Education Research Conference, August 13–15, 2018, Espoo, Finland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3230977.3230997

Brian Harrington, Marzieh Ahmadzadeh, Nick Cheng, Eric Heqi Wang, and Vladimir Efimov

1 INTRODUCTION

With growing enrolment in computer science programs worldwide, it is becoming increasingly necessary at many institutions to rely on larger and larger numbers of teaching assistants to mark assignments and examinations, particularly in larger introductory courses. There are many advantages to having a large number of TAs on a teaching team including enhanced flexibility and coverage of possible hours of help for students. However, one major disadvantage is the increased effort required to manage logistics. In the University of Toronto Scarborough's Department of Computer and Mathematical Sciences, we regularly hold "marking parties" in which the entire teaching team for a course meet in a single room to mark exams, usually handing out stacks of papers at random, and having students mark questions in teams, moving on to the next question once they have finished marking all exams. As the number of TAs has increased, it has become increasingly difficult to find times when they are all free for a large enough timespan to mark an entire test or examination. As a department, we have attempted to persist with this model despite the logistical headache, on the basis that these marking parties resulted in a higher quality of marking with more consistency between markers and improved accountability resulting in fewer errors. However, as the number of students, and therefore TAs, has increased, many colleagues have

ICER '18, August 13–15, 2018, Espoo, Finland

ACM ISBN 978-1-4503-5628-2/18/08…$15.00
https://doi.org/10.1145/3230977.3230997

moved to a more traditional "solo" marking system, where TAs mark exams on their own time at home or in a shared workspace.

The assumption, that marking parties are better for the students than simply giving TAs access to the exams and allowing them to mark on their own time, seems intuitive. But one could argue that perhaps the marking parties are actually causing the TAs to be less focused, chatting with their friends and colleagues, and being distracted or uncomfortable in a large full room. And it may be possible that the added overhead of the marking parties is slowing down the marking process, resulting in more hours for the TAs. Furthermore, we have received occasional complaints from TAs that the system was unfair, as some questions were easier to mark than others, and no one was paying attention to how many papers each TA had marked. So some markers would work diligently, while others would chat with their friends or take more frequent breaks. It is also possible that the marking setup has no effect on the quality or efficiency of the marking, and thus instructors should feel free to choose whichever version they find most convenient.

In this paper, we detail an experiment in which we set out to discover whether there is in fact any reality to these assumptions about marking parties. Our primary research questions are:

- Do marking parties result in better inter-marker consistency than solo marking?
- Do marking parties result in better intra-marker consistency than solo marking?
- Do marking parties complete marking in more or less time than solo marking?
- Do TAs prefer marking parties over solo marking?

For the purposes of this study, we define a marking party as: A gathering of teaching assistants at a predetermined time and place to mark an examination, with the ability for the TAs to easily and continually communicate during the marking process. We define solo marking as: A system whereby TAs can mark papers at a time and in an environment of their own choosing, with access to ask for clarification only in an asynchronous manner such as e-mail (without the presumption of immediate and continuous feedback).

While marking the final examination of our CS1 course, we split our TA team into 2 groups: a "marking party" group and a "solo marking" group and randomly assigned 10 TAs to each group. Each TA was randomly assigned the same number of exam papers to mark, and given the same rubric. Unbeknownst to the TAs, they each received 2 additional exam papers in their pile which were not written by students in the course, but which were instead created by the experimenters, containing subtle errors that would be easy to overlook, or questions that could easily be given an inaccurate mark if the marking rubric was not followed carefully. After all papers were marked, the TAs were asked to self-report the total time it had taken them to mark all exams, and to fill in a short survey asking questions about their personal opinions on the marking setup.

Our results were surprisingly clear given the small sample size. Not only did the TAs prefer the marking party, and felt it was better for both them and the students, the marking party group took less time to complete their tasks than the solo marking group. Most interesting of all was the analysis of our 'fake' exams. We found that the marking party group made fewer errors and were, on average, closer to the expected mark than their solo marking counterparts.

2 BACKGROUND

Although a reasonable amount has been published documenting the effectiveness of TA training [2, 5, 6, 12], and a wealth of research has been done on qualitative analysis of marking [7], research focusing on consistency of TA's marking has been much more sparse.

