Discussions on University Science Teaching: Proceedings of the Western Conference on Science Education

Volume 1
Issue 1 *Proceedings of the 2015 Western Conference on Science Education*
Article 11

2017

Inter- and Intra-Rater Consistency: Armies of Graduate TAs Grading in First Year

Michael Moore
University of Guelph, mmoore01@uoguelph.ca

Daniel F. Thomas
University of Guelph

Inter- and Intra-Rater Consistency: Armies of Graduate TAs Grading in First Year

Michael Moore and Daniel F. Thomas
Department of Chemistry, University of Guelph

Abstract

In the introductory chemistry course for first year students at the University of Guelph, written answer questions are included on midterm and final exams, despite the logistical hurdles involved in grading 1800-2400 students' solutions. The process was improved with the development of a customized Scantron® form on which students wrote their answers. Teaching assistants (TAs) graded students' work and bubbled in the grades on the sheet which was then read into the computer. As well as improving the grade entry process, it also allowed for students to have their results (and grader comments) conveniently emailed to them. For this study, we have used one of the Scantron® fields as a "TA identifier" to correlate grading with the specific grader. This allows for comparison between TAs' grading rate, grade averages, variances, and distributions. Group and individual trends were also observed over time. Averages for four questions increased by 0.1-2.4 % points between the first and second hour of grading, and the probability of the more extreme findings occurring without a correlational link was 53-92%, based on χ^2 tests. In the same manner, the probabilities that differences in distributions between the population and a particular TA's sampling occurred by chance ranged from ~0-99 %. We discuss these results and how they may impact our confidence in the final grade assigned to a particular student. We also aim to use these results to develop new statistical treatments of inter-rater consistency for large sample sizes that require minimal to no exams to be graded multiple times, resulting in saving time when studying a large number of graders.

Keywords: grading, inter-rater reliability, marker accuracy, large courses

Introduction

In large courses at medium and large universities, questions requiring human graders are logistically difficult for several reasons: the personnel required to grade these questions, the returning of the students' work, and concerns about consistency among, and within, graders. The University of Guelph has made efforts to alleviate these issues. The introductory chemistry courses at Guelph (CHEM*1040 and CHEM*1050) are offered to approximately 2400 students in the fall semester and 1800 students in the winter semester. In many large courses, this grading problem is entirely addressed by resorting to a multiple-choice-only exam format. While our exams do consist of multiple choice questions, we have also retained a couple of pages of written answer questions which we feel addresses an important aspect of student assessment. This presents a challenge for courses that involve gathering about 30 Graduate Teaching Assistants (TA) together on the day following the exam and organizing them into marking teams to grade different pages of the exams. Previously, the grading was recorded by hand, the exams alphabetized, and the marks then entered manually into a spreadsheet to be collated with the multiple choice grades which were read from a Scantron® page completed by the student. The exams were subsequently placed in a common room in alphabetically organized piles and students invited to go to the room to pick up their exam.

Corresponding author: Daniel Thomas: dfthomas@uoguelph.ca

Recently, a modification to a standard Scantron® Test Scoring Answer Sheet was developed (Jones, 2013) that allowed for students to record their written answers on a Scantron® page that would be graded by a TA who would bubble in the student's grade for each question into a designated field that could be read by a Scantron® scanner (Figure 1). The images of the sheets would also be recorded as PDF files and sent electronically to the students, removing the need to sort and return the physical pages.

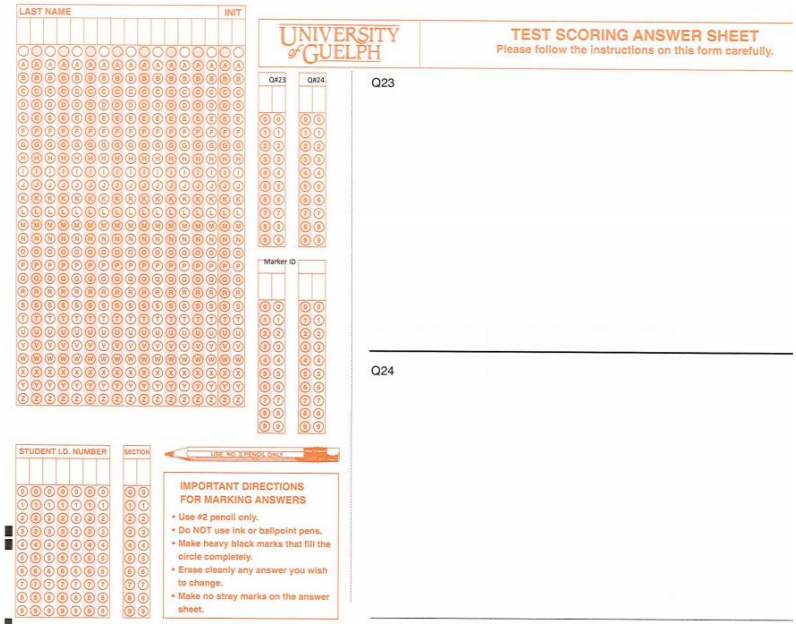


Figure 1. A sample test scoring answer sheet for short answer questions. Questions are listed in the examination booklet, and students record their work and answers in the white space on this page, to be graded by teaching assistants. The TAs bubble in the grade for each question in the fields on the left, denoted Q#23 and Q#24, as well as their assigned ID code in the Marker ID field.

