Using Transformer Language Models to Validate Peer-Assigned Essay Scores in Massive Open Online Courses (MOOCs)

Wesley Morris
wesley.g.morris@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Scott A. Crossley
scott.crossley@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Langdon Holmes
langdon.holmes@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Anne Trumbore
TrumboreA@darden.virginia.edu
University of Virginia
USA

ABSTRACT

Massive Open Online Courses (MOOCs) such as those offered by Coursera are popular ways for adults to gain important skills, advance their careers, and pursue their interests. Within these courses, students are often required to compose, submit, and peer review written essays, providing a valuable pedagogical experience for the student and a wealth of natural language data for the educational researcher. However, the scores provided by peers do not always reflect the actual quality of the text, generating questions about the reliability and validity of the scores. This study evaluates methods to increase the reliability of MOOC peer-review ratings through a series of validation tests on peer-reviewed essays. Reliability of reviewers was based on correlations between text length and essay quality. Raters were pruned based on score variance and the lexical diversity observed in their comments to create sub-sets of raters. Each subset was then used as training data to finetune distilBERT large language models to automatically score essay quality as a measure of validation. The accuracy of each language model for each subset was evaluated. We find that training language models on data subsets produced by more reliable raters based on a combination of score variance and lexical diversity produce more accurate essay scoring models. The approach developed in this study should allow for enhanced reliability of peer-reviewed scoring in MOOCS affording greater credibility within the systems.

CCS CONCEPTS

• **General and reference** → **Validation**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → **E-learning**; **Collaborative learning**.

KEYWORDS

transformers, natural language processing, rater reliability, moocs

ACM Reference Format:
Wesley Morris, Scott A. Crossley, Langdon Holmes, and Anne Trumbore. 2023. Using Transformer Language Models to Validate Peer-Assigned Essay Scores in Massive Open Online Courses (MOOCs). In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3576050.3576098

1 INTRODUCTION

Ever since their introduction in 2008, Massive Open Online Courses (MOOCs) have provided opportunities for skill development and recognition for students that may not be able to attend traditional education courses [30]. The versatility of MOOCS has led to their quick growth and, in 2016, the popular MOOC hosting site Coursera had over 17.5 million registered users [39]. Although concerns have been raised regarding the retention of students in the courses [34], students continue to perceive participation in MOOCs as a way to develop their cognitive interests, career goals, and interpersonal relationships [15, 20].

Assessment results generated by these courses provide a wealth of data for researchers, teachers, and administrators [8, 12, 36]. However, much of this behavioral data is based on click-stream logs, and data about learning is generally based on closed assessment such as multiple-choice items. The use of open responses such as essays are difficult to manage in MOOCs because the sheer number of students makes personalized teacher feedback difficult. One solution to incorporating open ended assessments in MOOCs has been for students in the course to review samples written by the other students and assign scores to those samples based on holistic or analytic rubrics [1]. Unfortunately, research indicates that these peer-assigned scores may have serious problems with reliability and validity [19, 26, 43].

This paper seeks to address these problems by providing a method to increase the reliability of peer-assigned scores. We examine a corpus of 27,909 essays produced as a capstone project to a MOOC on design principles hosted by Coursera. We generate subsets of the essays by pruning raters suspected of providing unreliable scores based on the reviewer score variance and the lexical diversity of their comments. We use the correlations between score and word count as a measure of criterion validity knowing that longer essays receive higher scores [13, 32]. We then develop large language models (LLMs) for each subset to predict the essay scores from the pruned subset of raters and tested the accuracy of the models by

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576098

subset. Our assumption was that LLMs trained on higher-quality data would produce higher scoring accuracy, leading to a measure of score validity for the generated subsets.

1.1 Peer-reviewed Assignments

Within a massive open online course, closed response items such as multiple choice and true or false problems have the advantage of being easily scorable by computer [39] and as a result are often the assessment types most relied upon by MOOCs [40]. Although they are often seen as more practical to implement, closed response items do not elicit ideas or knowledge from the students [5] and research indicates essay questions measure different competencies from multiple choice items [31]. Additionally, closed response items do not accurately reflect the types of tasks that the learners may be expected to face in the workplace [28]. As a result, most assessment experts suggest that learners produce open-ended responses like essays in order to track and understand learning [11, 44].

Open response assessments, though, can pose a difficulty in MOOCS because the high enrollment of these courses makes it impossible for a single instructor or even a small group of instructors to score written samples in an efficient manner [42]. One solution has been to have these extended response items peer-reviewed by other students in the course [18]. This solution not only provides valuable feedback to the author of the item, but also gives the reviewers the opportunity to hone their own skills by critically analyzing the work of other students [9, 41]. The very practice of reviewing another students' work may engage cognitive processes that can enhance a deeper learning experience [19].

However, it can be very difficult to obtain reliable scores using peer-reviews because peer reviewers may have little incentive to take the time or effort to deeply analyze the work that they review and some may give scores that have little to no relationship with the text they are reviewing [19, 43]. Additionally, the reviewers rarely have the standardized, rigorous training that is expected of reviewers in scientific studies which may increase rater reliability [1, 18]. While some platforms include ways to assess and train raters by comparing their scores against benchmark essays [2], some researchers have called into question the entire project of using peer-assigned scores as a valid measure of assessing student writing because of the very low observed reliability [19].