Shannon et al. studied the correlation of TA training and students' perceived effectiveness of the TA by providing a questionnaire to the TAs asking if they had prior teaching experience [8].

Wiley and colleagues, studied student and TA perception of grading before and after group discussion, and found that both groups felt that they've been marked more fairly and received better quality feedback when the TAs had first discussed the marking as a group [9]. Further research also showed that group discussions improved both consistency of marks (as measured by difference in average mark between tasks), and confidence of the markers [10, 11].

Huang et al. developed a profile for a marker over many courses, and found that assigning markers to different groups based on attributes of their profiles resulted in a small decrease in marking errors [3].

One study focusing on the inter-marker consistency is similar to the solo-marking portion of the experiment detailed in this study. A group of 5 TAs was assigned the same paper to grade with no communication or discussion, and the experimenters found that the standard deviation was worryingly high [4].

3 METHODOLOGY

Exam Design

CSCA08: Introduction to Computer Science I is a CS1 course offered at University of Toronto for students planning on pursuing a major or specialist degree in computer science. In the Fall term of 2017 564 students wrote the final examination. 20 teaching assistants (TAs) were assigned to the course, all of whom were senior undergraduate CS students or graduate students in the department. All TAs taught weekly tutorial sections and supervised labs, and also underwent training in both course content and teacher training that included advice and guidance on marking [2]. All TAs marked 5 weekly quizzes and 3 programming assignments throughout the term as well as 2 term tests prior to the final exam and had participated in at least one 'marking party' and one 'solo marking' session prior to the commencement of this study.

The exam consisted of 4 questions[1] worth a total of 54 marks. The questions were:

- **ADT (10 marks)**: Students were provided with code to implement a standard Queue abstract data type, and asked to explain why the given code failed to provide good quality abstraction, and then to fix the code to avoid abstraction leaks.
- **Tracing (10 marks)**: Students were given code (focusing on understanding of OOP in Python) and asked to reproduce the output of the code.

[1] There was actually also a 'Question 0' which was worth 1 mark, and awarded for correctly filling out identification information on the cover sheet, but this question was ignored for the purpose of this study

- **Mangler (14 marks)**: Students were provided code that had been mangled by the “Code Mangler” [1], and asked to recreate the original code.
- **UML (20 marks)**: Students were provided with user requirements as a simple text dialogue, and asked to draw a UML class diagram for a proposed system.

A copy of the exam as well as the raw marking data can be found at http://uoft.me/markingparties.

Experimental Setup

The 20 TAs were randomly assigned to either the marking party group or the solo marking group. The solo marking group was told that they could pick up their exams for marking from a secured drawer with a combination lock at any time during a 4 day period. They could mark the exams in the TA office or take them to one of the study rooms in the building, but the exams could not leave the campus. Once the exams were marked, they were to enter the marks into a database and return their papers to the drawer. The marking party group was given a time and place to meet (after finding a time that was amicable to all the TAs involved), and all marking would be done over the course of a single day.

Members of both groups were each provided with a stack of 28-29 exam papers, and a marking scheme with instructions on how to mark each question. They were provided with identical instructions for how to mark, including the line: *“If you are unsure about anything, please ask for clarification rather than guessing."*. Since the course instructors were part of the experiment, in order to avoid undue bias, the TAs in the marking party were told that they had urgent business to attend to elsewhere and so would not be present during the marking party as is the normal custom. But one instructor would check in periodically during the party, and was available via e-mail to markers in both groups.

In addition to the regular student exam papers, each marker received 2 ‘fake’ exams (one placed 2 tests from the top of their pile and one placed 2 tests from the bottom). These fake exams, were hand written and assigned a fake name and credentials [2], and created according to a specific rubric. The rubric was designed to ensure that all fake tests should receive the same mark and make the same errors, but still allowed for enough variation in answers so as not to appear suspiciously similar. The rules for creating fake tests are given in Figure 2[3]. The corresponding marking scheme given to TAs is shown in Figure 1.