The concern about grading personnel is unlikely to change in the foreseeable future, and the second concern has been effectively addressed. This allowed our group to focus on the third problem: consistency among graders. Specifically, we aimed to gauge intra-rater consistency (how consistent a TA grades during the marking session) and inter-rater consistency (how consistent a TA's marking is with respect to other TAs). In this work, "consistency" is used as a gauge of how similar assigned grade distributions are, often on a grader-by-grader basis. It is important to note that the quantitative values from this study (χ^2 values, and their corresponding p values) are conceptually related to, but not interchangeable with, reliability coefficients.

Intra-rater reliability (and consistency) has been studied in a variety of contexts and time schemes. Highly subjective ratings, such as judgement of Rorschach card results, were found to decrease in reliability and accuracy within only a 20-minute period (Cummings, 1954). Simpler tasks, such as sorting vegetables into categories based on preference, were found to not be affected within similar time periods (Bendig, 1955). Generally, grader fatigue is increased by repetitiveness of task, complexity of task, and time spent on task (without breaks) (Ling, Mollaun, & Xi, 2014).

In the context of grading English tests, grader fatigue was found to manifest itself through (a) an increased mean assigned score, (b) an increased variance of scores, and (c) a decreased acceptability threshold (Sprouse, 2007). With 6-8-hour work days (with variable break schemes), it was found that working for shorter times with more frequent breaks yielded both better grading speed and increased reliability.

A second consideration is the "warming-up" period. This is the time in which the grader is becoming accustomed to the work that they are doing, and results in an increase in grading speed with time. A typical work curve shows several stages, highlighted in Figure 2 (Anastasi, 1979).

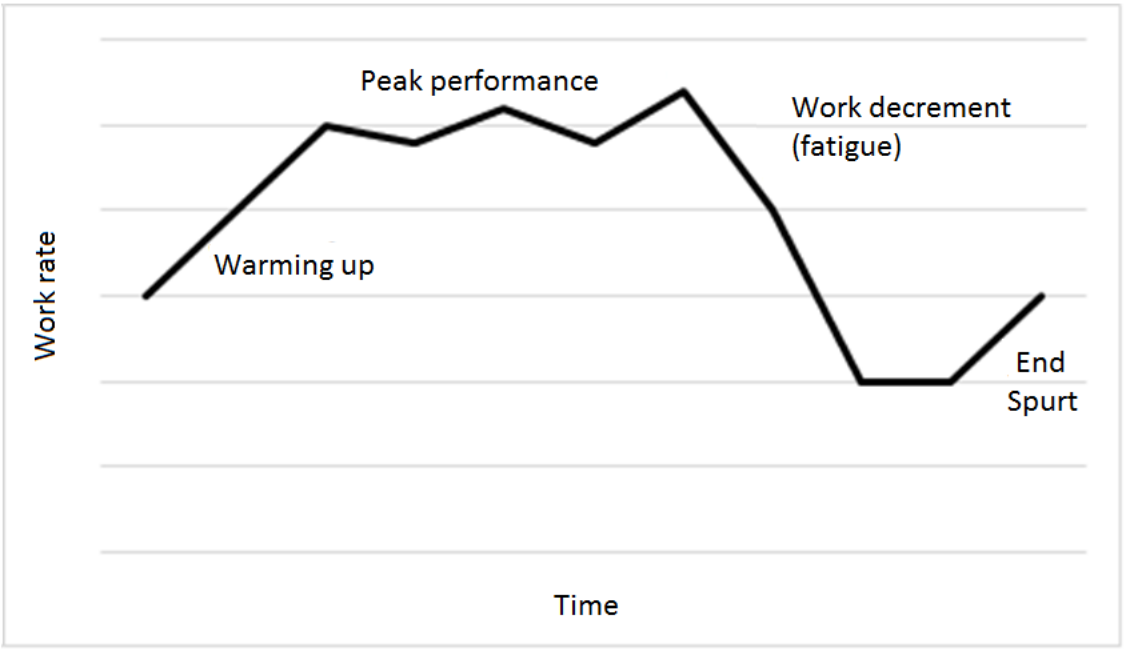


Figure 2. A theoretical work curve, illustrating common psychological trends (Anastasi 1979).

Inter- and intra-rater reliability are often measured using inter-rater agreement or Kappa coefficients. All of these require a crossed design: multiple raters rating identical items. Inter-rater agreement is typically calculated as a percentage, and is simply the number of times two or more raters' ratings matched, divided by the maximum number of potential matches. One limitation of using inter-rater agreement is that chance agreement can inflate one's confidence in the raters. Kappa coefficients (κ) normalize out chance agreement, and can be interpreted as reliability coefficients, as shown in equation 1 (McHugh, 2012):

$$\kappa = \frac{True\ variance}{True\ variance + variance\ from\ error} \quad (1)$$

In an educational context, this could be calculated by having each grader grade every, or every one of a subset of exams. A kappa value for each question could then be calculated using equation 2, in the case of Cohen's kappa (Cohen, 1960):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

Where:
p_o is the relative observed agreement between raters, and
p_e is the expected/chance agreement between raters.
Fleiss' kappa is a variation of Cohen's kappa for more than two graders (Fleiss, 1971).

