Optimal clustering of item responses to determine scoring inconsistencies by human raters

Kinansa Husainy

Introduction

Educational measurement aims to quantify students' abilities in a fair and accurate way, providing the foundation for decisions about learning progress. To fully capture the complexity of the students' abilities, exams are oftentimes constructed to go beyond the multiple-choice items and include open-ended questions. This way, their high levels of cognitive understanding (like reasoning) can be assessed, and leaves little room for guessing (Gharehbagh et al., n.d.). Responses to open-ended questions are typically scored by one or more human raters. Ideally, raters should give consistent scores for answers that are alike or comparable. In practice, however, similar answers do receive different scores across raters, threatening the so-called inter-rater reliability.

As human scoring takes time, adds workload to the teachers, and is still prone to inconsistencies, there has been research into automated scoring of open-ended items. Various methods have been explored in the context of exam grading in education, namely through natural language processing (NLP) approach like word embeddings and transformer architectures (Haller et al., n.d.; Gupta, n.d.), or even large language models (LLMs) (Mansour et al., n.d.). With well-designed prompts, appropriate rubrics, and sometimes data augmentation, LLMs and transformer models can achieve high quality automated scoring (Wu et al., n.d.; Xie et al., n.d.; Gupta, n.d.). The scoring for any types of open-ended answers in different educational settings could be explored with these methods, provided that we continuously evaluate the performance and appropriateness of each method.

In automated scoring of short answers with NLP, the embeddings of item responses together with their scores are used to train a classifier model. As these scores come from the raters, the quality (e.g., consistency) of the grading will affect the system's training. When the noise is random, it is suggested that the system performed just as well as a system trained on a "cleaned" dataset (Reidsma and Carletta 2008). However, that is not the case when the noise is from human grading, thus not always random. As rater disagreement (inconsistent human grading) introduces label noise that affects model training and evaluation, Loukina et al. (2020) proposed using the metric of proportional reduction in mean squared error (PRMSE). Therefore, training various models on an inconsistent dataset is undesirable as it will lead to unreliable outcomes.

To come up with a consistently labeled dataset, it is important to be able to detect inconsistencies in the data. Therefore, a proper grouping of similar answers is needed, as we expect these to get the same score. Some attempts with NLP method that have been tried to cluster similar responses were edit similarity, cosine similarity, Jaccard similarity, and normalized word count (Ahmad and Why, n.d.) , while others explored clustering through word embedding and deep learning methods (Chang et al., n.d.).

This study will explore which method is the best practice in detecting and filtering these similar answers (and to what extent of similarity). To explore this, it is important to consider the situation we have at hand and which alternative would give an applicable insight. We then come up with a research question: what is the optimal procedure for clustering open item responses to determine scoring inconsistencies by human raters?

Methods

We address the problem at hand through a three-stage analytical process: (1) text preprocessing, (2) response clustering, and (3) evaluation of labeling consistency.

Analytical Strategy

Preprocessing standardizes textual responses to minimize styling (punctuation, capitalization, and articles) variation while retaining semantic information. Some common approaches are tokenization, stop word removal, and lemmatization, and effective preprocessing is argued to improve the system performance in text classification (Gabon 2025). There is no one size-fits-all solution in doing text preprocessing. Several literatures have explored different selection of techniques and in different order. Depending on the task performed on the data, one technique would be more appropriate than another. Centre for Language Studies, Radboud University, The Netherlands et al. (2019) explored different techniques on different datasets, on the task of question similarity. In doing so, the techniques and ordering performed differently. Siino, Tinnirello, and La Cascia (2024) also reviewed the performances of these techniques in different NLP tasks. In general, it is suggested to go with minimal text normalization first like lowercasing early and then move to more complex transformation later on. It is also suggested that for morphologically rich language, it is better to do the lemmatization first and then the stop words removal last (Kestemont et al. 2017).

For response clustering, we try out alternatives: the cosine similarity of the word embeddings and string matching with a correction model. For cosine similarity, we determine the acceptable threshold to examine the similarity. It will be run by using the spaCy package in R. Variances in the clusters shall be analyzed.

In those selected clusters, we measure the consistency, or rather, the inconsistency of the labeling. This metric is done through the method of measuring uncertainty with similarity-sensitive entropy, using the entropy package, which quantifies how scattered or concentrated the label distribution is within a cluster (Cheng and Vlachos 2024). Clusters with high entropy will be flagged as they indicate high inconsistency. Through the metric results, we will be able to determine which is the best practice in detecting the inconsistency.

Lastly, the highly inconsistent clusters will then be relabeled to achieve a gold-standard dataset that is of high consistency, through generative artificial intelligence (AI) LLM prompting, as it shows promising performance (Flodén 2025; Zhang et al. 2024; Poličar et al. 2025).

Dataset

These analyses will be applied on the open-ended answers from the students taking the Cito national exam, administered from 2022-2025. The exams were for the basic vocational pathway (vmbo-bb), and the intermediate vocational pathway (vmbokb). The dataset was pre-processed, clustered with NLP for visual grouping, and each response was given a grade of either 0 or 1. The subjects ranged from geography, biology, economy, history, nature and chemistry, society, to mathematics.