40 total exams were hand written by the authors using the rules shown in Figure 2. Independent analysis by two experienced TAs not associated with the study confirmed that a marker carefully following the marking scheme and instructions should have awarded these exams a final score of 34/54. These exams were then randomly assigned to the TAs, with each TA receiving two fake exams to mark.

Once all tests had been marked and collated, the markers were asked to complete a short survey asking them to report the total time they had taken to complete all exams, not including breaks or preparation, as well as a series of questions about the marking setup. The marking party and solo marking arrangements were defined and explained (no mention was made at this point of the fake tests, only the logistics of the two testing setups were included), and the markers were asked a series of questions including:

- Which marking setup would you have personally preferred?
- Which marking setup is faster?
- Which marking setup is more fair from a marker’s perspective?
- Which marking setup is more fair from a student’s perspective?
- Which marking setup should we use in future?

Markers were also given a freeform text section in which they were told they could explain or expand upon any of their answers.

Research Ethics

As this project involved not only the use of student exam information, but also the willful deception of teaching assistants, and the addition of extra work on the part of the TAs, ethical considerations were important in our experimental design.

There was an additional concern that, should the results show a systematic bias towards students marked by one of the marking groups, an adjustment would need to be made in the final exam marks.

For the handling of student data, normal protocols at the University of Toronto were followed. Students in the course had opted into an experimental consent form at the beginning of the term in which they consented to their exam grades being used in an aggregate fashion, properly anonymized, for pedagogical and educational studies.

In order to gain the consent of the teaching assistants, they were told they would be participating in a study on marking effectiveness. They consented to self-report their own marking time and to participate in a post-study survey a priori. They also consented to a potentially increased marking load, not to exceed 5% of their regular exam marking duties (this also fell well within their contractual time allocation for the course). They were not provided with the details of the purpose of the study, as this would have possibly biased their performance.

Upon completion of the study, all parties were informed of the purpose of the study, and the resulting data was made available to both the students and the TAs.

A post hoc analysis of the “real” student exams found that the bias towards leniency in the solo marking group was still present, but was not statistically significant. A secondary marking check was conducted for elements of the marking scheme found to be regularly mis-marked, but no global adjustment was made. It appears that the mistakes made by the students were generally not nuanced enough to be falling through the gaps in the marking scheme at the same rate as our targeted fake exams.

4 RESULTS

The experiment ran as expected for the most part. No markers showed any signs of suspecting that the fake submissions weren’t genuine or appeared to have treated them any differently than the

[2] In an earlier experiment, we attempted to use photocopied examinations, but this immediately raised suspicions among the markers, we also found that it was important to match the name to the writing style, as one subject mentioned ‘there’s no way this is female hand-writing’

[3] Some of the rules have been paraphrased slightly from what was given to the TAs in order to provide details that would have been clear to the markers in context.

Q1:

- 1 mark for mentioning private variables, 1 mark for correcting (adding underscores)
- 1 mark for explanation of why private variables are important, 1 mark for example
- 1 mark for explanation of why private variables are important, 1 mark for example
- 1 mark for mentioning abstraction, one mark for explaining what it means
- 2 marks for creating error classes
- 2 marks for correctly raising errors
- 1 mark for modifying docstring

Q2:

- 1 mark off for every deviation from solution (maximum of 1 mark for formatting)
- block deviations should count as a single deviation. e.g., if all A's output comes before B, that's only 1 mark off
- repeated errors should only be penalized once. e.g., if they swapped X with Y, don't penalize repeated instances of Y where there should be X.

Q3:

- Docstring
  - 1 for type contract
  - 1 for examples
  - 2 for description (1 for clear/concise, 1 for having all necessary info)
  - Watch out of examples/description copied directly from handout
- Internal Commenting
  - 2 marks for nested loops
  - 1 mark for explaining if structure
  - 2 marks for explaining looping structure
- Code
  - 1 mark off for each deviation from solution
  - Block deviations count as a single deviation

Q4:

