# Mechanical TA: Partially Automated High-Stakes Peer Grading

James R. Wright
Dept. of Computer Science
University of British Columbia
Vancouver, B.C., Canada
jrwright@cs.ubc.ca

Chris Thornton
Dept. of Computer Science
University of British Columbia
Vancouver, B.C., Canada
cwthornt@cs.ubc.ca

Kevin Leyton-Brown
Dept. of Computer Science
University of British Columbia
Vancouver, B.C., Canada
kevinlb@cs.ubc.ca

## ABSTRACT

We describe Mechanical TA, an automated peer review system, and report on our experience using it over three years. Mechanical TA differs from many other peer review systems by involving human teaching assistants (TAs) as a way to assure review quality. Human TAs both evaluate the peer reviews of students who have not yet demonstrated reviewing proficiency and spot check the reviews of students who have. Mechanical TA also features "calibration" reviews, allowing students to quickly gain experience with the peer-review process. We used Mechanical TA for weekly essay assignments in a class of about 70 students, a course design that would have been impossible if every assignment had had to be graded by a TA. We show evidence that it helped to support student learning, leading us to believe that the system may also be useful to others.

## Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education—*Collaborative learning*

## Keywords

Computer Science Education, Peer Grading, Peer Review, Calibration, Scalability

## 1. INTRODUCTION

This paper describes our experience with software-supported, anonymous peer grading in a fourth year undergraduate course ("Computers and Society"). The course focuses on reasoning critically about the importance and social implications of computational advances. In earlier offerings of the course, students had to write three essays: on the midterm, final, and for a term project. However, shorter and more frequent essay writing assignments are both a more effective way to teach writing skills [12], as well as providing more opportunities to evaluate and improve writing skills and critical reasoning skills. Our past three offerings (2011, 2012, and 2013) thus shifted to a assigning students a total of 14 essays of about 300 words (11 weekly assignments, plus one essay on the midterm exam and two essays on the final exam). Manually marking essays is very expensive in terms of teaching assistant (TA) time. Furthermore, it can be difficult for students to learn to write such essays well. Peer grading offers a solution to both problems.

Peer grading is far from a new idea. However, students are often concerned that the quality and fairness of the evaluation that they receive from peer grading is lower than it would be from TAs [11, 8, 13]. Most systems (surveyed at the end of this article) attempt to address these concerns by evaluating the quality of the peer reviews in an automated way, whether by reweighting reviews based on some criterion [1, 5], by "review the reviewer" schemes in which students rate the feedback they have received [8, 3, 4, 2], by evaluating how close a review is to the combined "consensus" grade for an assignment [3, 5], or by some combination of these ideas.

We wanted to use peer grading to make more efficient use of TAs, not to replace them entirely. We thus designed a new system, dubbed "Mechanical TA."[1] Our system leverages (human) TAs in three ways. First, students start out in a "supervised" state, in which all of their reviews are marked by a TA. They are only promoted to an "independent" state when they demonstrate that they understand the grading rubric and are able to apply it competently. Second, students may use the system to appeal any peer grade that they consider unfair. (We reduce abuse of this feature by requiring a 100 word explanation of why a student believes that a review was unfair.) Finally, every independent review is eligible to be randomly spot checked by a TA, who can retroactively mark a reviewer's past reviews if they uncover a poor review. We found that students had surprisingly few concerns about fairness in Mechanical TA, and believe that the visible involvement of human TAs in marking assignments—especially in the early part of the class, when most students are supervised—was a major reason why.

Our system of random spot checks and appeals allows students to be persistently promoted. That is, once a student has been promoted, they can remain independent for the remainder of the class (i.e., if they are not demoted again due to a spot check or an appeal). This contrasts with systems such as Calibrated Peer Review (CPR) [1], in which students' review skills are retested at the beginning of each assignment. The time required to complete such calibration was a source of complaints in one study of CPR [13].


*SIGCSE* '15 Kansas City, Missouri USA
ACM 978-1-4503-2966-8/15/03 ...$15.00.
http://dx.doi.org/10.1145/2676723.2677278.

[1]Mechanical TA is freely available by contacting the authors.

Our implementation of calibration has a strong element of automated practice rather than just evaluation.[2] To our knowledge, this is a unique feature of our system of calibration. Students receive immediate feedback about their performance on calibration essays, and may optionally choose to perform many more than the required number of calibrations. In Section 4.2, we present our finding that calibration practice significantly improved students' review performance. In Section 3.3 we present evidence that it also improved students' writing performance, as measured by exam scores.