Inter- and intra-rater reliability is prominent in both rubric-evaluation and medical diagnostics. The psychology department in Carolina State University developed a rubric to assess undergraduate students' final papers (Stellmack, Konheim-Kalkstein, Manor, Massey, and Schmitz, 2009). Two graduate students graded the same twenty papers each. There was generally strong inter-rater agreement, the areas of disagreement were able to be identified, and those aspects of the rubric further developed iteratively. A similar study, taking place in the Department of Engineering at Rowan University was successful in iterating a rubric to yield high inter-rater reliability (Newell, Dahm, & Hewell, 2002). The field of medicine has extensively explored inter- and intra-rater reliability. Examples include two raters judging the same 13 patients' lengths and angles of lunges (Bennell et al., 1998) to two psychologists diagnosing the same patients' social and communication disorders (Wing, Leekham, Gould, & Larcombe, 2002).

These situations deal with far fewer "items" (things being diagnosed, graded, or generally, rated) than our grading environment. And while day-to-day hospital operation deals with large numbers, the variability among their patients is great. In short, these situations have far more potential variability than our grading situations. This requires inter-rater reliability studies in those fields to use crossed studies, where the same item is rated by multiple graders. In a grading area containing up to 30 graders, whose times are often split between many tasks, having a large number of exams graded multiple times is less feasible.

This study aims to develop methods to evaluate inter- and intra-rater consistency without grading the same exam multiple times. It takes advantage of the relatively low variability in potential results of the short answer questions on chemistry exams, and relies on a fundamental assumption: as each TA grades an increasing number of exams, the grade distributions that they assign should resemble the total grade distribution that all TAs assign since each TA is randomly drawing papers to grade from the entire pool of students' papers. In particular, it looks to evaluate the comparison of each TA's:

(a) assigned grade distributions at times 2 and 3 to their own grade distribution at time 1 (intra-rater consistency)
(b) assigned grade distributions at times 1, 2, and 3 to their own overall grade distribution (intra-rater consistency)
(c) assigned grade distributions at times 1, 2, and 3 to the population of TAs' overall assigned grade distribution (intra-rater consistency)
(d) overall assigned grade distribution to the population of TAs' overall assigned grade distribution (inter-rater consistency)

Method

Data Collection

A total of 1603 exams with four sets of short answer questions were graded by a total of 25 TAs. Each question was worth 4 to 8 marks. The worth was written on the test scoring sheet shown above (see Figure 1) with space for additional questions on the back. TAs were assigned

unique identifying numbers (TAIDs) that they filled into a specified grading field on each. These TAIDs allowed for each grade for each question on the exam to be linked to the TA who graded it. Each TA had their ID number changed twice (50 minutes into grading and 1:45 minutes into grading) in order for changes over time to be observed. The marked sheets were then read by a Scantron® device which recorded the student grades, the marker ID, and produced a PDF file image of each side of the page. The grades were quickly collated with the rest of the exam marks, but also provided a rich set of data for this analysis.