- Classes
  - 1 mark off for each missing class
  - Do not take marks off for extra classes if they are reasonable
- Relationships
  - 1 mark off for each missing/incorrect relationship
- Formatting
  - 1 mark off for missing/non sensible relationship names
  - 1 mark off for missing/incorrect cardinalities
  - 1 mark off for missing/incorrect inheritance arrows
  - 1 mark off for missing privacy
  - 1 mark off for missing type contracts
- Methods
  - 1 mark off for each missing/incorrect method
  - Max 1 mark off for missing initialization methods
  - Don't take marks off for extra methods if they are reasonable
  - Do take off marks if method appears in child class when it should only be in parent

Figure 1: Marking Scheme Provided to TAs

rest of their papers. There were several requests for clarification from both the marking party and the solo markers, and in all cases e-mails clarifying points were sent to all members of both groups. None of the questions seemed to be specific to any of the fake papers from our study.

One issue that did arise was that three of the markers in the solo marking group decided to mark during the same time as the marking party was being held, and decided to join the marking party. Since the original pretence offered to the TAs for why some were marking solo while others were in the marking party was logistic in nature (since many of the TAs are friends outside of school, the information was bound to leak out, so we simply told those in the marking party that the others weren't there due to scheduling conflicts), there was no good reason to exclude them.

Q1:

- Add underscores to the variables and say that the code "breaks the ADT" or "violates the abstraction" or some combination of the two, but do not explain how or why
- Include an error class, but don't give Error as the parent class instead of Exception
- When raising the exception, don't include an error message

Q2:

- Replace every `A` with either a `B` or a `C` in the 3rd line of the output
- Move the parent section of the last line to the end so it appears after the grandparent output

Q3:

- Copy the description from the handout *almost* exactly, just piecing together the words we gave them, but in a way that is very clearly not "your own words"
- Have your comments repeat exactly what each line of code is doing without providing any additional context or information
- Un-indent the second for-loop so that it could be mistaken for not being inside the outer loop

Q4:

- Have type contracts in most of your methods, but skip 2-3
- Include at least one class from a word that is in the handout, but shouldn't really be a class (e.g., hospital, needle)
- Put the cardinalities backwards on the relationships
- Don't include any initialization methods

Figure 2: Rules for Creating Fake Exams

Table 1: Self Reported Time to Grade 40 Exams

	Party	Solo
Avg. Reported Time	7 hr 30 min	8 hr 45 min
Max. Reported Time	8 hr 30 min	10 hr
Min. Reported Time	7 hr	5 hr

This means that rather than having 10 TAs in the marking party group and 10 in the solo marking group, we ended up with 13 in the marking party group and 7 in the solo marking group. The potential problems this may have caused for our data collection and our analysis are further discussed in section 5.

Time Taken

The markers did not strictly monitor their time, but they were asked to self-report how long the entire marking process took in 15 minute increments, excluding breaks or distractions. This is obviously not a rigorous result, but post study discussion indicates that the TAs felt confident that their answers were accurate to within a window of between 15 to 30 minutes. Furthermore, it may be argued that the actual time taken to mark exams is less important than the perceived time from the graders' perspective, if the goal is to analyze marking setups in terms of preferences and TA morale.

The average reported time to mark all 30 exams (28 real and 2 fake) for the solo marking group was **8 hours 45 minutes**, the time reported for the marking party group was **7 hours 30 minutes**. The times reported for the marking party were much more consistent, with all reports being between 7 hours and 8 hours 30 minutes, while the times for the solo marking group ranged from 5 hours 30 minutes to 10 hours. The details are summarized in Table 1.

While it is difficult to verify the veracity of this data, as it was not rigorously measured, and students may have motivations to over/under report the time taken for grading, it is interesting to note that the average time in the solo marking group was over an hour longer than the marking party. This runs counter to our initial hypothesis that the added distractions of group marking would slow down the process, making the marking party group take longer to complete their tasks.

TA Surveys

In the survey administered at the end of the study, it was clear that the students believed that the marking party setup was superior in almost every respect. There was very clear personal preference for marking parties, as well as a belief that it was faster and more fair from the student perspective. The only question that did not show a clear preference for the marking party was *"Which marking setup is more fair from a markers perspective?"*. In post study discussion, it was clear that this question was interpreted as ambiguous by the markers. Some of them were unclear as to whether we were asking about the "traditional" marking party where students just worked on arbitrary stacks of papers and no one kept track of who marked which papers, or the style of marking party we were using in the experiment, where each TA had a set number of papers to mark. Many of the TAs also expressed uncertainty as to what exactly was meant by "fair" in this context. The results of the survey can be found in Table 2.