We begin by describing our particular peer review model in detail in Section 2. We survey our three years of experience with this model in Section 3, and compare some outcomes between different offerings of the course, paying particular attention to the differences between the most recent offering—which included automated calibration—and the previous two offerings. In Section 4 we analyze the data from our most recent offering. In addition to showing that calibration practice improved students' review skills, we also demonstrate that the persistent division of students into independent reviewers and supervised reviewers was an effective strategy. After reviewing some related work in Section 5, we conclude in Section 6.

# 2. PEER EVALUATION MODEL

In brief, our peer review system works as follows. Students submit their essays as free-form text in the Mechanical TA system.[3] After the essay submission deadline, each student is assigned three essays for double-blind peer review. After the deadline for submitting reviews, each essay is assigned the median peer-review mark. Students can register a request for a TA to regrade their essay if they believe that they received an unfair grade. The use of medians to compute grades means that an appeal is only worthwhile if the student believes they received two unfair reviews.

A review consists of a configurable set of text fields and multiple-choice questions. In our "Computers and Society" class, students were asked to rate each essay on a scale of 0–5 along four dimensions—following a detailed rubric that described what an essay would look like to justify each score in each dimension—and provide a textual justification of their scores. The grade assigned to an essay by a review is the sum of the scores in each dimension.[4]

Mechanical TA automatically assigns a selection of essays to the TAs for "spot checking", in which the TA reads the essay and evaluates its reviews. Every essay has a chance of being randomly selected, and *every* essay whose mark is above a configurable threshold is selected. This addresses a potential incentive problem, where a student could avoid the work of properly reviewing essays by simply giving every essay a high grade.

## 2.1 Supervised and Independent Reviewers

As mentioned above, we classify students as either *supervised* or *independent* reviewers. Every student begins as a supervised reviewer. Every essay that they review is also reviewed by a TA, and peer reviews are disregarded in this case for the purpose of grading: supervised essays are assigned grades from TA reviews. Furthermore, supervised students' reviews are also marked by TAs. Students are promoted from supervised to independent when their average review marks crosses a configurable threshold. Once promoted to independent status, a student automatically receives 100% on each of their reviews unless it is subsequently checked by a TA as described earlier, in which case it is graded.

Supervised reviewers are assigned only the essays of other supervised reviewers; similarly, independent students are only matched with each other. This is important in terms of TA workload: indeed, it minimizes the number of essays that must be read by the TAs who evaluate the supervised reviewers' reviews. If independent and supervised students could review each others' essays, then potentially *every* submitted essay would have at least one supervised reviewer and would hence need to be read. Conversely (and for the same reason), our scheme maximizes the number of essays that are fully peer graded.

## 2.2 Calibration

In addition to giving them the opportunity to learn by reviewing the work of their peers, Mechanical TA also allows students to practice reviewing via *calibration essays*. A calibration essay is an essay from a past offering of the course[5] which was carefully evaluated by multiple TAs to establish a "gold standard" review. At any time during the course, a student can request a calibration essay from Mechanical TA. The student then enters a review in the usual way. However, immediately after the review is submitted, Mechanical TA shows the student the gold standard review, and highlights the dimensions in which the student's review differed from the gold standard. If the student's review is within a configurable distance of the gold standard review, the student is given a "review point". After the student has collected enough review points over a configurable (potentially decaying) time window, they are promoted to independent status. This makes it possible for students to become independent before a TA has evaluated any of their reviews.

# 3. EVOLUTION OF OUR DESIGN

Our design of the Mechanical TA system evolved over time. Analyzing data from three consecutive offerings of Computers and Society allows us to argue that our current design helps to achieve better student outcomes. We have described the most recent version of the peer review process in Section 2. In the initial 2011 offering, each essay was reviewed by only two students; its mark was the average of the two reviews. In 2012, we switched to using the median of three reviews. In the most recent 2013 offering, we added the calibration process.

One of the major differences that calibration required was an extensively reworked rubric for reviewers. In the 2011 and 2012 offerings, reviewers were asked to rate each essay along 4 dimensions (Argument, Subject, Evidence, English) on a scale from 0 to 2. We offered minimal guidance about what separated 2/2 on a given dimension from 1/2. We found

[2]Indeed, our evaluation suggests that this aspect was the main benefit offered by calibration in our most recent course offering.