1.2 Automatic Essay Scoring

One solution to assessing open ended writing samples has been to remove humans from the loop entirely and rely on automated essay scoring (AES) software [4, 35]. AES systems are algorithms that automatically provide summative feedback to users about the quality of written samples. The simplest AES would predict essay quality using text length because longer essay generally receive higher scores. In fact, previous research has indicated strong correlations between essay length and score, with correlations between 0.42 and 0.79 common [32]. More advanced AES systems rely on using linguistic features automatically produced by NLP tools that go beyond text length. These features include lexical diversity [24], syntactic complexity [23, 27], cohesion features [14], lexical sophistication [22], or argument tree depth [25]. Newer AES systems use neural network based Transformer Large Language Models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT) [16] to provide feedback [7]. Large language models that have been finetuned using labelled data (e.g., scored essays) in a target domain have been found to be strongly correlated with essay scores produced by expert raters [37, 45, 46]. As a result, LLM AES system may be a potential solution to assessing open-ended assessments in MOOCs [25]. However, the large amount of compute time needed to run these models may make them impractical to implement [29]. Just as significantly, a transformer model is only as good as its training data. In order to have predictive power, these models should be finetuned using data from the target language domain which have labelled scores that reliably reflect the quality of the essay [33]. In contexts where the reliability of the student-assigned scores are questionable, such as in the case of MOOC peer-review data, models trained on that data will also be in question.

1.3 Current Study

This study addresses the problem of peer-rater scores by assessing methods to prune raters thought to be unreliable and then validating the scores provided by the remaining raters by building large language models (i.e., AES systems) to predict the scores. We develop ten subsets of peer-review data from a dataset of 27,909 students in a massive open online course (MOOC) on design principles hosted by the educational company Coursera. We prune reviewers that are unreliable based on their score variance and the lexical diversity of comments that they left within the MOOC. We assess the strength of the pruned reviewer subsets based on correlations with essay length, which is a strong predictor of writing quality. We then build ten large language models to predict the essay scores from each subset. Model accuracy is then compared to evaluate whether more reliable subsets of data produce better performing models. This study answers the following two questions:

RQ 1: Can the reliability of peer-reviewed rating be increased by pruning reviewers based on variance in scores and the lexical variety of written feedback?

RQ 2: Can LLMs be used to develop automatic essay scoring systems that provide convergent validation for the pruning processes found in research question 1.

2 METHODS

2.1 Participants

This study uses data provided by 27,909 students in a massive open online course (MOOC) on design principles hosted by the educational website Coursera. Users were invited to either audit the course for free or pay to receive a certificate of completion at the end. Within the MOOC, students completed a four-week course on design thinking with a graded quiz at the end of each week. Additionally, they completed a capstone final essay which was graded by their peers. Students had to pass all graded assignments in order to complete the course. In addition, students were invited to use discussion forums hosted within the Coursea website to connect with peers.

Students in the course were given the opportunity to fill out a demographic survey which collected information about the students' gender, year and country of birth, country of residence, race and ethnicity, education level, current educational enrollment status, industry, and their proficiency in English and other languages. However, very few students completed the demographic survey (n=198) making it unusable. Location stamps collected from clickstream data were used to determine the geographic location of the students. The largest group of users logged in from India, making up 44% (n=6,931) of the total users in the baseline data set. The next largest groups logged in from the United States, Mexico, and Brazil, making up 14.8% (n=2,337), 5% (n=795), and 4.7% (n=734) of the total number of users respectively. The rest of the users came from 153 different countries.

2.2 Capstone Project

The most important component of the MOOC was a written student reflection on applying one of four design tools they learned about in the course – visualization, storytelling, mind mapping, or learning launch – to a challenge of their choice. Students were asked to describe the challenge, explain why they selected the particular tool, describe how they applied the tool to their challenge, describe any insight they gained from the assignment, and describe how they might change their approach in the future. Students were told that each of these elements would receive an individual score based on peer-reviews. The original corpus consisted of 27,909 reflection essays, each written by a separate, unique author. However, 5,006 (18%) were found to be direct copies of other essays in the corpus. The number of copies ranged widely, with 661 unique essays having two copies in the corpus and one essay having 720 copies (M = 4.91, SD = 25.89). Investigation revealed that many of these duplicate essays were directly copied from example essays given within Coursera to guide the students' writing. If an essay was found to have identical duplicates in the corpus, all instances of that essay were removed. This left us with a corpus of 22,903 essays. Additionally, only essays that had been reviewed by at least two peer-reviewers were retained. Thirty-two essays did not meet this criterion, leaving a baseline corpus of 22,871 essays.

2.3 Peer Review Scoring

Each remaining essay was reviewed by between 2 and 59 students in the same course (M = 4.94, SD = 2.8) on an analytic rubric. Peer-review focused on how well the author responded to each of the five required elements described in the previous section: challenge, application, selection, insight, approach, and organization. In addition, students were also asked to rate the organization of the essay, making a total of six criteria. Each criterion received a score between 1 and 3, giving a total score for each essay of between 6 and 18. Descriptive statistics for the scores on the analytic rubric can be seen in Figure 1.

