Analysis of Grading Times of Short Answer Questions

Michael Yen
Turnitin
Oakland, CA, USA
myen@turnitin.com

Sergey Karayev
Turnitin
Oakland, CA, USA
skarayev@turnitin.com

Eric Wang
Turnitin
Oakland, CA, USA
ewang@turnitin.com

ABSTRACT

We present an analysis of factors correlated to grading speed in short answer questions from college level STEM courses using a novel dataset collected by an online education company. By analyzing timestamp data, we were able to estimate how long instructors grade individual student responses, which we typically found to be less than 10 seconds. This dataset provides us with a unique opportunity to determine which steps in the grading workflow could benefit from intervention. We found that sorting responses by rubric similarity has the potential to drastically reduce grading time by up to 50% per response. We plan to follow this work by implementing an intelligent agent to present responses in a sorted order to minimize grading time.

Author Keywords

learning at scale; short answer questions; grading speed; grading time; examinations;

CCS Concepts

•**Applied computing → Computer-assisted instruction;**

INTRODUCTION

One of the challenges instructors face when teaching a large class is grading free response questions. Free response questions are a necessary assessment tool since they require students to recall rather than recognize information, therefore probing knowledge at a higher level in Bloom's taxonomy [2]. As a result, students express their answers in a large number of unique ways which forces the instructor to individually grade each response.

Solutions to this problem have taken the form of peer assessment [5], semi-automated grading [1], fully automated grading [6], as well as streamlined user interfaces [1, 4] across a wide variety of curricula. Many of these solutions utilize either supervised or unsupervised machine learning in order to scale the instructor's effort to be sub-linear with the number of students. In many of these solutions, the instructor is not required to inspect each student's response. This leads to concerns about accuracy and security, since students may learn to "game" the system. In contrast, we seek to identify interventions that reduce the time it takes to grade an individual response, mitigating such concerns.

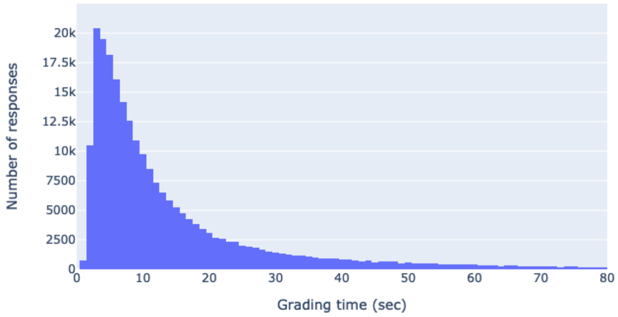


Figure 1: *How much time do instructors spend grading each response?* 50% of responses are graded in less than 9 seconds and 90% of responses are graded in less than 41 seconds.