[3]This makes it easy for us to check all essays for plagiarism using TurnItIn, which we do.

[4]The full text of the rubric we used is available at `http://www.cs.ubc.ca/~jrwright/sigcse15/rubric.pdf`.

[5]Mechanical TA allows students to flag whether or not submissions may be used anonymously; we chose essays whose authors had permitted anonymous reuse.

that students were extremely reluctant to give 1/2 marks in this scheme, and received many comments that students did not want to deduct half the possible marks for a dimension. In the 2013 offering, we reworked the rubric in two ways. First, we expanded each dimension's scale to run from 0 to 5. Second, students were given explicit descriptions about what sort of essay deserved each score for each dimension.

In the remainder of this section, we first describe the process of setting up to offer calibration essays for the first time. Offering calibration reviews made a substantial impact, both on students' achievement, and on the workload for the TAs. In the final two subsections we compare 2013 to the earlier offerings by when students were promoted to independent reviewer, and by exam performance.

## 3.1 Calibration Setup

Constructing a library of calibration essays was a time-intensive process. We started by considering every essay from the previous offering that students had flagged as available for anonymous reuse. We then hand-selected 27 candidate essays. Each of these essays was reviewed by the same four TAs. The review marks were reconciled during in-person meetings, and every essay where the TAs reached consensus was selected as a calibration essay, whereas the other essays were discarded.

One extremely valuable (and unintended) benefit of the process of creating calibration essays was calibrating the TAs themselves. With the exception of the lead TA, our course is run by a new contingent of TAs every year, most or all of whom have no particular past experience in evaluating essays. The meetings and discussions to determine marks for the calibration essays constituted an opportunity to give the TAs extensive extra training.

A one-time benefit of the initial process of creating calibration essays was that it pointed out opportunities to improve our rubric. The rubric went through multiple iterations during the process of calibrating the TAs, as they discovered various ambiguities.

## 3.2 Independent Reviewers

One bottleneck in our original Mechanical TA design was that all students begin in the supervised pool, requiring extensive TA work at the beginning of term. One of our main motivations for introducing an automated calibration process to reduce this TA workload by encouraging students to be promoted to the independent pool before the first assignment was marked. We were unsuccessful in achieving this goal in our 2013 course offering: no students were promoted to independent before the first assignment, and hence TAs needed to mark every student's essay.[6] However, the opportunity to practice reviewing that the calibration essays provided appears to have had a large effect on students' review skills. More students were promoted to independent early in the 2013 offering than in either of the earlier offerings, and a larger overall proportion of the class (100%!) became independent during the term.

Figure 1 shows how many students reviewed independently over the course of the term in each of our past three offerings. The criteria for becoming independent in 2011 were much more lenient than in 2012, leading to a large number of students becoming independent fairly quickly. However, this lenience seems to have resulted in the promotion of many unreliable reviewers, and so many of these students were later moved back to the supervised pool as a result of spot checks and appeals. In contrast, all but one of the many students who became independent in 2013 stayed in the independent pool throughout the class. Nearly a third of the students became independent after just one assignment; by the end of the course, *every* student was reviewing independently. The criteria for becoming independent based on review quality were identical in 2012 and 2013; the only differences between the two years were the introduction of our calibration system and the improvements we made to the review rubric to support calibration.

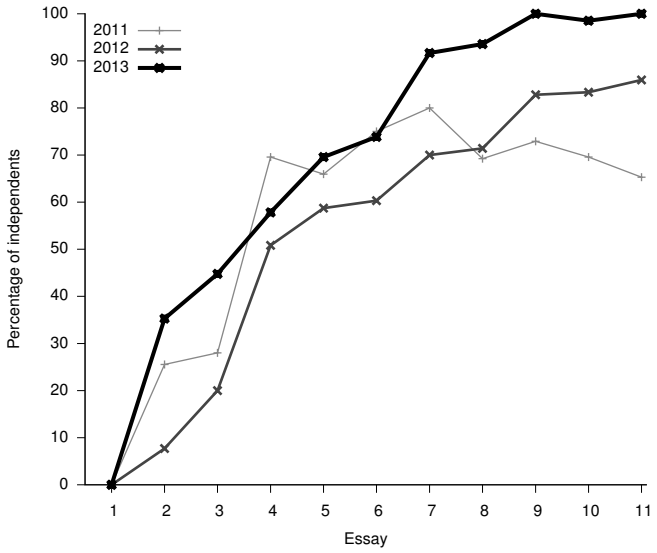


**Figure 1: Proportion of independent reviewers at the beginning of each assignment.**

## 3.3 Exam Performance

It would be nice to compare assignment marks between years; however, this is difficult because we made dramatic changes to the rubric. In 2011 and 2012, we marked essays out of 8, and an "acceptable" essay received 8/8. In 2013, we marked essays out of 20, and gave an "acceptable" essay 16/20. Thus, we do not present an analysis of how assignment marks varied from one year to the next.