In the freeform text section of the survey, many of the markers discussed the difference between this style of marking party and the traditional style. With several commenting to the effect that while this style is more fair (with each marker being responsible for the same number of papers), the traditional style allowed them

Table 2: Marker Answer Counts for the Prompt: *"Which marking setup..."*

	Marking Party	Solo Marking	Don't Know/No Opinion
would you have personally preferred?	17	2	1
is faster?	13	5	2
is more fair from a marker's perspective?	5	7	8
is more fair from a student's perspective?	20	0	0
should we use in future?	20	0	0

- *"Once we get a good feel for the questions, it's more efficient to do it solo without distraction, but staying in a group ensures that any last quirks in the questions can be sorted out with a full consensus."*
- *"Being able to mark on own time allows for more flexibility. Being in a marking party allows for quicker consensus on how questions are marked."*
- *"If this were a course I had TAed many times, I might prefer to mark on my own, but I think the marking party is really good for first timers so we can be sure that we're doing things the right way"*

Figure 3: Selected Comments from Marker Feedback

to focus on a smaller number of questions for a larger number of papers, which was generally interpreted to be faster and require less effort overall. Selected comments can be found in Figure 3.

Inter-Grader Consistency and Examination Errors

In order to analyze whether the marking setup has an effect on the quality of the marks received, we first looked at the total marks given by each grader, to see which marking setup would result in our fake students receiving grades that deviated from what they should have received had the marking scheme been strictly followed.

The question error was defined as the absolute value difference between the grade given on each question, and the expected grade provided by the experienced TAs carefully following the marking scheme. Each question error was then summed for each test, and for each marker, their two errors were then added to produce a total error for each participant.

Some of the total errors are alarmingly large [4]. However, it is important to remember that the fake papers were designed specifically to contain as many borderline, ambiguous, or easy to miss errors as possible, and that the total error is a sum of the absolute errors, so in practice many of these would cancel out (as we will later see). The total error for all graders is given in Table 3.

It is immediately obvious that the total errors in the solo marking group are consistently higher than those in the party group, with an average error 5.8 marks higher. A boxplot of the two distributions is shown in Figure 4.

Due to the low sample size, we first ensured that both groups passed the Shapiro-Wilk test for normality ($W_{solo} = 0.86$ $p_{solo} = 0.16$, $W_{party} = 0.96$ $p_{party} = 0.7$), and an F-test for homogeneity of population variance ($F = 1.48$, $df_{num} = 6$, $df_{denom} = 12$, $p = 0.53$).

Table 3: Total Examination Errors (Sum of absolute errors on all questions)

Grader	Test 1 Error	Test 2 Error	Total Error
Solo1	4	10	14
Solo2	12	10	22
Solo3	11	10	21
Solo4	5	13	18
Solo5	8	3	11
Solo6	7	5	12
Solo7	11	11	22
		AVG:	17.14
Party1	8	9	17
Party2	5	3	8
Party3	4	8	12
Party4	3	4	7
Party5	2	2	4
Party6	6	5	11
Party7	5	6	11
Party8	5	11	16
Party9	3	6	9
Party10	3	6	9
Party11	8	8	16
Party12	6	6	12
Party13	5	10	15
		AVG:	11.30

The total error distribution in the two groups was shown to be statistically significant by both a t-test with pooled variance ($t = 2.95$, $df = 18$, $p = 0.0087$), and a Wilcoxon rank sum test with continuity correction ($W = 75$, $p = 0.021$). From this data, it is clear that the marking party was statistically less likely to produce examination errors than the solo marking group.