Data Analysis

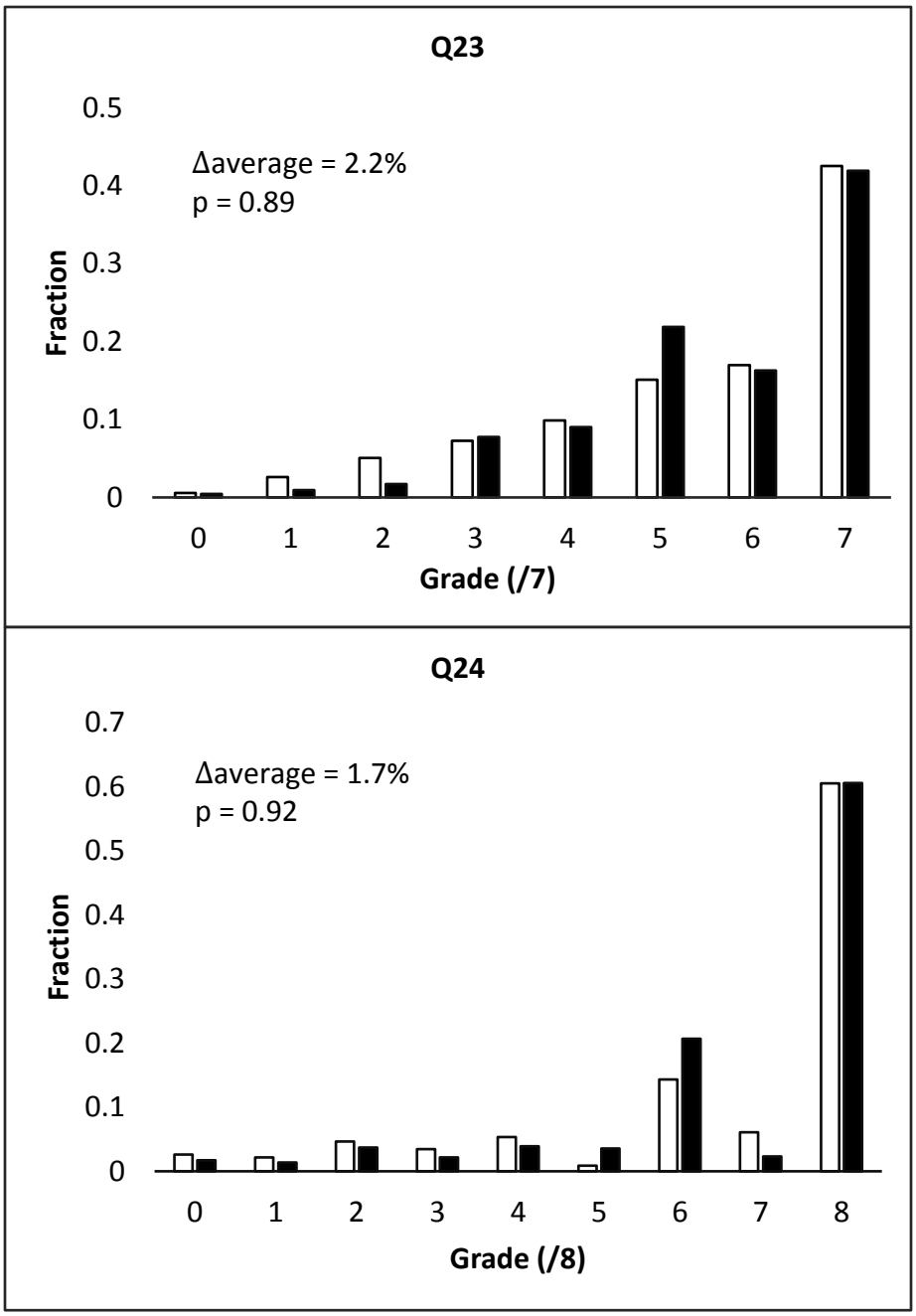
The bulk of this paper involves determining the probability that the grade distribution generated by a particular TA's marking matches that of the overall grade distribution. This is done by the calculation of a χ^2 value and the associated p value (for the particular sample size/degrees of freedom) using Microsoft Excel. The χ^2 test used compares two distributions of nominal variables. The associated p values represent the probability that, compared to a reference distribution, a distribution as different (or more different) as the sample distribution would be found. More loosely, the p value of the χ^2 test represents the probability that the difference between the sample and reference distribution could happen just by chance instead being the result of changes in rater performance as discussed above.

For this to be a valid inference, the TA must have graded randomly from the population, and graded a sufficient number of exams. In this study, the first assumption is somewhat true: through the transport, combination (exams were written in several rooms), and redistribution of exams, order initially present (e.g., polarized quality of early submissions, due to particularly strongly-prepared or under-prepared students) was mitigated. The latter assumption, which is necessary to reasonably expect that TAs' assigned grade distributions will reflect the overall grade distributions, is not discussed in this study.

Results and Discussion

General Results

The intra-rater consistency of the group was analyzed using data from the first two time periods. The third time period was excluded, as some TAs had left or began grading sections of the exam to which they were not initially assigned. The exam had four written answer questions – two answered on the front and two on the back. A given TA would be assigned to mark either the front or the back, with a different TA marking the other side later in the process. The average grade increased by 2.2, 1.7, 0.1, and 2.4% points for these four questions (labelled Q23 through Q26) between these two time periods. The grade distributions for each question are shown in Figure 3.


Q23
Δaverage = 2.2%
p = 0.89
Fraction
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7
Grade (/7)
Q24
Δaverage = 1.7%
p = 0.92
Fraction
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8
Grade (/8)