In contrast to assignments, we marked essays on the midterm and final exams in a very similar way across all three offerings, and indeed offered very similar exams. This makes exams a more suitable target for analysis. Figure 2 gives the cumulative distributions of marks on the midterm and final exams across the three years. We observe that the mark distributions for both exams were strikingly higher in 2013 than in the prior years.[7] We thus conclude that one or both of the improvements associated with our calibration system had a positive impact on student performance.

[6] We've since tweaked our calibration threshold and the number of calibration essays required of students, and believe that this will yield a vastly different outcome in our current, 2014 offering.

[7] A Mann-Whitney rank test confirms this. Both the midterm and final exam distributions for 2013 are significantly higher than the corresponding distribution for both 2011 and 2012 ($p < 0.001$).

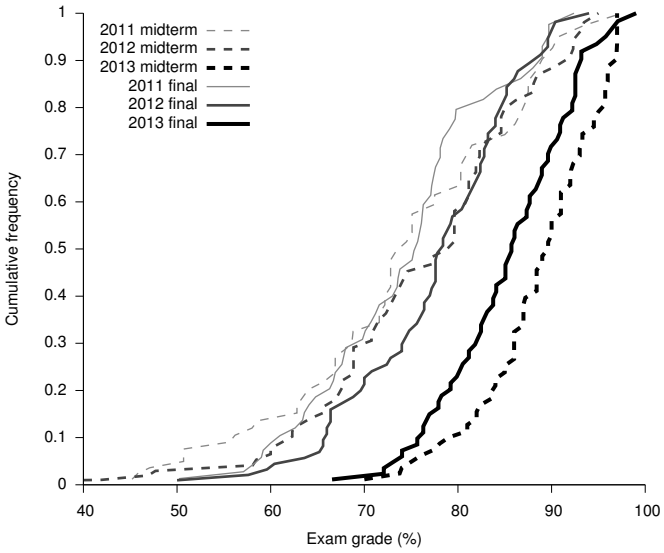


**Figure 2: Cumulative distributions of final exam and midterm exam marks.**

# 4. ANALYSIS OF OUR CURRENT DESIGN

We now turn to a deeper analysis of data from the latest offering of Computers and Society. We first confirm that the division of supervised and independent reviewers meaningfully reflects differences in review quality. We then consider the effect of reviewing practice on calibration performance.

## 4.1 Review Quality

Mechanical TA is designed on the premise that independent reviewers can be trusted to reliably review peer work without oversight, whereas supervised reviewers cannot. An important question is therefore whether the two pools really do differ in terms of review quality. To answer this question, we followed the basic strategy of estimating the average review quality of supervised reviews and the average quality of independent reviews, and checking whether these averages differed significantly. The quality of supervised reviews is easy to estimate, since all of them get marked by a TA. For independent reviews, we had access to TA marks of reviews that were randomly spot checked or appealed, unfortunately without a label indicating which criterion had led to their selection. The spot check selection criterion adds a complication, however: all essays that receive a grade of 80% or higher get spot checked automatically; all other essays are spot checked at random. If the quality of an essay is independent of the quality of its reviews, then this does no harm. However, if high-quality essays are easier to grade, then this selection criterion could add an upward bias to the estimate of the independent reviews' quality, since our sample of independent reviews would contain disproportionately many easily graded essays. We address this by subdividing the independent and supervised reviews into those that were associated with essays that got a mark over 80% and those that did not, giving a total of four groups of observations. This allows us to detect the situation where the supervised reviews have significantly different quality from the independent reviews of high-mark essays, but not significantly different quality from the independent reviews as a whole. Another possible source of bias is appeals, as low-quality reviews may be appealed more frequently than average. We do not attempt to correct for this bias, for two reasons. First, the bias is downward for independent reviews; if we find a statistical difference between the two pools in the presence of this bias, correcting for it will not change our finding. Second, we cannot distinguish retroactive spot checks that were triggered by random spot checks from those that were triggered by appeals.