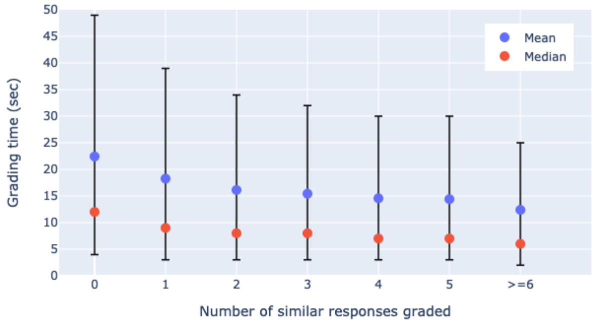
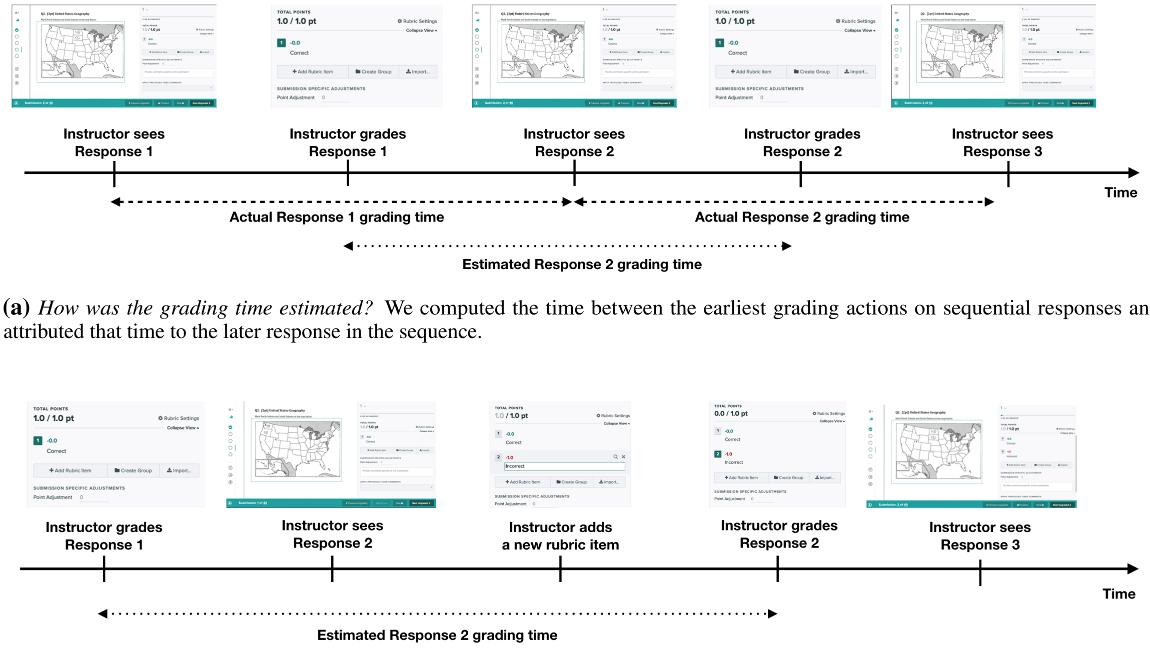


Figure 2: *Does grading a sequence of similar responses speed up grading?* Instructors graded faster when encountering a sequence of similar responses. On the extreme left side of the plot, the median grading time for a response that was different from the previous response is 12 seconds. This is twice the length of the median grading time for a response that was graded after seeing five responses similar to it — 6 seconds, on the extreme right side of the plot. The bars represent the 10-90 percentiles. Pearson $R = -0.10$.

Our contribution is an in depth correlation analysis of the instructor's grading process, based on a large dataset of graded questions from a number of subjects and institutions, to identify the operations that can most benefit from full or partial automation. We observed that instructors speed up as they grade sequences of similar responses. This observation points to a future intervention where responses are pre-sorted by similarity before the instructor begins grading, which we hypothesize can drastically reduce grading time.

L@S '20, August 12–14, 2020, Virtual Event, USA.

ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3406748


Instructor sees Response 1
Instructor grades Response 1
Instructor sees Response 2
Instructor grades Response 2
Instructor sees Response 3
Time
Actual Response 1 grading time
Actual Response 2 grading time
Estimated Response 2 grading time
Instructor grades Response 1
Instructor sees Response 2
Instructor adds a new rubric item
Instructor grades Response 2
Instructor sees Response 3
Time
Estimated Response 2 grading time

(a) *How was the grading time estimated?* We computed the time between the earliest grading actions on sequential responses and attributed that time to the later response in the sequence.

(b) *How did we count the number of rubric items created?* For consistency between our measurements, we counted the number of rubric items created within the estimated grading time window for each response.

Figure 3: Illustration of how we estimated different measurements from a sequence of timestamps.

DATA COLLECTION

Gradescope is an educational company that provides tools to facilitate the online grading of exams and homework. Since its inception Gradescope has been used by over 10,000 instructors to grade over 1,000,000 students. The majority of instructors using Gradescope are located at tier 1 universities in the United States and teach STEM courses.