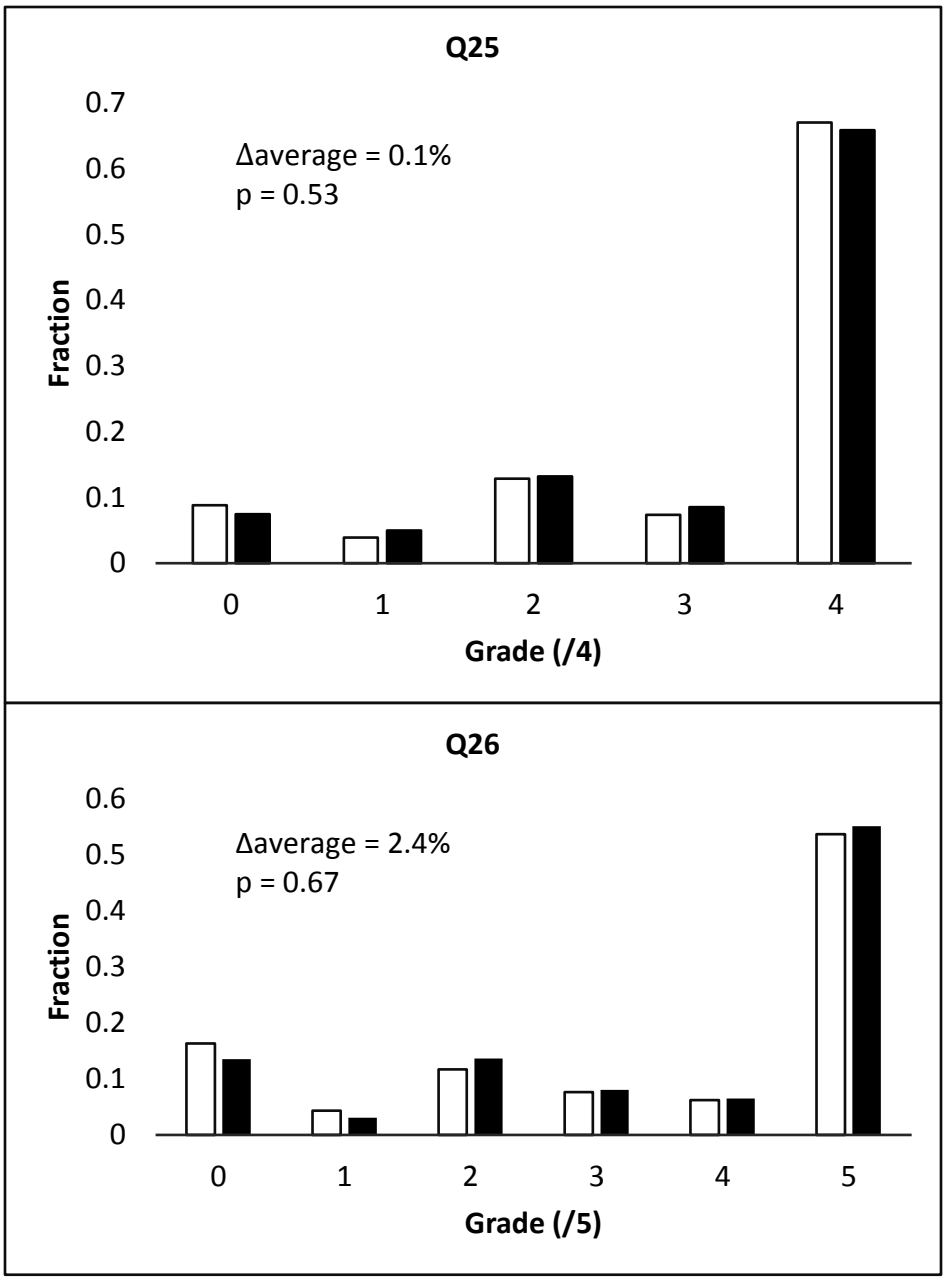


Figure 3. Summary of grade distributions of four midterm questions for the first grading time (blue, left) and the second grading time (orange, right). The change in the average performance (Δaverage), and the likelihood of the difference between the distributions being due to randomness (p, based on a χ^2 test) are shown. $\chi^2(7, N = 1333) = 2.93$, $p = 0.89$, $\chi^2(8, N = 1335) = 2.57$, $p = 0.92$, $\chi^2(4, N = 1335) = 3.20$, $p = 0.53$, $\chi^2(5, N = 1335) = 3.20$, $p = 0.67$ for questions 23-26, respectively.

While no group trends for intra-rater consistency were immediately obvious, several statistical methods were applied for further analysis.

To compare differences in distributions that would not be captured without a changing average (e.g., if 10 students got 4/5 in the first time period, and then five students got 3/5 and another five received 5/5 in the second), a χ^2 test was used. The p values in this study are the

probability of the difference between the observed and reference assigned grade distributions, or a greater difference, occurring by chance. As such, high p values indicate that changes in grade distributions are happening by chance (i.e., TAs are grading consistently). Based on the χ^2 tests, the group intra-rater consistency looked reasonably stable, with no p values below 50 %.

Several methods of quantifying individual intra-rater consistency were tested, and will be discussed here. In this preliminary analysis, four TAs' data were analyzed as a small-scale test to determine which types of further analyses are likely to be useful. Q23 is used arbitrarily for this analysis, and the number of question 23s each TA has graded are listed in Table 1.

Intra-Rater Consistency Relative to TA's First Time Period's Distribution

The first type of analysis compared each TA's grade distributions in the second and third time periods to their respective first period. The results are summarized in Figure 3. This kind of analysis has the advantage of being sensitive to any changes that occur through each TA's grading. It has two primary disadvantages: (a) it would be skewed if the first time period of grading is not indicative of the TA's true grading habits, and (b) it leaves only two time periods to be compared. The former problem could have been caused by the TA learning and adapting to the grading key without retroactively correcting mistakes, or by pulling an improbable distribution of tests (e.g., several blank tests). These problems would appear as both the second and third time period distributions being inconsistent with the first time period, but consistent with each other. This method of testing is the most sensitive to grader fatigue, or other factors varying consistently with time.

The grade distributions in time period 2 are significantly different from those in time period 1 for TAs 2 and 4, with p values of 0.00093 and 0.0067. No other p value for the χ^2 tests for any of the four TAs tested was under 20 %. Contrary to the implication of TA 2's second time period deviating significantly from the first, the grade distribution of the tests they graded in the third time period were very similar to those from the first time period.

Quantitative conclusions from these data should be made with caution. This kind of analysis assumes that in every time period, every TA graded enough questions for their personal grade distribution to be very similar to the population grade distribution, which has yet to be proven.