For each of our four groups of reviews, we estimated a Bayesian joint posterior distribution over the following model:

$$\begin{aligned} \mu_g &\sim \text{Uniform}[0, 10] \\ \sigma_g &\sim \text{Uniform}[0.0001, 10] \\ q_{g,r} &\sim \text{N}(\mu_g, \sigma_g) \quad \text{truncated to } [0, 1], \end{aligned}$$

where $q_{g,r}$ is the quality of review $r$ in group $g$. We normalized all marks to lie within $[0, 1]$. The quality of each review in group $g$ is assumed to be drawn from the same Normal distribution, truncated at 0 and 1. We estimated the posterior distributions over the parameters $\mu_g, \sigma_g$ for each group using a Metropolis-Hastings sampler [10] to simulate $12,000$ samples after a burn-in period of 4000 samples.[8]

Figure 3 gives the cumulative posterior distribution over the average review quality for each group.[9] The 95% central credible interval for each distribution is shown as a bar on the $x$-axis.[10]

We observe that there is no overlap between the credible intervals of either independent group with either supervised group. This answers our question: we have strong evidence that independent reviewers perform higher quality reviews.

We are also able to examine whether high-quality essays are easier to review well. In both the independent and supervised groups, the quality of the reviews of essays that received grades of at least 80% did indeed appear to be higher, although not substantially (nor significantly; note that the credible intervals for the above- and below-80% groups intersect). The effect was more pronounced in the supervised group than in the independent group, although again not statistically significant.

## 4.2 Calibration Performance

We have described two benefits offered by calibration: assessing students' review quality without TA intervention, and providing an opportunity for students to practice reviewing with immediate feedback. In this section, we evaluate whether students benefit from such practice by asking whether students' calibration marks improved as they completed more calibration reviews.

We begin by plotting the performance of each calibration that was completed. We index calibrations by the time of promotion to the independent pool; that is, the last calibration review performed before a student was promoted is calibration number 0, the calibration review completed just before that is calibration number -1, etc. We then perform a Bayesian linear regression by estimating the joint posterior

[8]We used version 2.3 of the PyMC package to implement the sampler; see `http://pymc-devs.github.io/pymc/`, retrieved September 4, 2014.

[9]Due to the truncation of the Gaussian distributions to the interval $[0, 1]$, this is not identical to the posterior distribution of the $\mu_g$ parameter.

[10]A central credible interval is a Bayesian counterpart to a confidence interval. The true value of a parameter lies within its 95% central credible interval with probability 0.95.

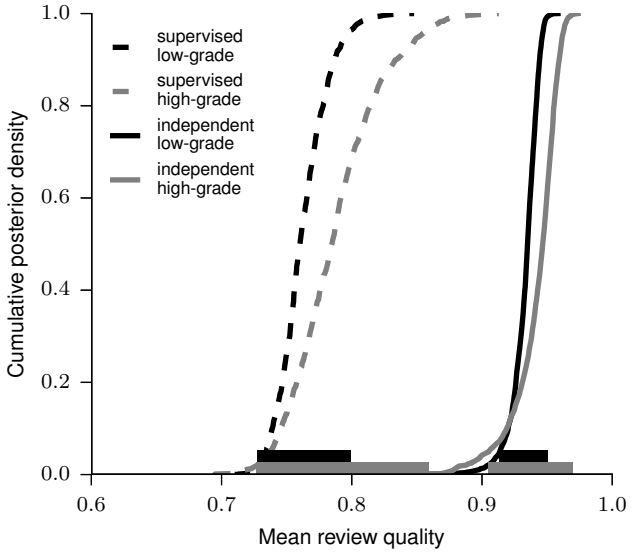


**Figure 3: Cumulative posterior distributions for the mean review quality of supervised reviews of low-marked essays, supervised reviews of high-marked essays, independent reviews of low-marked essays, and independent reviews of high-marked essays. 95% central credible intervals for each of the distributions are shown as bars on the $x$-axis.**