In the typical workflow, instructors create assignments as a PDF and students write their answers on printed out copies of the assignment which are later scanned and uploaded to Gradescope. Once student submissions are uploaded, instructors can create a rubric and grade each response to each question, grading one question at a time. We note that instructors have the ability to skip and revisit responses while they are grading, however we found that the vast majority of instructors grade in the originally presented order. For more detail on the interface and workflow see [4].

While grading, the instructor must complete four tasks: (1) read the hand written student response, (2) understand what has been written, (3) if needed, add to or edit the rubric, and (4) apply the rubric. Steps 3 and 4 are logged by Gradescope as distinct grading actions which we used to construct our dataset.

The timestamps associated with each action allow us to estimate the time spent grading each student response and count the number of rubric items created while grading each response. We illustrate how these estimates are made in Figure 3 and show the distribution of estimated grading times in Figure 1.

DATA

The data presented here is a subset of the data collected via the Gradescope platform between August 2013 and September 2018. The majority of instructors who use Gradescope teach STEM courses, so our data is biased towards STEM instruction. The data was randomly drawn from a database of exam and quiz questions with at least 100 student responses. We further refined our dataset to only "short answer" questions, meaning that the typical student response is limited to a few words to a paragraph of hand written text. This refinement was done by a team of annotators who classified the type each question. See [3] for details about the annotation effort.

We decided to focus on short answer questions because they present a challenge for grading in large scale courses and are the most frequent question typed used by instructors on Gradescope. This was a necessary step to control for the differences across response formats.

Our final dataset, after removing outliers, contained 242,775 graded student responses from 1,358 questions, 338 courses, and 43 institutions.

Outlier Identification

Due to the way we were estimating grading time and how instructors use Gradescope, we needed to identify and remove responses with incorrectly estimated grading times from our data. A common scenario is when an instructor grades half of the responses, takes a break, and then finishes grading. This break would manifest itself in our data as a single response with a very large grading time.

	Grading Time (sec)		
Credit	*Mean*	*Median*	*N*
Correct	16.07	8	126,157
Incorrect	15.45	8	46,824
Partial	25.05	14	56,947

Table 1: *Do instructors grade correct responses faster than incorrect or partially correct responses?* On average instructors take 56% more time to grade a partially correct response than a correct or incorrect response.

Questions in our data spanned a wide range of median grading times from 2 seconds to 226 seconds, which reflects the complexity of the question and the rubric used to grade it, and the behavior of different graders. Therefore, we used a dynamic definition of outliers. Within each question, we removed all responses with grading time greater than 10 times the median grading time for that question. This definition allows us to keep responses that naturally take a longer time to grade, which is not possible with a static definition of outliers, such as removing all responses that take longer than 90 seconds to grade. Following this procedure we removed $12,537$ student responses.

In addition to long grading times, some questions in our dataset were graded by more than one instructor. For consistency within individual questions, we only kept responses graded by the instructor who graded the most responses within each question. Any other responses were removed. This step removed $48,773$ student responses.

ANALYSIS

Our timestamped data provides us with a unique opportunity to analyze how long it takes instructors to grade. Note that even if a factor has a small effect per-submission, the effect can be large over the grading session, as hundreds of submissions have to be graded. In particular, we identify a good candidate for intervention, which we discuss last.

Correctness

We labeled each response with one of three labels, correct, incorrect, or partially correct. Correct responses received a normalized score of 100%, an incorrect response received a score of 0%, and partially correct as any other response. As seen in Table 1, instructors are about two times faster when grading correct or incorrect responses compared to partially correct responses.

Number of Responses Graded

We hypothesized that instructors would speed up as they graded more responses. We aggregated the grading times across all questions and show the mean and median grading times in Figure 4. We can see that the grading time per response rapidly decays with the number of students graded. This result is consistent with Singh et al. 2017 [4].