Intra-Rater Consistency Relative to TA's own Overall Distribution

A second, similar method to describe intra-rater consistency is to compare each TA's grade distribution at each time period to their own total grade distribution. This approach yields three data on all three time periods instead of only two, and is not skewed by a large error in the first time period. The disadvantage of this form of analysis is that it is somewhat self-referential: the TA's total grade distribution is composed of the sum of the distributions from the three time periods which would lead to increased p values of the χ^2 tests. This information is summarized in Figure 4.

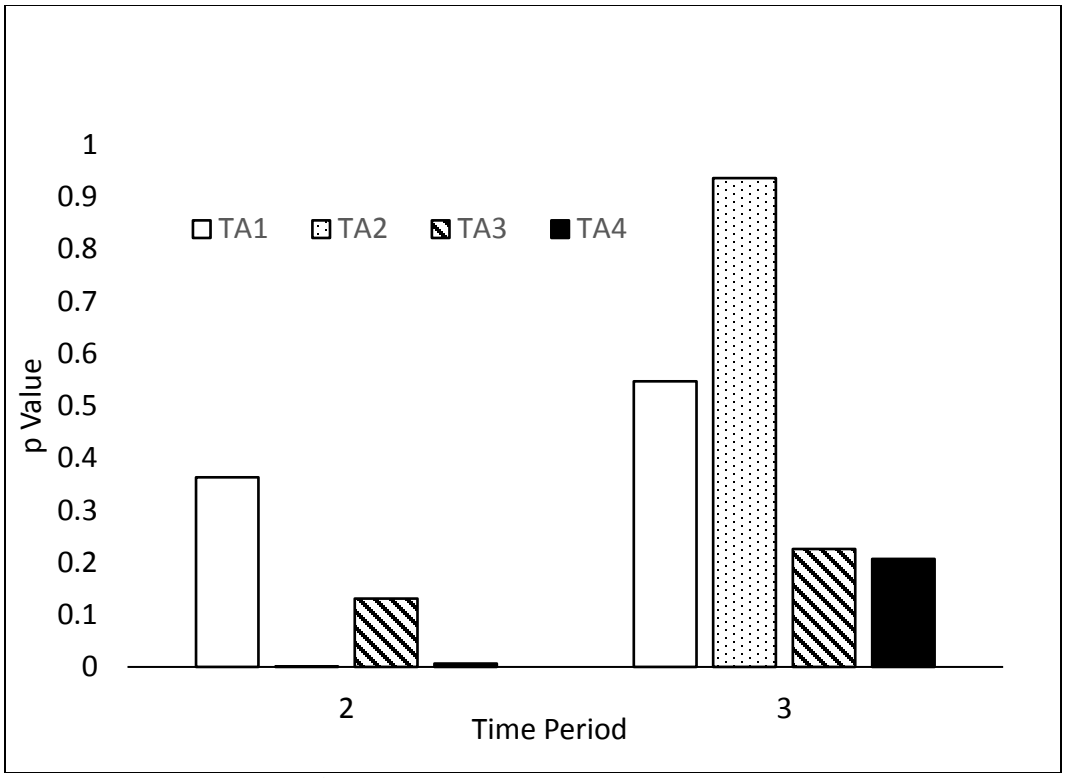


Figure 4. p-values based on χ^2 tests done on each TA's grade distribution from their second and third grading periods, relative to the first.

The graph shown in Figure 4 can roughly be interpreted as the level of self-consistency throughout the entire grading session a TA has exhibited. Again, quantitative comparison should be done with caution – each set of data for each time period is in reference to the TA's total grade distribution. The TAs graded at different rates throughout the session, and therefore quantitative comparison between TAs' data within the same time periods is inappropriate, as there will be different levels of self-reference between TAs. Qualitative analysis, especially involving trends that persist in all three time periods, is appropriate.

From these results, it is clear that TAs 1 and 3 were more self-consistent during the grading session than TAs 2 and 4. TAs 1 and 3 show little to no deviation or trending. TA 4 shows some deviation through the grading session, but no obvious trend. Conversely, TA 2 shows deviation, but a trend toward self-consistency.

Intra-Rater Consistency: Each Time Period Relative to Population's Overall Distribution

A final way of considering intra-rater consistency is through comparing inter-rater consistency over time. As this method of analysis requires inter-rater consistency information, and would be difficult to interpret without first understanding those data, we will now develop a few ideas regarding inter-rater consistency.

To compare inter-rater consistency, each TA's grade distribution was compared against the population grade distribution, and p values of χ^2 tests. The corresponding p values for these tests are depicted in Figure 5. These p values indicate that the agreement between each of these TAs' grade distributions and the population grade distribution is weak. None of the TAs had a p value greater than 50%, and most of them had p-values of under 20%. The probability

of finding a difference as extreme (or more) as TA 4's grade distribution from the population simply by chance was under 5%, assuming the null hypothesis.