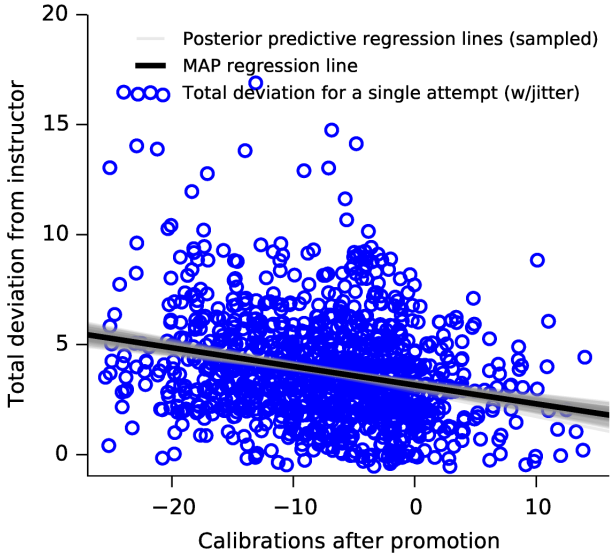


**Figure 4: Total deviations from gold standard review on calibration reviews, versus number of calibrations after promotion. The bold line is the maximum a posteriori linear fit with Normal error. The gray lines are samples from the posterior predictive distribution of linear fits.**

distribution of the following model:

$$b \sim \mathrm{N}(0, 5) \qquad \sigma \sim \mathrm{Uniform}[0, 10]$$
$$m \sim \mathrm{N}(0, 2) \qquad y_i \sim \mathrm{N}(mx_i + b, \sigma),$$

where $m$ and $b$ are the slope of the regression line, $x_i$ and $y_i$ are the number and performance for each calibration review, and each datapoint $(x_i, y_i)$ has zero-mean Gaussian noise with variance $\sigma$. Performance is measured as sum of absolute differences (i.e., the $L_1$ distance) from the instructor review; smaller performance values thus represent better performance. We again used Metropolis-Hastings sampling to estimate the posterior.

Figure 4 shows a plot of the number and performance for each calibration (with a small amount of jitter). The maximum a posteriori regression line is plotted as a bold line; this is the line whose slope and offset have the highest posterior probability. To illustrate the range of possible fits, we also plot the lines corresponding to 100 samples from the posterior distribution.

The MAP estimate of the slope is $-0.085$, with a 95% central credible interval of $[-0.104, -0.066]$. The credible interval does not contain 0, so we conclude that students showed a significant improvement in their calibration performance as they practiced. Our rubric grades essays out of 20, so a slope of $-0.085$ represents an average improvement of approximately 4% with each calibration.

## 5. RELATED WORK

Now that we have described Mechanical TA in detail, we give a more thorough survey of related work and describe how our own system differs. By far the most widely used online peer review system is Calibrated Peer Review (CPR) [1, 11]. After submitting their own essays, students evaluate three instructor-provided calibration essays of varying known quality. They then anonymously review the essays of other students. Each review is weighted according to the reviewer's performance on the calibration task. Reviewers who do not pass the calibration task on the first two tries "flunk out" of the assignment and are not permitted to review at all. Review quality is further assessed by students' reviewing other students' reviews. The initial calibration essays are entirely for the purpose of evaluating students' reviewing skill, and form a portion of the students' grade. This contrasts with our calibration essays, which do not directly impact a student's grade, and which allow students to practice reviewing in addition to demonstrating reviewing competence.

Kulkarni et AL. [6] combine algorithmic assessment of written answers with peer review in a large online course. A learning algorithm first estimates both the assessment and its confidence in the assessment. These estimates are used to determine how many peer reviews are required for a given item. Other students then assess the peer reviews' accuracy.

Mechanical TA focuses on evaluating the final version of an assignment. SWoRD [2], PRAZE [7], and CaptainTeach [9] allow students to incorporate feedback from peer reviews during the course of an assignment.

CrowdGrader [3] dynamically assigns reviews to reviewers in an online fashion, in an attempt to provide an approximately equal number of reviews to each submission. Similarly to CPR, the quality of each review is assessed by comparing it to the "consensus" (trimmed average) review of the assignment; reviews that are further from the consensus are penalized. The Aropa system [5] combines consistency scoring and weighting by reweighting reviews until a fixed point of weights and consistency with the weighted average is reached. Both systems thus assess review quality "auto-

matically", whereas Mechanical TA assesses review quality directly via TAs. Many other systems use a "review the reviewer" system to evaluate review quality, in which students rate the quality of the reviews they have received [4, 2, 7, 8].