Analyzing Intra-Grader Consistency

In addition to evaluating the inter-grader reliability and average error of the marking setups, we also wanted to evaluate the effect on the marking within an individual marker's papers. It was for this reason that each marker received two fake papers, one near the top of their pile and one near the bottom. The assumption was that the markers would evaluate each paper in the order they were presented in the bundle. However direct observation as well as self reporting in post study discussion revealed a slight flaw in that assumption. Generally the markers completed marking all of the

[4] Author's Note: We have set a personal wager on how long it will be before the first student of ours tries to use this paper to justify an increase to their exam grade

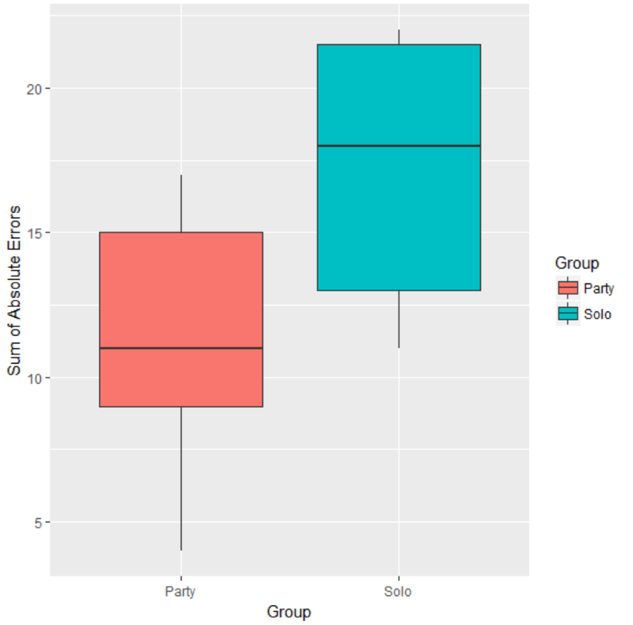

Figure 4: Total Examination Errors vs Marking Group

first question before moving on to the second, and as each paper was removed from the top of the pile and marked, it would be transferred to a second pile. Some of the markers would keep the papers oriented the same way and some would turn them over. Thus when marking the second question, some markers would be marking the papers in the same order, and some would mark in the opposite order. For this reason, the actual order of the fake papers within the pile did not produce any significant results. However, the fact that they were well separated within the piles seemed to have the desired effect in that none of the markers wound up marking two of the same 'fake' questions in quick succession.

Our exam error evaluation only took into account the absolute error of the final mark and ignored the direction. However, it is possible that markers in one group were either consistently over or under-marking, while markers in the other group were being less consistent. For this reason, we analyzed the mark range of each marker.

For this study, we define the mark range as the difference in mark assigned to a question on the two papers evaluated by a single marker. If a marker is consistently awarding marks above the expected average or consistently below, they will have a high error, but a low range. If a marker is awarding some marks above the expected average and some marks below, they may have a lower error, but will have a higher range. The mark range for all markers is given in Table 4.

Once again, the average range for the solo group was found to be higher than the party group. Though in this case, the party group appears to have a larger variance. A box-plot for the two distributions is given in Figure 5.

We attempted to discover whether the difference in total mark ranges was statistically significant for the two groups. Due to the small sample size, we again ensured that the distributions passed

Table 4: Mark Ranges
(Difference in marks awarded on separate tests)

Grader	Q1	Q2	Q3	Q4	Total
Solo1	4	1	2	1	8
Solo2	2	2	0	2	6
Solo3	1	0	8	0	9
Solo4	3	4	2	1	10
Solo5	1	5	0	1	7
Solo6	2	0	6	4	12
Solo7	4	2	1	1	8
				AVG:	8.57
Party1	1	2	2	0	5
Party2	0	0	3	3	6
Party3	1	1	6	4	12
Party4	2	0	3	2	7
Party5	1	0	1	2	4
Party6	1	0	3	3	7
Party7	1	0	0	0	1
Party8	2	9	0	3	14
Party9	1	2	1	5	9
Party10	0	0	2	1	3
Party11	2	2	0	0	4
Party12	0	0	1	1	2
Party13	2	0	6	1	9
				AVG:	6.38