Students were required to review at least three essays but were allowed to review as many essays as they liked, with many writers using the discussion forum to ask others to review their essays. Raters were given the opportunity to complete an optional bench-marking assignment in which they practiced rating a sample essay using the analytic rubric. However, no data was collected to reflect how many, or which students went through the bench-marking assignment. In addition to numerical scores, raters were also encouraged to leave comments for the essays they reviewed. Reviewers could leave a comment for each of the criteria as well as a single comment for overall feedback, creating a total of seven possible comments for each essay.

2.4 Peer-Review Reliability

Initial analysis of the peer-review scores raised questions about their instrument validity, as a correlation test revealed a low Pearson's product-moment correlations between holistic score and essay word count (r=0.165). As noted earlier, research has indicated strong correlations between word count and score (e.g., r 0.50- 0.80) [13, 17, 32, 47, 48]. The relatively low correlation between peer-review score and word count indicated low reliability for reviewer ratings. Closer analysis of rater scores revealed that many of the untrained peer-reviewers gave scores that appeared arbitrary and unconnected to the quality of the essays. For example, out of 41,242 reviewers, 12,731 (30.8%) always gave the same score to every essay they reviewed regardless of the writer or topic. Of these, 11,172 (87.8%) gave full points to every essay they reviewed. Thus, the final criterion for success in the MOOC is likely imprecise.

2.4.1 *Generating Data Subsets.*

To investigate the potential to derive more reliable peer-review ratings, we considered several different techniques and metrics to prune unreliable raters and only retain reviewers who gave reliable scores. We also set a threshold that each essay should retain at least two reviewers per essay. Our techniques for investigating reviewers were related to score variance and linguistic variation in comments left by the reviewers. Score variance was calculated as the sum of the squared difference from each reviewer's mean score. Variance of zero indicates that the reviewer gave the same score to every essay they reviewed while a larger variance means that the reviewer gave a wider range of scores. We hypothesized that reviewers with higher score variance may have been assigning scores more systematically, rather than simply giving the same or similar score to every essay they reviewed. In terms of linguistic variation, we examined the type/token ratio (TTR) for the comments left by the reviewer. Reviewer TTR was calculated by dividing the number of unique words (types) by the total number of words (tokens) for the aggregate comments of each reviewer. A high TTR means that the reviewer used a more diverse vocabulary relative to the number of words that they wrote, indicating greater lexical diversity. Reviewers with lower lexical diversity in their comments may have left the same or similar comments in each of the comment fields available to them. Low lexical diversity would indicate that the reviewer spent less time and effort in the review process and did not provide individualized feedback. As a result, these reviewers may have provided less systematic scores. While the Coursera platform requires reviewers to leave comments for each of the criteria on the analytic rubric, 2.5% of reviewers (N=1,043) left responses that did not include any alphanumeric characters, including leaving reviews consisting only of spaces, hyphens, or emojis. Of those who did leave substantive comments, some raters left detailed critiques while others repeated the same word or phrase multiple times on the papers they reviewed. The distribution of TTR for all raters can be seen in Figure 1.

Variance in scores and lower lexical diversity in comments taken in isolation may be insufficient to capture peer-rater reliability. Raters who only scored a few essays may have a low score variance

Table 1: Criterion and Total Scores Descriptive Statistics

Domain	n	mean	std	min	max
Application	22,871	2.62	0.63	1	3
Approach	22,871	2.54	0.69	1	3
Challenge	22,871	2.68	0.6	1	3
Insight	22,871	2.6	0.64	1	3
Organization	22,871	2.62	0.64	1	3
Selection	22,871	2.6	0.64	1	3
Total	22,871	15.65	3.25	6	18

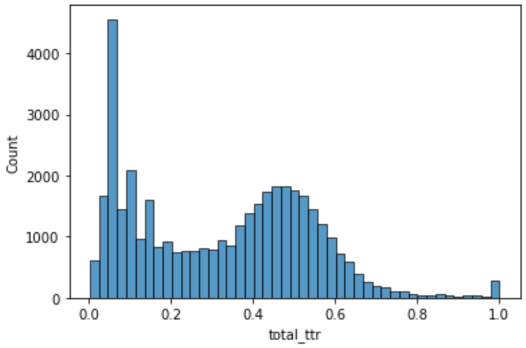


Figure 1: Histogram of Reviewer Type/Token Ratio

despite giving ratings that accurately reflect the quality of the assignment. Similarly, raters who only left a few comments might have a deceptively high type/token ratio because of the low word count of their comments. In order to take these ideas into account, we also looked at combinations of variance and lexical diversity measures.

To investigate how controlling peer-reviewed ratings based on variance and lexical variation might increase rater reliability, we generated data subsets by pruning reviewers who fell below given thresholds in these metrics, generating three subsets based on variance thresholds and three based on TTR thresholds. These thresholds were set based on an analysis of the data with the intention of a selection that would maximize the correlation between score and text length while minimizing the number of essays that would have to be removed. The three thresholds were categorized as low, medium, and high. For reviewer variance, we set a low threshold of variance greater than 0, a medium threshold of variance greater than 1, and a high threshold of variance greater than 2.5. For lexical diversity, the low threshold was set at 0.05, the medium threshold was set at 0.2, and the high threshold was set at 0.4. In addition, we also generated three data subsets by combining the two strategies where we retained reviewers who were above thresholds on variance and/or TTR. This included one subset which included reviewers with variance of greater than zero or TTR of greater than 0.05, one which included reviewers with variance of greater than zero or TTR of greater than 0.4, and one which included reviewers with variance of greater than 2.5 or TTR of greater than 0.4. In total, we created nine subsets and had ten sets in total, including the full set of data which was retained as a control.