Results

i. Text Preprocessing

In doing the text preprocessing, we first apply the cleaning of the raw data first, because some responses were still in HTML format. This step is rather a cleaning process, and not a normalization itself. This is to ensure that subsequent preprocessing steps operate on comparable plain-text input. We then do the minimal preprocessing first – (i) the lowercasing and (ii) the removal of punctuation. Punctuation is critical for finding boundaries of things and for identifying some aspects of meaning, but the NLP algorithm is usually run only inside words, thus not merging across word boundaries (Jurafsky and Martin 2026). This is why we remove white space and punctuation. Afterwards, we do (iii) the lemmatization of the words and then (iv) the stop words removal last. We apply Levenshtein's edit distance for stage 1 and 2 to measure the minimum number of singlecharacter edits (insertions, deletions, or substitutions) required to change a word into another (Levenshtein, n.d.). Word-level edit distance is applied for stage 3 (as lemmatization affects words as a whole, not character-level), and word count change in stage 4.

We first group items of similar type of responses, simply through the character length. This method of grouping serves as a proxy for response complexity, and allows us to detect item effects, before we move on to the next step that is response clustering. We calculate the median character length of the response, normalize and scale the metric from 0 to 1, all within subjects. After being scaled, the median character length is then separated into 3 bins – again, within subjects. This grouping will help us to evaluate whether each preprocessing step behaves the same or consistently across item groups, or whether the items show different patterns. The assignment of this character length grouping into the responses is done twice, before and after doing all the preprocessing stages, allowing us to inspect whether preprocessing itself alters the relative positioning of responses across length-based groups.

We want to have a clear picture of the changes caused by the preprocessing stage, thus we need to properly detect the outliers. However, we cannot simply define outliers from the whole dataset, as we have multiple subjects. The nature of each subject will result in particular patterns of answers, so such factors must be taken into account to consider the difference caused by the item behavior, also known as item effects. For this reason, all outlier detection procedures are performed within subjects.

Here, outliers are defined as responses that are outside the quantile - median character length higher than the 95th percentile and lower than 5th percentile. The proportion is the number of outliers compared to the total responses. We do not do any exclusion yet to the outliers before the preprocessing. For comparison, we see a considerable amount of reduction of outliers before and after we do the preprocessing, meaning the preprocessing stages made the response item length shift more towards the mean. Before preprocessing, the proportion of outliers is about 30-60% of the total responses, with group 2 of item length group having the highest proportion of 67.2%. After preprocessing, all groups have substantially reduced outlier proportion of less than 10%.

Table 1: Outlier Proportion Before and After Preprocessing

Item Length Group	n of All Responses	n of outliers (before preprocessing)	Proportion of outliers (before preprocessing)	n of outliers (after preprocessing)	Proportion of outliers (after preprocessing)
1	38,577	12,424	32.2%	3,396	8.8%
2	23,485	15,777	67.2%	2,047	8.7%
3	15,360	9,806	63.8%	1,501	9.7%

The flagged outlier responses after the preprocessing stages are now being excluded from further visualizations. This is because outliers are treated as responses that are likely driven by extreme formatting, or non-representative input patterns, rather than substantive semantic content.

Since I cannot share my thesis data here, I will be placing a dummy figure using an available `college` dataset from the package `ISLR`.

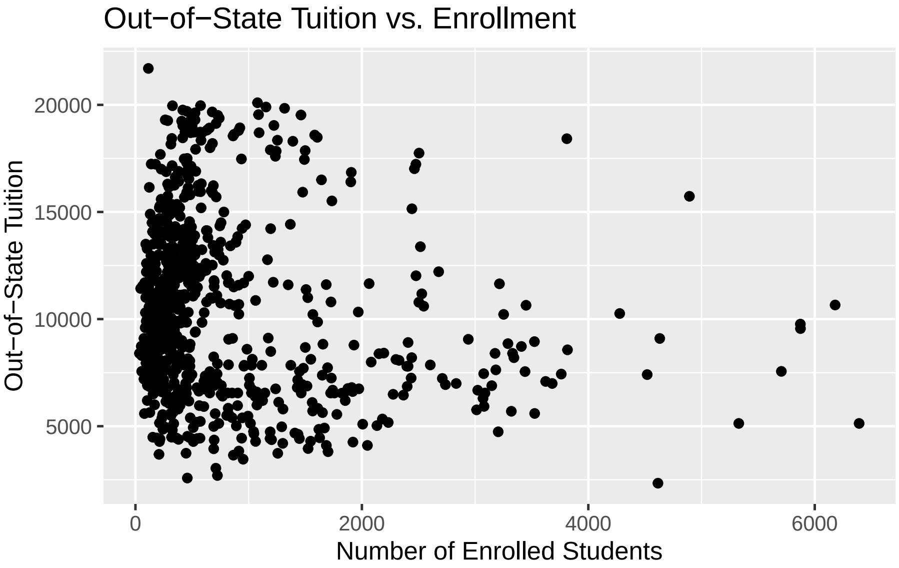

Figure 1: Out-of-State Tuition vs. Enrollment

A simple linear regression model can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

References

Ahmad, Ayaan, and Dr. Ng Kok Why. n.d. "Automated Grading Using Natural Language Processing and Semantic Analysis." https://doi.org/10.2139/ssrn.4999531.

