Measuring Argument Quality in Technology-Mediated Peer-Instruction

Sameer Bhatnagar
Antoine Lefebvre-Brossard
Michel C. Desmarais
Amal Zouaq
{sameer.bhatnagar,antoine.lefebvre-brossard,michel.desmarais,amal.zouaq}@polymtl.ca
Ecole Polytechnique Montreal
Canada

ABSTRACT

TO DO

CCS CONCEPTS

• Applied computing → Computer-assisted instruction; **• Computing methodologies → Natural language processing**.

KEYWORDS

Peer Instruction, Learnersourcing

ACM Reference Format:
Sameer Bhatnagar, Antoine Lefebvre-Brossard, Michel C. Desmarais, and Amal Zouaq. 2021. Measuring Argument Quality in Technology-Mediated Peer-Instruction. In *Proceedings of LAK '21: Learning Analytics & Knowledge (LAK '21).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.000/0000.0000

1 INTRODUCTION

Technology-mediated peer instruction *(TMPI)* platforms [2][20] expand multiple choice items into a two step process. On the first step, students must not only choose an answer choice, but also provide an explanation that justifies their reasoning. On the second step, students are prompted to revise their answer choice, by taking into consideration a subset of explanations written by their peers for another answer choice. In the case that the student wants to keep their original answer choice, but may be unsure of their own explanation, they are also shown peer-explanations for their original answer choice. The student now has three options:

(1) Change their answer choice, by indicating which of their peer's explanations was most convincing
(2) keep their answer choice, but *change explanations* by choosing one that is for the same answer as their own
(3) choose "I stick to my own", indicating that their own explanation is best from amongst those that are shown.

Whenever the student goes with either of the first two scenarios above, we frame this as "casting a vote" for the chosen peer explanation.

The design and growing popularity of TMPI is inspired by three schools of thought: firstly, prompting students to explain their reasoning is beneficial to their learning [5]. Second, classroom based *Peer Instruction*[6], often mediated by automated response systems (e.g. clickers), has become a prevalent, and often effective component in the teaching practice of instructors looking to drive student engagement as part of an active learning experience [3]. In discussing with peers *after* they have formulated their own reasoning, students are engaged in a higher order thinking task from Bloom's taxonomy, as they evaluate what is the strongest argument, before answering again. Thirdly, by capturing data on which explanations students find most convincing, TMPI affords teachers the opportunity to mitigate the "expert blind spot" [15], addressing student misconceptions they might not otherwise have thought of.

We suggest that the "vote" data collected on each explanation, is a proxy for argument quality, along the dimension of *convincingness*, as judged by peer learners. These votes can be aggregated into a *convincingness* score, as a measure of how effective that explanation is in persuading peers to change their own answer. Instructors and students could benefit from analytics with respect to the most convincing explanations on a list ranked long such a score.

Peer-review platforms most often ask students to provide a score based on a rubric, but the difficulty lies in the ability of novices to generate useful feedback before they have gained expertise with the content. A growing number of peer-review platforms address this issue with pairwise *comparative* judgments. Notable examples include ComPAIR[19] and JuxtaPeer[1], both of which present students with a pair of their peers' submissions, and evaluate them with respect to one another. TMPI falls in this category as well, as students apply a comparative judgment of their chosen explanation, relative only to the subset that was shown to them on their review page. However, one of the challenges that arises in these contexts is how to construct the subset of peer-items in the database which will be presented to the current student.

This opens the door to our two central research questions for this study:

RQ1 Since each student's "vote" in this context represents an incomplete evaluative judgement, which *rank aggregation* methods are best suited for TMPI?

LAK '21, April 2021, UCI,CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.000/0000.0000

RQ2 How can these aggregate *convincingness* scores be used to select the subset of explanations that are shown to each student in TMPI, so as to promote deeper reflection during the review step?