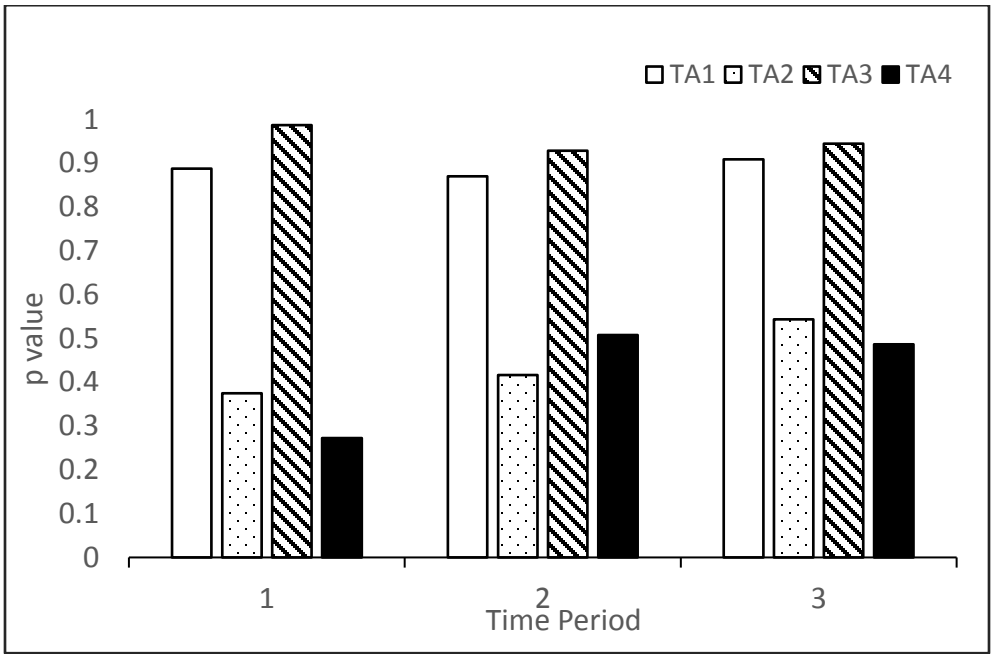


Figure 5. p-values based on χ^2 tests done on all of each selected TA's grading periods, relative to their total grade distribution.

Inter-Rater Consistency

Inter-rater consistency can be broken down by time period to compare how TAs' grade distributions evolve relative to the total population grade distribution. This information is summarized in Figure 6.

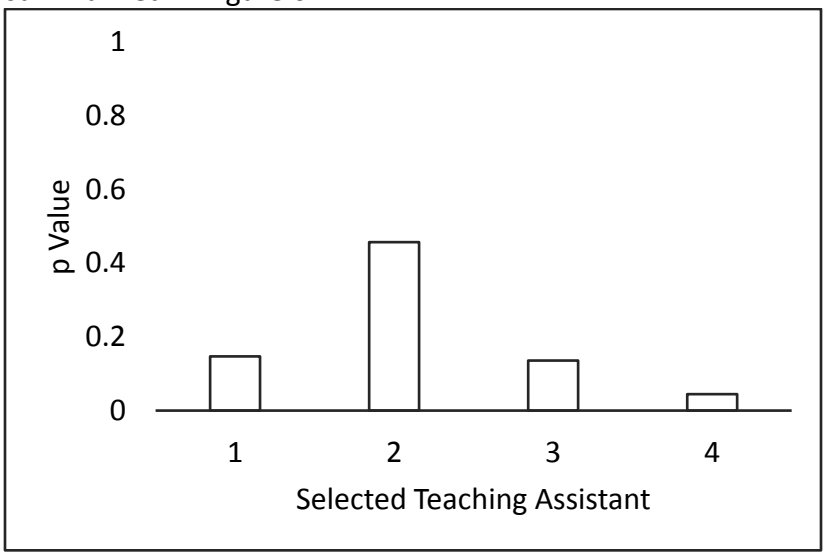


Figure 6. p-values based on χ^2 tests done on each selected TA's total grade distributions, relative to the total population grade distribution.

This method of analyzing inter-rater consistency allows for intra-rater consistency to also be considered. An advantage to this method over those previously shown is that it allows for all three time periods to yield useful, non-self-referencing data. Two disadvantages to this kind of analysis are that the trends are less sensitive to changes in intra-rater consistency, and that the p-values calculated using this method are not indicative of intra-rater consistency. Instead, only differences in these p-values are useful in discerning intra-rater consistency. These difference values are summarized in Figure 7.