Once we pruned unreliable raters, we compiled mean scores for those essays that retained two or more raters. The correlation between these scores and word-count for those essays were used as measures of criterion validity for the data in the pruned subsets. However, because pruning raters leads to fewer raters and fewer raters leads to fewer essays with two or more raters, each of these options for pruning reviewers involves a trade between sample size and instrument validity. Accordingly, the decision procedure to determine the best subset was non-trivial.

2.4.2 *Evaluating Data Subsets using Large Language Models.*

Our first approach to assessing the reliability of the pruned datasets was a simple correlation between essay scores and essays lengths. We would expect a stronger, positive correlation for more reliable data. However, this approach does not examine the content of the essays themselves. Thus, we used a second approach to determine whether the scores from our pruned datasets were assigned systematically rather than arbitrarily. Our second approach evaluated whether peer-reviewed scores can be predicted by LLMs that examine the semantics of the texts themselves and act as AES system. If LLM AES systems can accurately predict the essays scores based on language features within the essays, then the scores provide by the pruned rater subsets are likely more systematic.

In order to investigate the potential for LLMs to predict peer-reviewed scores, we generated encoder-only transformer language models for all nine of the groups as well as a baseline group in which all reviewers were retained. Transformer language models, such as Bidirectional Encoder Representations from Transformers (BERT) [16], are neural networks which tokenize input text into words or subwords, then convert each token into a high-dimensional vector based on the distribution of the word in the training corpus. These vectors are fed into a large pretrained neural network model as a matrix where they are transformed according to weights learned during training. The transformer's final state is a vector of numerical values which semantically represents the original text and can be used for classification, regression, or other tasks. Although the pretrained models can be seen as representing the language present in their training sets, they can be finetuned, a process by which labelled data from the target domain is used to further train the model for a specific purpose. Transformer models have been used extensively to successfully predict writing quality scores [7, 10, 25].

Each of the ten subsets of essays were split into train/validation/test groups using a 70/15/15 split. We used the texts and their associated scores in the training and validation sets to finetune distilBERT pretrained models and sequence classification heads, configured for linear regression. DistilBERT is a relatively light-weight pretrained language model which uses the English language Wikipedia as a training corpus. It is 60% faster and 40% smaller than its more well-known parent model BERT, while providing 97% of its BERT's language capabilities [38]. We chose this model over BERT in order to conserve computational resources in light of the environmental impact of pervasive computing [21] as well as for efficiency in adaptation into learning systems. Root mean squared error (RMSE) was used as a metric during the training process so that, during training, the model weights always moved toward a configuration with the lowest RMSE through gradient descent. After the ten models had been generated, each trained on its own subset of the data, each finetuned model was used to predict scores in the test group of that subset of essays. Pearson's product-moment correlation scores were calculated to show the correlation between the scores as predicted and the actual scores. These correlations were used to evaluate which of the subsets produced the most accurate models. More accurate models would indicate that peer-reviewed scores assigned in that dataset are more systematically related to linguistic features within the texts and thus more reliable.

3 RESULTS

3.1 Assessing Rater Reliability

3.1.1 Reviewer Score Variance. The first metric we examined was the variance of the reviewer's scores. Panel A of Figure 2 shows the results of pruning reviewers based on the variance of their scores. Eliminating reviewers with a variance of 0 led to an increased correlation with essay length (*r*= .265) while retaining 19,336 essays with at least two reviewers. If we increase the threshold to greater than one, we observe a higher correlation between essay length and score (*r*= .314) at the cost of losing 45% of the total number of essays (n = 12,574). When the minimum reviewer variance is increased to 2.5, a stronger correlation with essay length is attained (*r*= .336), but only 8,066 essays with at least two reviewers can be retained (i.e., 35% of the original set of essays).

As with reviewer score variance, we investigated a low, medium, and high threshold for response TTR. Our low threshold of .05 resulted in a correlation score of r=0.196 with essay length and retained 93% of the essays (n = 21,157). Our medium threshold of 0.2 resulted in textitr= .270 with essay length and retained 74% of essays (n = 16,875). Our high threshold of 0.4 results in *r*= .314 with essay length and retained 49% of essays (n = 11,234). The results of pruning reviewers at these three thresholds on score correlation with text length and number of authors retained can be seen in Panel B of Figure 2.