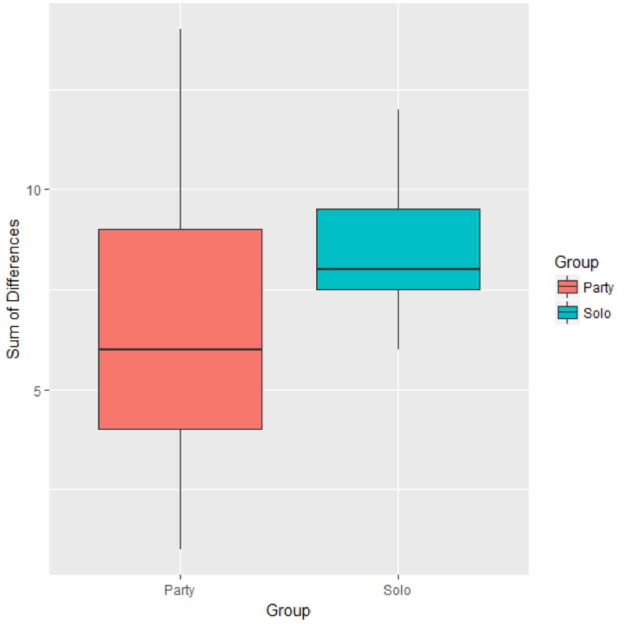

Figure 5: Total Mark Ranges vs Marking Group

Table 5: Question Error (Deviation From Expected Question Mark)

	Solo Marking			Marking Party		
	Min Error	Max Error	Average Error	Min Error	Max Error	Average Error
ADT	0/10	7/10	2.21/10	0/10	4/10	0.93/10
Tracing	0/10	5/10	2.43/10	0/10	5/10	1.31/10
Mangler	1/14	5/14	3.21/14	0/14	5/14	2.31/14
UML	0/20	4/20	2.00/20	0/20	4/20	1.96/20

both the Shapiro-Wilk test for normality ($W_{solo} = 0.97\ p_{solo} = 0.88$, $W_{party} = 0.96\ p_{party} = 0.73$), and an F-test for homogeneity of population variance ($F = 0.27, df_{num} = 6, df_{denom} = 12, p = 0.12$). However, in this case it was not possible to establish statistical significance of the difference between the two populations by either a t-test with pooled variance ($t = 1.40$, $df = 18$, $p = 0.18$), or a Wilcoxon rank sum test with continuity correction ($W = 65$, $p = 0.13$).

While it does appear that the marking party group may have been generally more internally consistent, due to the higher average, the small sample size and high variance, it is not possible to say that the effect was statistically significant.

Question Analysis

Examination error and mark range both look at the papers overall. However it is possible that the differences between the groups came from individual questions rather than being a consistent and persistent problem. For this reason we also evaluated individual questions. The minimum, maximum and average error for each question by each marking group can be found in Table 5.

Once again, we find that the marking party has lower average per-question errors on all questions except for the one on UML. A two-way ANOVA failed to show any statistically significant results, which is not surprising given the very small sample sizes. However, it is clear from these results that the differences found in the previous tests were not isolated to a single question or a specific error, but rather appear to be systematic across multiple elements of the marking scheme.

5 THREATS TO VALIDITY & FUTURE WORK

This study was as rigorous and as authentic as possible under the circumstances. However, there are several possible threats to validity. The marking party in the study was slightly different from our traditional group marking in that the TAs were not allowed to 'trade' marking in a 'I'll mark your question A if you mark my question B manner', and the solo markers were not allowed to take the papers out of the building. This means that our findings may not generalize to more traditional marking setups where group marking is less regulated, and individual graders are allowed to work from home with all the comforts and distractions that may imply.

The fact that some of the markers assigned to the solo marking setup decided to attend the marking party could be a contaminating factor in some of our analysis. As it is possible that either the imbalance of the group sizes led to difficulties in the statistical analysis, or that there is something about the personalities of the markers who chose to attend even though they were not assigned to the marking party could be introducing bias, as the participants were no longer randomly assigned to groups. The data does not seem to show any obvious difference between these three participants and others in the experimental group, but this could simply be due to data sparsity.