To our knowledge, we are among the first to examine rank aggregation methods as applied to these student "votes" in TMPI, in order to propose meaningful and reliable measurements of *convincingness.*

We suggest that the results of our work can inform the design of TMPI platforms, in how feedback is generated by teachers, and presented to students. In a broader context, we aim to contribute the growing body of research surrounding peer-review contexts, specifically where students do not provide holistic scores, but generate their evaluative judgments in a comparative setting.

2 RELATED WORK

2.1 Learnersourcing student explanations

This modality is a specific case of *learnersourcing*[21], wherein students first generate content as part of their own learning process, that is ultimately used to help their peers learn as well. Notable examples include PeerWise [8] and RiPPLE [13], both of which have student generate learning resources, which are subsequently used and evaluated by peers as part of formative assessment activities.

One of the earliest efforts to leverage peer judgments of peer-written explanations specifically is from the AXIS system[22], wherein students solved a problem, provided an explanation for their answer, and evaluated explanations written by their peers. Using a reinforcement-learning approach known as "multi-armed bandits", the system was able to select peer-written explanations that were rated as helpful as those written by an expert. Our research follows from these studies in scaling to multiple domains, and focusing on how the vote data can be used more directly to model argument quality as judged by peers.

2.2 Ranking Arguments for Quality

Rank aggregation is the task of combining the preferences of multiple agents into a single representative ranked list. It has long been understood that obtaining pairwise preference data may be less prone to error on the part of the annotator, as it is a simpler task than rating on scales with more gradations. (This is relevant in TMPI, since each student is choosing one explanation as the most convincing in relation to the subset of others that are shown.)

A classical approach for rank aggregation from pairwise preference data is using the Bradley-Terry model, which has been extended to incorporate the quality of contributions of different annotators in a crowdsourced setting when evaluating relative reading level in a pair passages [4].

When evaluating argument convincingness, one of the first approaches proposed is based on constructing an "argument graph", where a weighted edge is drawn from node A to node B for every pair where argument A is labelled as more convincing than argument B. After filtering example pairs that lead to cycles in the graph, PageRank scores are derived from this directed acyclic graph, and the PageRank scores of each argument are used as the gold-standard to rank for convincingness [11].

More recently, a relatively simpler heuristic Win-Rate score has been shown to be competitive alternative, wherein the rank score of an argument is simply the (normalized) number of times that argument has been chosen as more convincing in a pair, divided by the number of pairs it appears in [18].

Finally, a neural approach based on RankNet has recently yielded state of the art results. By joining two Bidirectional Long-Short-Term Memory Networks in a Siamese architecture, and appending a softmax layer to the output, [10] show that we can jointly model pairwise preferences and overall ranks publicly available datasets.

We will explore two of these options as part of our methodology in our rank aggregation step: the probabilistic Bradley-Terry model, and the simple heuristic scoring model. (We leave the neural approach for future work, as the additional work required to address make the models interpretable enough for the educational context is out of the scope of this study)

3 METHODOLOGY

We borrow our methodological approach from research in argument mining (AM), specifically related to modelling argument quality along the dimension of *convincingness.* A common approach is to curate pairs of arguments made in defence of the same stance on the same topic. These pairs are then presented to crowd-workers, who label which of the two is more convincing. These pairwise comparisons can then be aggregated using rank-aggregation methods so as to produce a overall ranked list of arguments. We extend this work to the domain of TMPI, and define prediction tasks that not only aim to validate this methodology, but help answer our specific research questions.

3.1 Rank Aggregation

The raw data emerging from a TMPI platform is tabular, in the form of student-item observations. The fields include the item prompt, the student's *first* answer choice, their accompanying explanation, the peer explanations shown on the review step, the student's *second* answer choice, and the peer explanation they chose as most convincing (`None` if they choose to "stick to their own").