We also developed subsets based on combining the reviewer score variance and TTR in the comments to remove unreliable reviewers. In this approach, we applied three threshold conditions to both TTR and variance and reviewers who met either criterion or both criteria were accepted. In other words, this approach retains all reviewers with high variance or high TTR. Results of these approaches can be seen in Panels C and D of Figure 2. In Panel C, any reviewer with a score variance of greater than zero is retained and, additionally, reviewers with TTR greater than a given value (in this case 0.05 and 0.40) were also retained. Both subsets also retained reviewers with variance greater than zero. At the lower TTR threshold of > 0.05 we retained 97% (n=22,244) of essays with a correlation of *r*= .191 with essay length and at the higher threshold of TTR > 0.4 we retained 88% (n=20,209) of essays with a correlation of *r*= .25 with essay length. Finally, Panel D shows the effect of retaining all reviewers with a TTR higher than 0.4 and selecting additional reviewers based on variance threshold of 2.5. This subset retained 70% of essays (n= 16,116) and reported a correlation of *r*= .297 with text length.

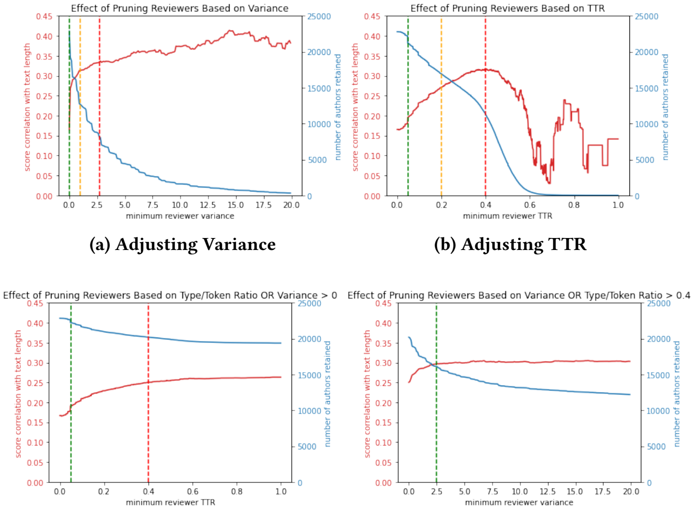
A full review of results for all the thresholds is presented in Table 2. Pruning reviewers led to increases in the correlation between peer-reviewed scores and word-count, which likely indicates increased reliability of the remaining reviewers. However, pruning also led to much smaller sample sizes because each subset had a different number of essays. The number of essays in the ten groups were variable with a maximum of 22,871 in the full set of essays and a minimum of 8,066 in the set which only included essays with score variance of 2.5 or higher.

3.2 Assessing Final Score Reliability Using distilBERT

To further test the validity of ratings in our pruned datasets, we used each of the datasets to build separate LLM AES systems by fine tuning the distilBERT pretrained model to predict the peer-reviewed scores. These models can provide increased validity for the subsets discussed above because they can map linguistic features in the essays to the peer-reviewed ratings of quality. After training the LLM AES systems, we used Pearson's product-moment correlations to compare the scores predicted by the models with the actual scores assigned by the reviewers in the test groups. Table 3 displays the numbers of essays in the three groups (test/validation/train) for each subset, as well as the Pearson's r value for the model's predictions in the test groups. While each of the pruned models improved on the performance of the baseline model, the best performance was seen in the high variance threshold model (variance > 2.5) followed by the medium variance threshold model (variance >1) and the combined variance > 2.5 or TTR > 0.4 model. The worst performing models was the low TTR threshold model (TTR > 0.05) which underperformed the control group consisting of all raters. Performance increases, however, came at a cost with the best performing model excluded 65% of the essays. The second strongest model excluded 45% of essays while the third strongest model excluded 30% of the essays. Figure 3 displays scatter-plots in which the actual scores assigned by raters are graphed on the y axis while the scores predicted by the models are graphed on the x axis. The distilBERT models that were finetuned on the pruned data were better at predicting scores than the model finetuned on the baseline data set, suggesting that the scores in those data subsets are more closely linked to linguistic features within the texts. The predicted scores were also highly correlated with the score to word count correlations in each subset (*r*= .943, p < 0.001).

4 DISCUSSION AND CONCLUSION

Low rater reliability poses a serious challenge to the validity of using peer-assigned scores in open online courses, limiting the quality of the feedback the students receive as well as calling into question grading criteria. To increase the quality of feedback students



(a) Adjusting Variance

(b) Adjusting TTR

(c) Adjusting TTR or Variance > 0

(d) Adjusting Variance or TTR > 0.4

Figure 2: Numbers of Essays Retained and Correlation between Text Length and Score after Pruning Raters

Table 2: Comparison of Ten Data Subsets

Model	N	proportion of essays retained	text length and score *r*
All reviewers	22,871	1.00	0.165
Variance > 0	19,336	0.85	0.265
Variance > 1	12,574	0.55	0.314
Variance > 2.5	8,066	0.35	0.336
TTR > 0.05	21,157	0.93	0.196
TTR > 0.2	16,875	0.74	0.270
TTR > 0.4	11,234	0.49	0.314
Variance > 0 or TTR > 0.05	22,244	0.97	0.191
Variance > 0 or TTR > 0.4	20,209	0.88	0.250
Variance > 2.5 or TTR > 0.4	16,116	0.70	0.297