Since we had to hand-create each fake paper, minor variations in answers could have accounted for noise in the data. Something as simple as different handwriting could impact some of the more subtle points of marking. Unfortunately, simply photocopying the results led to immediate detection in our pre-study, and even having identical answers with different handwriting led to graders becoming suspicious of possible plagiarism. This threat to validity is also a barrier to replication, as our team spent approximately 25-30 hours hand-crafting the fake papers. Future work will look into ways that this study could be replicated without the need for manual creation of so many papers. If possible, this could allow us to replicate the study across various types of exams, groups of TAs, and levels of details in the marking scheme.

Another avenue for future pursuit is to analyze marker's behaviours over time. Do TAs eventually get better at collaborating in marking parties and enhance the differences from the solo markers? Or do the solo marking TAs eventually get better at avoiding distraction, and therefore reduce the difference between the setups? To evaluate this we would need to assign TAs to a particular marking setup, and have them continue with that setup throughout the year in order to track the change in the various effects.

6 CONCLUSION

In this paper we set out to evaluate marking parties and compare them to solo marking on four metrics. We were able to provide a convincing qualitative analysis that TAs preferred marking parties, and at least perceived them as being more time efficient. Our experimental results showed a possibility that marking parties resulted in better intra-marker consistency, however the data was not conclusive. We were able to show clear, statistically significant evidence that the marking parties resulted in better inter-marker consistency, with fewer errors.

This study was performed on a single group of TAs, for a single course. It is not obvious that the results will generalize to other institutions which may have different marking requirements and TA cultures. However, we feel that this study does provide valuable evidence to support our initial supposition that the benefits of marking parties, are worth the extra setup time, logistical effort, and strain on the departmental pizza budget.

REFERENCES

[1] Nick Cheng and Brian Harrington. 2017. The Code Mangler: Evaluating Coding Ability Without Writing Any Code. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE '17)*. ACM, New York, NY, USA, 123–128. https://doi.org/10.1145/3017680.3017704

[2] Francisco J. Estrada and Anya Tafliovich. 2017. Bridging the Gap Between Desired and Actual Qualifications of Teaching Assistants: An Experience Report. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '17)*. ACM, New York, NY, USA, 134–139. https://doi.org/10.1145/3059009.3059023

[3] Zhuhan Jiang1and Jiansheng Huang. 2017. Improving Fairness On Students'overall Marks Via Dynamic Reselection Of Assessors. *International Journal on Integrating Technology in Education (IJITE)* 6, 3 (June 2017).

[4] E. A. Jackson. 1988. Marking Reliability in B.Sc. Engineering Examinations. *European Journal of Engineering Education* 13, 4 (1988), 487–494.

[5] Sasha Nikolic, Peter James Vial, Montserrat Ros, David Stirling, and Christian Ritz. 2015. Improving the laboratory learning experience: a process to train and manage teaching assistants. *IEEE Transactions on Education* 58, 2 (2015), 130–139.

[6] Chris Park. 2004. The graduate teaching assistant (GTA): Lessons from North American experience. *Teaching in Higher Education* 9, 3 (2004), 349–361.

[7] D Royce Sadler. 2010. Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education* 35, 5 (2010), 535–550.

[8] David Shannon, Darla Twale, and Mathew S. Moore. 1998. TA Teaching Effectiveness: The Impact of Training and Teaching Experience. *The Journal of Higher Education* 69 (07 1998), 440–466.

[9] Keith Willey and AP Gardner. 2010. Improving the standard and consistency of multi-tutor grading in large classes. In *ATN Assessment Conference*. Institute for Interactive Media and Learning, University of Technology Sydney, Sydney.

[10] Keith Willey and Anne Gardner. 2010. Perceived differences in tutor grading in large classes: Fact or fiction?. In *Frontiers in Education Conference (FIE), 2010 IEEE*. IEEE, S2G–1.

[11] Keith Willey and AP Gardner. 2011. Building a community of practice to improve inter marker standardisation and consistency. In *SEFI 2011 Annual Conference: Global Engineering Recognition, Sustainability, Mobility*. Instituto Superior de Engenharia de Lisboa.

[12] Stacy L Young and Amy M Bippus. 2008. Assessment of graduate teaching assistant (GTA) training: A case study of a training program and its impact on GTAs. *Communication Teacher* 22, 4 (2008), 116–129.