Rubrics

Rubrics are a critical part of the grading process. In Gradescope, rubrics are comprised of individual rubric items that can be turned on or off for each response, thereby assigning a corresponding amount of credit. Instructors can create, edit, or delete any rubric item at any time. Additionally, rubrics can be composed of any number of rubric items which allows for instructors to grade with any level of granularity.

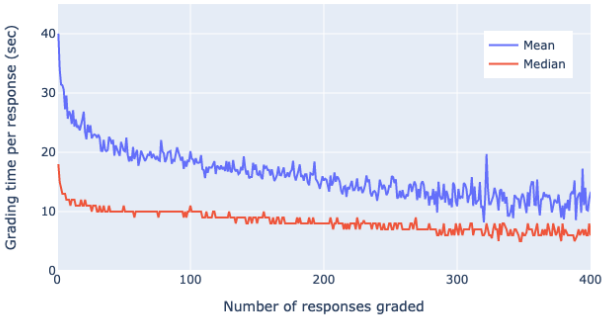


Figure 4: *How does time spent grading a response change with the total number of responses graded?* Mean and median grading time per response quickly decreases with the number of responses already graded.

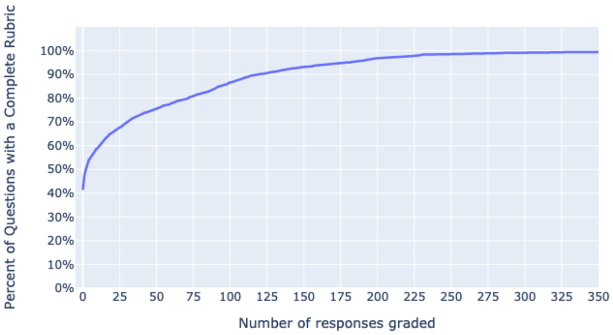


Figure 5: *How quickly do instructors complete their rubric while grading?* The number of questions with a completed rubric rapidly grows with the number of responses graded. It is not common to begin grading with a completed rubric.

Rubric Item Creation

On Gradescope, the majority of instructors begin grading with an incomplete rubric, and add rubric items to it as they grade. As shown in Figure 5, only 41% of questions have a completed rubric before grading, the other 59% have their rubric completed at some point in the grading process. From Figure 5 we can see that rubric item creation is heavily concentrated at the beginning of the grading process.

In total, 1.4% of all student responses have at least one rubric item created while being graded. We observe in Figure 6 that the estimated grading times for these responses are greater than for responses where no new rubric items are created during grading. As would be expected, there is a positive correlation, Pearson $R = 0.12$, between the number of rubric items created while grading a response and the grading time for that response.

Rubric Size

The number of rubric items used to grade a question can be indicative of many things, including question complexity, the relative importance of the question within the assignment, the typical student response length, and the instructor's willingness to provide detailed feedback. While we do not believe

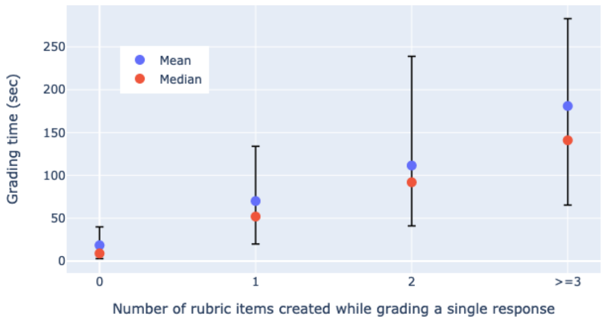


Figure 6: *How does creating rubric items while grading a response affect grading time?* We observe that time spent on grading a response increases with number of rubric items created while grading the response. We show both mean and median times, with the bar representing 10-90 percentiles.