Table 3: Correlation Between Actual and Predicted Scores in LLMs

Model	N	train n	valid n	test n	predict r
All reviewers	22,871	16,009	3,431	3,431	0.192
Variance > 0	19,336	13,535	2,900	2,901	0.320
Variance > 1	12,574	8,801	1,886	1,887	0.360
Variance > 2.5	8,066	5,646	1,210	1,210	0.456
TTR > 0.05	21,157	14,809	3,174	3,174	0.189
TTR > 0.2	16,875	11,812	2,531	2,532	0.307
TTR > 0.4	11,234	7,863	1,686	1,685	0.335
Variance > 0 or TTR > 0.05	22,244	15,570	3,337	3,337	0.217
Variance > 0 or TTR > 0.4	20,209	14,146	3,032	3,031	0.281
Variance > 2.5 or TTR > 0.4	16,116	11,281	2,418	2,417	0.358

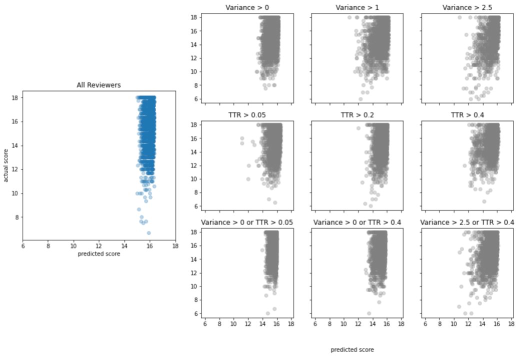


Figure 3: Scatterplot of Predicted/Actual Scores

receive and to increase student trust in the scoring mechanisms used in MOOCs, peer-review ratings need to be more reliable. This study addressed this by analyzing a large sample of peer-reviewed essays submitted as a capstone project in a MOOC by identifying unreliable raters based on measures of lexical diversity and score variance. Data subsets were generated by pruning unreliable raters, which increased correlations between peer-reviewed scores and essay length. The scores for the subsets were further validated by finetuning LLM AES systems using each of the data subsets as training and test sets. Including only raters with higher comment TTR or variance resulted in models that were more predictive than models built using the entire set of rater data.

Increasing the threshold for reviewer score variance had the strongest effect on our measures of score validity, including both correlation between score and word count as well as the predictive power of the language models. However, this method also greatly reduced the number of essays with at least two raters, retaining only 35% of essays at the highest threshold (variance > 2.5). Our best subset model was likely the low threshold (variance > 0), which retained 85% of the variance and reported a correlation with essay length of .27. This same subset when used to train an LLM AES system reported a correlation between actual and predicted essay score of .32. As a result, this approach may not be a practical solution in many educational or research contexts because many reviewers are categorized as unreliable. Using variance alone may also penalize reviewers who only reviewed a few essays. For example, if a reviewer only rated two essays that happened to be of the same quality, that reviewer might have a low score variance despite giving valid and reliable scores.

Pruning reviewers by using type/token ratio as a measure of lexical diversity in the comments also increased the reliability of the scores over the full dataset and allowed us to retain more essays as compared to using score variance, although the correlation with essay length was not as strong. Our best subset were for reviewers in the middle threshold (variance > 2) which reported a correlation with essay length of *r* = .27 while retaining 74% of the reviewers. In terms of the LLM AES systems, the best performing model was for the middle threshold for which a correlation of r = .31 was reported. There are many factors that might create a disconnect between a reviewer's comment and score quality. For example, some reviewers might have been reluctant to give lower than full points despite leaving detailed comments while others may have graded carefully and fairly according to the rubric while not taking the time to leave more than one-word comments.

The combined method in which we retained reviewers who had either a high score variance or high lexical diversity in their comments provided the strongest results and seems to address potential limitations of both of the methods in isolation. This approach retained reviewers who graded fairly but left minimal comments as well as reviewers who only reviewed a few essays but were determined to be reliable reviewers based on their detailed comments. The combined method led to measures of reliability that were almost as good as variance only, while being able to retain more essays. This is particularly true at the highest threshold of 2.5 variance or 0.4 TTR which retained 70% of the essays and reported a correlation of *r*= .30 with essay length. An LLM AES model for this subset reported a correlation of *r*= .36 between predicted and actual essay scores. This subset is likely the sweet spot for retaining as much data as possible that reports strong reliability with essay length and essay quality.

4.1 Implications

These results have implications to the fields of learning analytics and learning engineering, particularly in the case of MOOCs that depend on peer-reviewing to assess open ended writing assignments. Providing high-quality feedback to written responses in MOOCs is critical to developing classes that assess student knowledge through student production. We know that it is virtual impossible for instructors to provide personalized feedback to open ended questions for the large number of students in the class. We also know that open-ended questions lead to greater learning and are better approaches to assessing knowledge [3, 6]. While AES systems can provide some level of reliability in assessing open-ended questions, AES systems cannot provide the granular feedback that students need and deserve. Additionally, AES systems can only approximate human scoring and generally need to be trained for each task and topic they are assigned. Training AES systems specific for each assignment is a luxury unavailable to the majority of MOOC instructors and course content providers. Thus, there is a strong need to include a human in the loop for open-ended MOOC assessments and the only real candidates are the students themselves.