(1) **win-rate**, defined as the ratio of times an explanation is chosen to the number of times it was shown (we add a correction for rationales that were never shown);
(2) **BT** score, which is the argument "quality" parameter estimated for each explanation, according to the *Bradley-Terry* model, where the probability of argument A being chosen over argument B is given by

$$P(a > b) = \frac{e^{\beta_a}}{e^{\beta_a} + e^{\beta_b}}$$

where β_i is the latent strength parameter of argument i.
We decompose each student-item observation into argument pairs, where the chosen explanation is paired with each of the other shown ones, and the pair is labelled with $y = -/+1$, depending on whether the chosen explanation is first/second in the pair. The latent argument strength parameters are estimated by maximizing the log-likelihood of the

m explanation-pairs:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{m}[w_{ij}ln\beta_i - w_{ij}ln(\beta_i + \beta_j)]$$

subject to $\sum_i \beta_i = 0$.

In order to evaluate these rank aggregation different scores, and address **RQ1**, we employ a time-series based cross-validation scheme: at each timestep, we calculate the aggregated argument *convincingness* scores from past students, and set out to predict:

(1) which arguments will be chosen as more convincing from the pairs constructed for the current student?
(2) will the current student choose a peer's explanation as more convincing than their own, or not?

For the latter binary classification task, we also include *surface level* features as a baseline, including the word count of explanation that the current student wrote; word counts of explanations that were shown on the review step; the number of explanations shown that were much shorter, or much longer than than the student's own explanation; and finally whether the student's first answer is correct.

Once these *surface level* and *convincingness* features are calculated for each student explanation, related statistics are assembled as well with respect to the *shown* explanations at each time step (e.g. maximum, minimum and mean **win rate** and **BT** scores of shown explanations, etc).

We build interpretable predictive models for this task, and inspect the relative feature importances in order to address our research questions from above.

We experiment with classification models that are interpretable, but limit our selection to the most linear (linear regression), to the most acceptably non-linear (Random Forests).

3.2 Modelling Argument Quality Scores

The goal **RQ1** is establish which rank aggregation methods are best suited for the context of TMPI, such that one can take the comparative preference data from many students who each see different subsets of peer explanations. We build on the results from the previous section to now predict these aggregate scores for each explanation, using linguistic properties of the explanations. We address **RQ2** with a regression task of predicting the argument *convincingness* scores using two different approaches to representing the student text: as an embedding inside a vector space models, or via a feature-rich document vector.

We experiment with vector space models with different document representations:

(1) LSA vectors (10,50,100 components) [7]
(2) Glove embeddings [16]
(3) BERT embeddings [9], out-of-the-box, and fine-tuned for the current classification task

The advantage of a feature-rich approach lies in the interpretability for teachers in their reporting tools, as well as generalizability to new items before vote data can be collected. The list of features included here are derived from related work in argument mining [11][17]on student essays, automatic short answer scoring [14]

	a	s	q	$\overline{a/s}$ (SD)	$\overline{wc}(SD)$	d	Δ
Bio	50667	1042	176	48 (45)	21 (49)	0.72	0.47
Chem	49905	2629	231	19 (17)	22 (21)	0.70	0.47
Phys	100873	4117	292	24 (25)	21 (27)	0.77	0.55

Table 1: Summary statistics of data, aggregated by discipline. The columns are a=number of answers, s=number of students, q=number of items, $\overline{a/s}$=mean number of answers completed by each student (with standard deviation), d=question difficulty, as defined by overall success rate of choosing correct answer choice on first attempt, and Δ=the fraction of answers where students chose an explanation other than their own on the review step.