# 6. CONCLUSIONS

Mechanical TA is a system designed to support a novel model of high-stakes peer grading, in which marks from trusted "independent" reviewers are binding (but can be appealed), but marks from untrusted "supervised" reviewers are replaced by grades from a TA. Students are promoted to independent status based on the quality of their reviews, and after promotion they typically remain independent for the duration of the term. We have successfully used this system to set weekly essay assignments in a class of approximately 70 students. This would not be possible if every assignment had to be graded by a TA, as essays are very time consuming to grade. We have focused here on grading essays, but our system is easily applicable to other domains such as coding assignments or code review.

A major bottleneck in our peer review approach is that the first assignment *does* require that TAs mark every submission along with all of the peer reviews. While we have found that TAs are willing to work hard at the beginning of term given assurances that they will subsequently have a much-reduced workload, this bottleneck nevertheless limits the scalability of our system. We thus introduced calibration reviews in the most recent offering, in which students review carefully chosen assignments with known "correct" gold standard reviews constructed by the instructor and TAs. Each student receives automated feedback comparing their review to the gold standard review, and if they match the gold standard closely enough on enough repetitions, they are automatically promoted to independent status. This calibration mechanism has multiple goals. First, it aims to allow students to become independent before the first assignment, without TA intervention, thereby reducing TA workload on the first assignment. Second, it allows students to practice the reviewing process, with immediate feedback about how well they did. We did not achieve the first goal in our most recent course offering. Nevertheless, offering students practice reviewing had a striking effect. Students in the 2013 offering were promoted sooner and received higher grades on roughly comparable exams than those in the 2011 and 2012 offerings. Students' average review performance improved by approximately 4% per attempted calibration essay.

One additional benefit of a calibration system is that it allows the systematic training of TAs in how to mark according to "subjective" rubrics. (We described how our TAs benefited from constructing calibration questions; we've asked our 2014 TAs to do the existing calibration exercises before the class starts.) We believe that this leads to higher quality marking by TAs and more consistency between TAs.

Calibrating reviewers before the first assignment is a key requirement for increasing the scalability of Mechanical TA's peer review model. In the next offering of the class, we will modify the system in several ways. First, we will modify the calibration promotion threshold, choosing a more appropriate value based on data from the latest offering. Second, we plan to experiment with including calibration essays in independent students' assigned reviews, without indicating which essays are from students in the course, and which are calibration essays; this will enable us to monitor the quality of independent reviews more closely. In particular, it will enable us to validate the calibration system by comparing the review quality of students who are promoted by automatic calibration to the review quality of students who are promoted by TA evaluations.

## Acknowledgements

We thank Jessica Dawson for pointers to related literature.

# 7. REFERENCES

[1] O. L. Chapman. Calibrated Peer Review™, 2001.

[2] K. Cho and C. D. Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3):409–426, 2007.

[3] L. de Alfaro and M. Shavlovsky. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, pages 415–420, New York, NY, USA, 2014. ACM.

[4] E. F. Gehringer. Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin*, 33(1):139–143, 2001.

[5] J. Hamer, K. T. K. Ma, and H. H. F. Kwong. A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian Conference on Computing Education - Volume 42*, ACE '05, pages 67–72, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc.

[6] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 99–108, New York, NY, USA, 2014. ACM.

[7] R. A. Mulder and J. M. Pearce. PRAZE: Innovating teaching through online peer review. In *ICT: Providing choices for learners and learning. Proceedings ascilite Singapore*, 2007.

[8] D. E. Paré and S. Joordens. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning*, 24(6):526–540, 2008.

[9] J. G. Politz, D. Patterson, S. Krishnamurthi, and K. Fisler. CaptainTeach: Multi-stage, in-flow peer review for programming assignments. In *ACM SIGCSE Conference on Innovation and Technology in Computer Science Education*, 2014.

[10] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[11] R. Robinson. Calibrated Peer Review™ an application to increase student reading & writing skills. *The American Biology Teacher*, 63(7):pp. 474–476+478–480, 2001.

[12] R. Seabrook, G. D. Brown, and J. Solity. Distributed and massed practice: from laboratory to classroom. *Applied Cognitive Psychology*, 19(1):107–122, 2005.

[13] M. E. Walvoord, M. H. Hoefnagels, D. D. Gaffin, M. M. Chumchal, and D. A. Long. An analysis of Calibrated Peer Review (CPR) in a science lecture classroom. *Journal of College Science Teaching*, 37(4):66, 2008.