This study demonstrates the potential to assess the reliability of peer reviewers using relative simple metrics Once peer-rater reliability can be accurately measured, peer-reviews can be more readily incorporated into the pedagogical framework of the MOOCS providing students with grades and feedback that better reflect their knowledge and effort. In turn, students will gain trust in the MOOCs and have greater confidence in their learning and motivation to continue learning. Additionally, students who provide helpful and valid feedback can be rewarded within the system not just in terms of knowledge transfer, but also medals, prizes, or extra credit. Such an approach would incentivize unreliable students.

4.2 Limitations and Future Directions

This study has demonstrated methods to access peer-rater reliability with confidence, but the results are subject to limitations. Notably, in our best-case scenario, 30% of the students in the MOOC do not have two scores assigned by reliable raters. A simple solution to this would be to increase the number of reviewers per assignment to limit the odds of pruning too many reviewers. A second solution would be asking reliable reviewers to peer-review assignments that need additional reviews and providing extra credit as incentive.

Another limitation to this study is the small number of features we used to assess peer-rater reliability. Lexical diversity and score variance are only two of many possible metrics that could be employed to ascertain the reliability of raters in MOOCS. Future research may take into account any number of other variables available in the system. These may include rater behavior on the platform obtained through clickstream data, linguistic features found in comments other than lexical diversity, and more sophisticated analysis of scores. Any of these may improve the reliability metrics discussed here and help determine the optimal decision criteria to prune raters.

Lastly, while the strong link between LLM AES system accuracy to score/text-length correlation (r=0.943) provides support for using LLM AES systems, these results should be interpreted with caution. The main concern is the limited interpretability of LLMs, which generally makes it impossible to know what criteria they used to predict scores. It is possible that scores were predicted partially based on word count itself, meaning that the two measures are collinear rather than two measures of different construct. Future research may establish a ground truth by having the essays rated by expert raters as well as peer raters, then testing the correlation between large language model predictions and expert scores for different subsets of essays.

REFERENCES

[1] Ramón Alcarria, Borja Bordel, Diego Martín de Andrés, and Tomás Robles. 2018. Enhanced Peer Assessment in MOOC Evaluation Through Assignment and Review Analysis. *International Journal of Emerging Technologies in Learning (iJET)* 13, 01 (Jan. 2018), 206. https://doi.org/10.3991/ijet.v13i01.7461

[2] Gabriel Badea and Elvira Popescu. 2022. A dynamic review allocation approach for peer assessment in technology enhanced learning. *Education and Information Technologies* (June 2022). https://doi.org/10.1007/s10639-022-11175-5

[3] Elizabeth Badger and Brenda Thomas. 1992. Open-Ended Questions in Reading. *Practical Assessment, Research and Evaluation* 3, 4 (1992). https://doi.org/10.7275/FRYF-Z044 Publisher: University of Massachusetts Amherst.

[4] Stephen P Balfour. 2013. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment* 8 (2013), 40–48.

[5] William E. Becker and Carol Johnston. 1999. The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record* 75, 4 (Dec. 1999), 348–357. https://doi.org/10.1111/j.1475-4932.1999.tb02571.x

[6] John Bennion, Brian Cannon, Brian Hill, Riley Nelson, and Meagan Ricks. 2020. Asking the Right Questions: Using Reflective Essays for Experiential Assessment. *Journal of Experiential Education* 43, 1 (March 2020), 37–54. https://doi.org/10.1177/1053825919880202

[7] Majdi Beseiso and Saleh Alzahrani. 2020. An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications* 11, 10 (2020).

[8] Mina Shirvani Boroujeni and Pierre Dillenbourg. 2018. Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, Sydney New South Wales Australia, 206–215. https://doi.org/10.1145/3170358.3170388

[9] Stephen Bostock. 2000. Student peer assessment. *Learning Technology* 5, 1 (2000), 245–249.

[10] Robert-Mihai Botarleanu, Mihai Dascalu, Laura K Allen, Scott Andrew Crossley, and Danielle S McNamara. 2022. Multitask Summary Scoring with Longformers. In *International Conference on Artificial Intelligence in Education*. Springer, 756–761.

[11] George A Brown, Joanna Bull, and Malcolm Pendlebury. 2013. *Assessing Student Learning in Higher Education* (0 ed.). Routledge. https://doi.org/10.4324/9781315004914

[12] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, Edinburgh United Kingdom, 6–14. https://doi.org/10.1145/2883851.2883931

[13] Scott Crossley, Qian Wan, Laura Allen, and Danielle McNamara. 2021. Source inclusion in synthesis writing: an NLP approach to understanding argumentation, sourcing, and essay quality. *Reading and Writing* (2021).

[14] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48, 4 (2016), 1227–1237.

[15] R Wes Crues, Nigel Bosch, and Carolyn J Anderson. 2018. Who they are and what they want: Understanding the reasons for MOOC enrollment. *Proceedings of the 11th International Conference on Educational Data Mining* (2018), 10.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan Keller, and Olaf Köller. 2020. Is a Long Essay Always a Good Essay? The Effect of Text Length on Writing Assessment. *Frontiers in psychology* 11 (2020), 562462.