- Surface Features: word count, sentence count, max/mean word length, max/mean sentence length;
- Lexical: uni-grams & bigrams, type-token ratio, number of keywords (defined by open-source discipline specific textbook), number of equations;
- Syntactic: POS n-grams (e.g. *nouns, prepositions, verbs,conjunctions,negation, adjectives, adverbs, punctuation*), modal verbs (e.g. *must, should, can, might*), contextuality/formality measure [12], dependency tree depth;
- Semantic: using LSA vectors trained on domain specific corpora, in this case an open-source textbook in the discipline, we calculate similarity to all other explanations in LSA space;
- Co-reference [17]: fraction of entities from the prompt mentioned in each sentence, averaged over all sentences (using neural Co-reference resolution) vector cosine similarity between student explanation and prompt, and answer choices;
- Readability: Fleish-Kincaid, Coleman-Liau, spelling errors

Features typical to NLP analyses in Learning Analytics that are not included here are cohesion, sentiment, and psycholinguistic features.

4 DATA

The data for this study come from myDALITE.org, which is a hosted instance of an open-source project, `dalite`[1], maintained by a Canadian researcher-practitioner partnership focused on supporting teachers develpping active learning pedagogy SALTISE.

Table 1 gives an overview of the dataset included in this study. The data is from introductory level university science courses, and generally spans different teachers at different colleges and universities in Canada.

5 RESULTS

TO DO

6 DISCUSSION

TO DO

[1]https://github.com/SALTISES4/dalite-ng

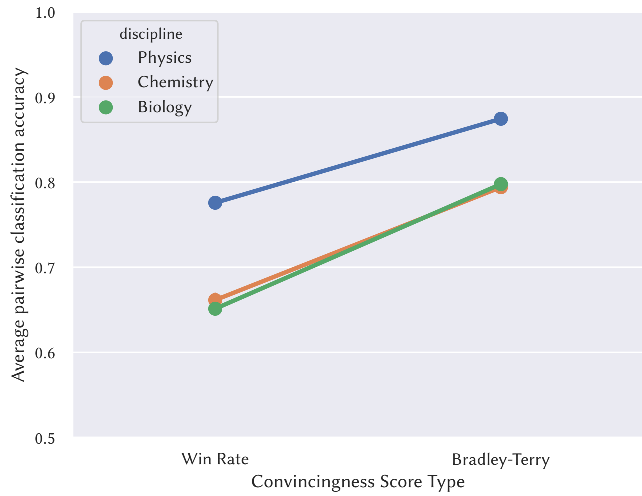

Figure 1: Comparing the classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair, under a time-series cross-validation scheme. Data is averaged across all questions and times steps, for different disciplinary subsets of TMPI data.

discipline	n	acc_LR	acc_RF
Biology	42408	0.67 (0.07)	0.67 (0.07)
Chemistry	41943	0.69 (0.06)	0.69 (0.07)
Physics	72160	0.72 (0.10)	0.73 (0.10)

Table 2: Average accuracy and F1 scores for models which aim to predict whether a student will choose a peer's explanation over their own, under a time-series cross validation scheme

7 LIMITATIONS AND FUTURE WORK

In many teaching contexts, however, teachers do not have the time to provide feedback to every student explanation for every question item. The feedback students receive is primarily based on the correctness of their first and second answer choices, not the *explanations* they write and choose.

The data at hand in TMPI environments enables scaling up how much feedback that can given.

TO DO

(1) Students are not explicitly directed on how to evaluate their peers' explanations. This may have an impact https://link.springer.com/article/10.1007/s10734-017-0220-3

ACKNOWLEDGMENTS

Funding for this research was made possible by the support of the Canadian Social Sciences and Humanities Research Council *Insight* Grant. This project would not have been possible without the