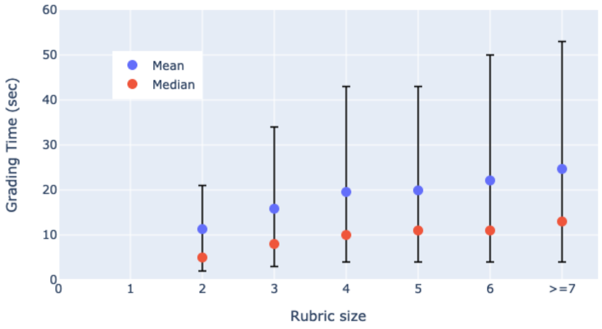


Figure 7: *Do instructors take longer to grade questions with a large rubric than a small rubric?* Looking only at responses that were graded after the rubric was completed, we find a slight positive correlation, Pearson $R = 0.11$, between grading time and number of rubric items used to grade. A question with 4 rubric items tends to take almost twice as long as a question with 2 items to grade. We show both mean and median, with bars representing 10-90 percentiles.

that rubric size is a good candidate for intervention, we are still interested to see how it is correlated to grading time.

As seen in Figure 7, we observe that the total number of items in a completed rubric, is positively correlated, Pearson $R = 0.11$, with the time it takes to grade a question. Questions graded with a smaller rubric tend to be graded faster than questions with a large rubric. We are not implying a causal relationship between the rubric size and grading time here, but simply noting the correlation.

Response Similarity

We wanted to see whether grading two similar responses in sequence tends to be faster than grading two dissimilar responses. For this analysis, we considered two responses to be similar if they had identically marked rubrics, which is a definition that can certainly be improved, as we discuss in the last section.

We refer to each sequence of thus defined similar responses as a *chain*, and each response can be referred to by its position in its chain. For example, position 2 represents a response with two similar responses preceding it, and position 0 represents a response that is different from the response preceding it. After removing all chains with ≥ 1 rubric item created, we found a total of $36,042$ chains. The relationship between position in the chain and grading time is shown in Figure 2. We observe a difference of 50%, 12 seconds to 6 seconds, between the medians of position 0 and position ≥ 6.

FUTURE WORK

Finding a way to scale instructor feedback to students is critical to maintaining high quality education for ever increasing class sizes. Our main finding is that instructors speed up as they grade sequences of similar responses. This points to a promising intervention of sorting responses by similarity in order to present the instructor with maximally long sequences of similar responses.

The definition of similarity used in our analysis also stands to be improved. We identified two specific weaknesses. First, rubric similarity can only be computed after grading. Second, depending on the rubric, responses that are very different in terms of length, language used, and other attributes can often be marked with the exact same set of rubric items, which makes them similar by our definition.

Due to these weaknesses, any future intervention will require the use of a different similarity metric, specifically one that takes into account the content of each response. Ideally, the similarity metric definition will additionally be influenced by the grading behavior of the instructor. This is the subject of our current work.

REFERENCES

[1] Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and Correct: Using Clusters to Grade Short Answers at Scale. In *ACM Conference on Learning @ Scale (L@S '14)*. 89–98. DOI: `http://dx.doi.org/10.1145/2556325.2566243`

[2] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. `https://doi.org/10.1207/s15430421tip4104_2`

[3] Paul Laskowski, Sergey Karayev, and Marti Hearst. 2018. How do professors format exams?: an analysis of question variety at scale. In *ACM Conference on Learning @ Scale*. DOI: `http://dx.doi.org/10.1145/3231644.3231667`

[4] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. 2017. Gradescope: A Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work. In *ACM Conference on Learning @ Scale*. `https://doi.org/10.1145/3051457.3051466`

[5] Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the Peer Assessment Experience on MOOC Platforms. In *ACM Conference on Learning @ Scale*. `https://doi.org/10.1145/2876034.2876043`

[6] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Association for Computational Linguistics: Human Language Technologies*.