[18] Dilrukshi Gamage, Thomas Staubitz, and Mark Whiting. 2021. Peer assessment in MOOCs: Systematic literature review. *Distance Education* 42, 2 (April 2021), 268–289. https://doi.org/10.1080/01587919.2021.1911626

[19] Felix Garcia-Loro, Sergio Martin, José A. Ruipérez-Valiente, Elio Sancristobal, and Manuel Castro. 2020. Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform. *Computers & Education* 154 (Sept. 2020), 103894. https://doi.org/10.1016/j.compedu.2020.103894

[20] Xinyue Guo, Feng Wu, and Xin Zheng. 2019. What Motives Learner to Learn in MOOC? An Investigation of Chinese University MOOC. In *2019 International Joint Conference on Information, Media and Engineering (IJCIME)*. IEEE, Osaka, Japan, 154–159. https://doi.org/10.1109/IJCIME49369.2019.00039

[21] Andreas Köhler and Lorenz Erdmann. 2004. Expected environmental impacts of pervasive computing. *Human and Ecological Risk Assessment* 10, 5 (2004), 831–852.

[22] Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly* 49, 4 (2015), 757–786.

[23] Kristopher Kyle and Scott A Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* 102, 2 (2018), 333–349.

[24] Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly* 18, 2 (2021), 154–170.

[25] Paraskevas Lagakis and Stavros Demetriadis. 2021. Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, Istanbul, Turkey, 1–6. https://doi.org/10.1109/CITS52676.2021.9618476

[26] Hongxia Li, ChengLing Zhao, Taotao Long, Yan Huang, and Fengfang Shu. 2021. Exploring the reliability and its influencing factors of peer assessment in massive open online courses. *British Journal of Educational Technology* 52 (2021), 2263–2277.

[27] Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics* 15, 4 (2010), 474–496.

[28] Robert Lukhele, David Thissen, and Howard Weiner. 1994. On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement* 31, 3 (1994), 234–250.

[29] Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring?. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Seattle, WA, USA → Online, 151–162. https://doi.org/10.18653/v1/2020.bea-1.15

[30] Rolin Moe. 2015. The brief & expansive history (and future) of the MOOC: Why two divergent models share the same name. *Current Issues in Emerging eLearning* 2, 1 (2015), 24.

[31] Yasuhiro Ozuru, Stephen Briner, Christopher A. Kurby, and Danielle S. McNamara. 2013. Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 67, 3 (2013), 215–227. https://doi.org/10.1037/a0032918

[32] Les Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing* 21 (July 2014), 104–111. https://doi.org/10.1016/j.asw.2014.05.001

[33] Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics* 110, 1 (April 2018), 43–70. https://doi.org/10.2478/pralin-2018-0002 arXiv:1804.00247 [cs].

[34] Justin Reich and José A. Ruipérez-Valiente. 2019. The MOOC pivot. *Science* 363, 6423 (Jan. 2019), 130–131. https://doi.org/10.1126/science.aav7958

[35] Erin Dawna Reilly, Rose Eleanore Stafford, Kyle Marie Williams, and Stephanie Brooks Corliss. 2014. Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *The International Review of Research in Open and Distributed Learning* 15, 5 (Oct. 2014). https://doi.org/10.19173/irrodl.v15i5.1857

[36] Zhiyun Ren, Huzefa Rangwala, and Aditya Johri. 2016. Predicting Performance on MOOC Assessments using Multi-Regression Models. http://arxiv.org/abs/1605.02269 arXiv:1605.02269 [cs].

[37] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. http://arxiv.org/abs/1909.09482 arXiv:1909.09482 [cs, stat].

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[39] Huma Shafiq. 2017. Courses beyond borders: A case study of MOOC platform Coursera. *Library of Philosophy and Practice* (2017), 1566.

[40] Thomas Staubitz. 2020. *Gradable team assignments in large scale learning environments: collaborative learning, teamwork, and peer assessment in MOOCs: Kollaboratives Lernen, Teamarbeit und Peer Assessment in MOOCs.* Ph. D. Dissertation. Universität Potsdam. https://doi.org/10.25932/PUBLISHUP-47183 Artwork Size: 16774 KB, 70133 KB, 122 pages Medium: application/pdf,application/zip Pages: 16774 KB, 70133 KB, 122 pages.

[41] Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*. ACM, Edinburgh Scotland UK, 389–398. https://doi.org/10.1145/2876034.2876043

[42] Thomas Staubitz, Hanadi Traifeh, Salim Chujfi, and Christoph Meinel. 2020. Have Your Tickets Ready! Impede Free Riding in Large Scale Team Assignments. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*. ACM, Virtual Event USA, 349–352. https://doi.org/10.1145/3386527.3406744

[43] Hoi K. Suen. 2014. Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning* 15, 3 (June 2014). https://doi.org/10.19173/irrodl.v15i3.1680

[44] Bruce W. Tuckman. 1993. The Essay Test: A Look at the Advantages and Disadvantages. *NASSP Bulletin* 77, 555 (Oct. 1993), 20–26. https://doi.org/10.1177/019263659307755504

[45] Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika* 48, 2 (2021), 459–484.

[46] Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. *arXiv preprint arXiv:2205.03835* (2022).

[47] Joshua Wilson and Jessica Rodrigues. 2020. Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology* 82 (2020), 123–140.

[48] Haoran Zhang and Diane Litman. 2021. Essay Quality Signals as Weak Supervision for Source-based Essay Scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. 85–96.

Received October 7, 2022; accepted December 1, 2022