Centre for Language Studies, Radboud University, The Netherlands, Florian Kunneman, Thiago Castro Ferreira, Tilburg center for Cognition and Communication, Tilburg University, The Netherlands, Emiel Krahmer, Tilburg center for Cognition and Communication, Tilburg University, The Netherlands, Antal Van Den Bosch, Centre for Language Studies, Radboud University, The Netherlands, and KNAW Meertens Institute, Amsterdam, The Netherlands. 2019. "Recent Advances in Natural Language Processing." In, 593–601. Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-056-4_070.

Chang, Li-Hsin, Iiro Rastas, Sampo Pyysalo, and Filip Ginter. n.d. "Deep Learning for Sentence Clustering in Essay Grading Support." https://doi.org/10.48550/arXiv.2104.11556.

Cheng, Julius, and Andreas Vlachos. 2024. "EACL 2024." In, edited by Yvette Graham and Matthew Purver, 21152128. St. Julian's, Malta: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.129.

Flodén, Jonas. 2025. "Grading Exams Using Large Language Models: A Comparison Between Human and AI Grading of Exams in Higher Education Using ChatGPT." *British Educational Research Journal* 51 (1): 201–24. https://doi.org/10.1002/berj.4069.

Gabon, Dennis C. 2025. "Automated Grading of Essay Using Natural Language Processing: A Comparative Analysis with Human Raters Across Multiple Essay Types." *Journal of Information Systems Engineering and Management* 10 (6s): 67–72. https://doi.org/10.52783/jisem.v10i6s.700.

Gharehbagh, Zahra Abdolreza, Azam Mansourzadeh, Atiyeh Montazeri Khadem, and Masumeh Saeidi. n.d. "Reflections on Using Open-Ended Questions."

Gupta, Kshitij. n.d. "Data Augmentation for Automated Essay Scoring Using Transformer Models." https://doi.org/10.48550/arXiv.2210.12809.

Haller, Stefan, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. n.d. "Survey on Automated Short Answer Grading with Deep Learning: From Word Embeddings to Transformers." https://doi.org/10.48550/arXiv.2204.03503.

Jurafsky, Daniel, and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed. https://web.stanford.edu/~jurafsky/slp3/.

Kestemont, Mike, Guy de Pauw, Renske van Nie, and Walter Daelemans. 2017. "Lemmatization for Variation-Rich Languages Using Deep Learning." *Digital Scholarship in the Humanities* 32 (4): 797–815. https://doi.org/10.1093/llc/fqw034.

Levenshtein, V. I. n.d. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals | BibSonomy." https://www.bibsonomy.org/bibtex/55f7ad93fcb9ae3ed999afaa6e24937d.

Loukina, Anastassia, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S. Johnson, Brian Riordan, and Daniel F. McCaffrey. 2020. "BEA 2020." In, edited by Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, 1829. Seattle, WA, USA → Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.2.

Mansour, Watheq, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. n.d. "Can Large Language Models Automatically Score Proficiency of Written Essays?" https://doi.org/10.48550/arXiv.2403.06149.

Poličar, Pavlin G, Martin Špendl, Tomaž Curk, and Blaž Zupan. 2025. "Automated Assignment Grading with Large Language Models: Insights from a Bioinformatics Course." *Bioinformatics* 41 (Supplement_1): i21–29. https://doi.org/10.1093/bioinformatics/btaf196.

Reidsma, Dennis, and Jean Carletta. 2008. "Reliability Measurement Without Limits." *Computational Linguistics* 34 (3): 319–26. https://doi.org/10.1162/coli.2008.34.3.319.

Siino, Marco, Ilenia Tinnirello, and Marco La Cascia. 2024. "Is Text Preprocessing Still Worth the Time? A Comparative Survey on the Influence of Popular Preprocessing Methods on Transformers and Traditional Classifiers." *Information Systems* 121 (March): 102342. https://doi.org/10.1016/j.is.2023.102342.

Wu, Xuansheng, Padmaja Pravin Saraf, Gyeonggeon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. n.d. "Unveiling Scoring Processes: Dissecting the Differences Between LLMs and Human Graders in Automatic Scoring." https://doi.org/10.48550/arXiv.2407.18328.

Xie, Wenjing, Juxin Niu, Chun Jason Xue, and Nan Guan. n.d. "Grade Like a Human: Rethinking Automated Assessment with Large Language Models." https://doi.org/10.48550/arXiv.2405.19694.

Zhang, Da-Wei, Melissa Boey, Yan Yu Tan, and Alexis Hoh Sheng Jia. 2024. "Evaluating Large Language Models for Criterion-Based Grading from Agreement to Consistency." *Npj Science of Learning* 9 (1): 79. https://doi.org/10.1038/s41539-024-00291-1.