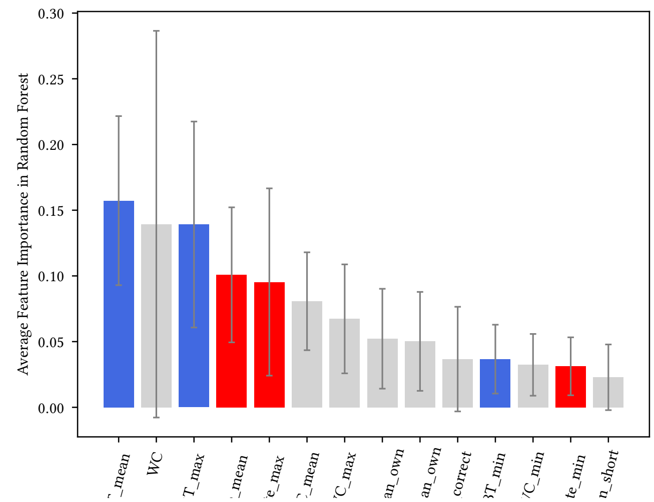

Figure 2: Average relative feature importances when predicting whether a student will choose a peer's explanation over their own (or not).

SALTISE/S4 network of researcher practitioners, and the students using myDALITE.org who consented to share their learning traces with the research community.

REFERENCES

[1] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–13. https://doi.org/10.1145/3173574.3173868

[2] Elizabeth S. Charles, Nathaniel Lasry, Sameer Bhatnagar, Rhys Adams, Kevin Lenton, Yann Brouillette, Michael Dugdale, Chris Whittaker, and Phoebe Jackson. 2019. Harnessing peer instruction in- and out- of class with myDALITE. In *Fifteenth Conference on Education and Training in Optics and Photonics: ETOP 2019*. Optical Society of America, 11143_89. http://www.osapublishing.org/abstract.cfm?URI=ETOP-2019-11143_89

[3] Elizabeth S. Charles, Nathaniel Lasry, Chris Whittaker, Michael Dugdale, Kevin Lenton, Sameer Bhatnagar, and Jonathan Guillemette. 2015. Beyond and Within Classroom Walls: Designing Principled Pedagogical Tools for Student and Faculty Uptake. International Society of the Learning Sciences, Inc.[ISLS].

[4] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

[5] Michelene TH Chi, Nicholas Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science* 18, 3 (1994), 439–477.

[6] Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics* 69, 9 (2001), 970–977.

[7] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS* 41, 6 (1990), 391–407.

[8] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. 2008. PeerWise: Students Sharing Their Multiple Choice Questions. In *Proceedings of the Fourth International Workshop on Computing Education Research (ICER '08)*. ACM, New York, NY, USA, 51–58. https://doi.org/10.1145/1404520.1404526 event-place: Sydney, Australia.

[9] 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).

[10] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. *arXiv preprint arXiv:1907.08971*

(2019). https://www.aclweb.org/anthology/P19-1093/

[11] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1589–1599. https://doi.org/10.18653/v1/P16-1150

[12] Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of science* 7, 3 (2002), 293–340. Publisher: Springer.

[13] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. 2019. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522* (2019).

[14] Michael Mohler and Rada Mihalcea. 2009. Text-to-text Semantic Similarity for Automatic Short Answer Grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 567–575. http://dl.acm.org/citation.cfm?id=1609067.1609130 event-place: Athens, Greece.

[15] Mitchell J Nathan, Kenneth R Koedinger, Martha W Alibali, and others. 2001. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the third international conference on cognitive science*, Vol. 644648.

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[17] Isaac Persing and Vincent Ng. 2016. End-to-End Argumentation Mining in Student Essays.. In *HLT-NAACL*. 1384–1394.

[18] Peter Potash, Adam Ferguson, and Timothy J Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*. 146–155.

[19] Tiffany Potter, Letitia Englund, James Charbonneau, Mark Thomson MacLean, Jonathan Newell, Ido Roll, and others. 2017. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry* 5, 2 (2017), 89–113.

[20] Teaching & Learning Technologies Univeristy of British Columbia. 2019. ubc/ubcpi. https://github.com/ubc/ubcpi original-date: 2015-02-17T21:37:02Z.

[21] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 405–416.

[22] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, Edinburgh, Scotland, UK, 379–388. https://doi.org/10.1145/2876034.2876042