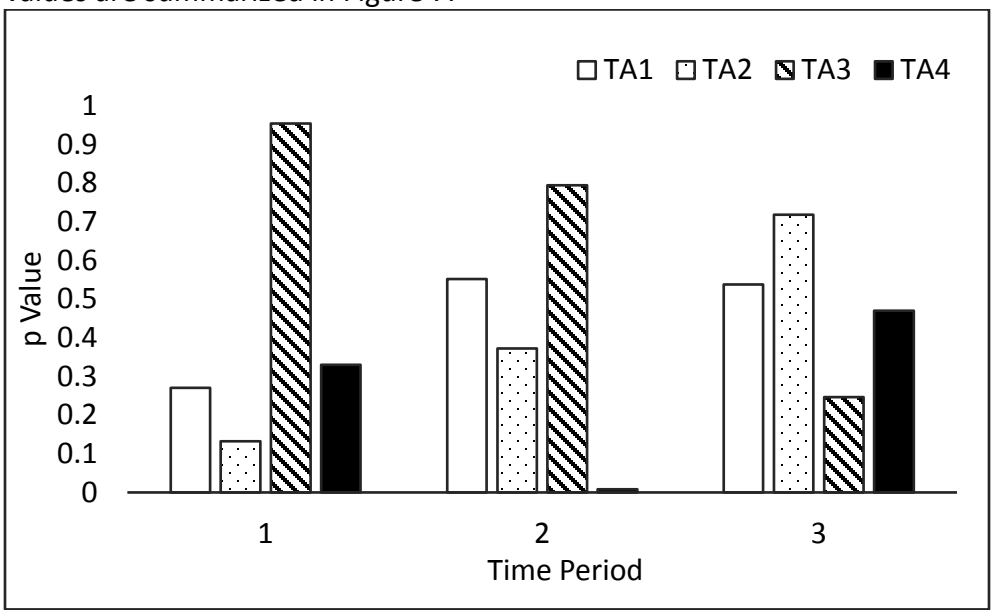


Figure 7. p-values based on χ^2 tests done on all of the grading periods for each selected teaching assistant, relative to the population grade distribution.

We can observe an individual's (TA1) grading pattern varying significantly between the second and third grading periods, which we might attribute to grader fatigue. On the other hand, we observe the other three TAs developing patterns that do not change during the same period, which could be interpreted as being indicative of their developing experiential consistency.

Conclusions

It should be emphasized that none of this work at this time directly addresses the ultimate question: "Is the TA grading correctly?" We are currently only addressing the question "Is the TA grading consistently?" We observe evidence where a TA's grading pattern changes with time, suggesting both the presence of marker fatigue (deviating from the norm through increased carelessness) as well experiential development (evolving towards the norm). We believe that our data supports our long held view that, on the whole, our mass marking scheme has been reasonably effective and consistent, though our final goal is to ensure that all students receive a fair and accurate assessment of their exams. Additional work will need to be done in this direction for there is some evidence that some substantial variation is observed with individual TAs though at this point we cannot say if that implies any faulty marking.

Three methods of analysis of intra-rater consistency were tested: evolution of each TA's grade distribution relative to both their distribution in the first time period and their overall grade distribution, and also changes in their inter-rater consistency as a function of time. Each

of these analyses yielded different information and trends, indicating that all three of them may have value in future work. These analyses showed grading fluctuation with time, which is consistent with the literature (Ling et al., 2014). Higher-level statistics, considering all TAs (instead of only three, as was done here) is necessary to be able to compare trends more reasonably with literature.

Future Work

This work requires a fundamental assumption to be fulfilled: that every TA graded enough items that their assigned grade distribution well-resembles the total grade distribution of the population. With varying grading speed, this may be a valid assumption for all, some, or none of the TAs sampled. The first priority of future work will be to check this assumption. This could be done through an iterative program, sampling the population distributions until a certain resemblance is reached. Once this information is obtained, quantitative conclusions would be possible.

An additional study using a small subset of student exams to be graded by all TAs would allow for the analyses used in this paper to be directly compared to those more prominent in literature. This kind of study has been planned, and is expected to be conducted this year. It will begin to provide some information about the accuracy of TA marking, in addition to the consistency. The current study was largely a proof of concept, considering only four of the 22 graders. Moving forward, more deliberate shuffling of exams will occur before distributing them to the graders. Additionally, higher-level statistics will allow for all of the TAs to be considered and at that point, multiple-comparison problems will be addressed. Other variables, including grading speed – which has been observed to vary between TAs by a factor of three or more – will also be considered.

References

Anastasi, A. (1979). *Fields of applied psychology* (2nd ed.). New York: McGraw Hill.

Bendig, A. W. (1955). Rater reliability and "judgmental fatigue." *The Journal of Applied Psychology*, *39*(6), 451-454.

Bennell, K., Talbot, R., Wajswelner, H., Techovanich, W., Kelly, D., & Hall, A. (1998). Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. *Australian Journal of Physiotherapy*, *44*(3), 175-180.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cummings, S. T. (1954). The clinician as judge: Judgments of adjustment from Rorschach single-card performance. *Journal of Consulting Psychology*, *18*(4), 243-247.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Jones, L. A. (2013). Technology assisted grade capture and electronic distribution of paper tests. *Teaching and Learning Innovations*, *16*. Retrieved from https://journal.lib.uoguelph.ca/index.php/tli/article/view/2788

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, *31*(4) 479-499.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-82.

Newell, J. A., Dahm, K., Hewell, H. (2002). Rubric development and inter-rater reliability issues. *Chemical Engineering Education*, *36*(3), 212-215. Retrieved from http://users.rowan.edu/~newell/Publications/16%20Rubric%20Development.pdf

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, *40*(2), 329-341.

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching Psychology*, *36*(2), 102-107.

Wing, L., Leekham, S. R., Gould, J., & Larcombe, M. (2002). The diagnostic interview for social and communication disorders: Background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry, 43*(3), 307-325.