Modelling Argument Quality in Technology Mediated Peer Instruction

Sameer Bhatnagar
Polytechnique Montreal

Michel C. Desmarais
Polytechnique Montreal

Amal Zouaq
Polytechnique Montreal

TO DO

Keywords:

---

1. INTRODUCTION

Technology-mediated peer instruction *(TMPI)* platforms (Charles et al., 2019)(Univeristy of British Columbia, 2019) 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, as shown in figure 1a.

On the second step (figure 1b), students are prompted to revise their answer choice, by taking into consideration a subset of explanations written by their peers.

The student now has three options:

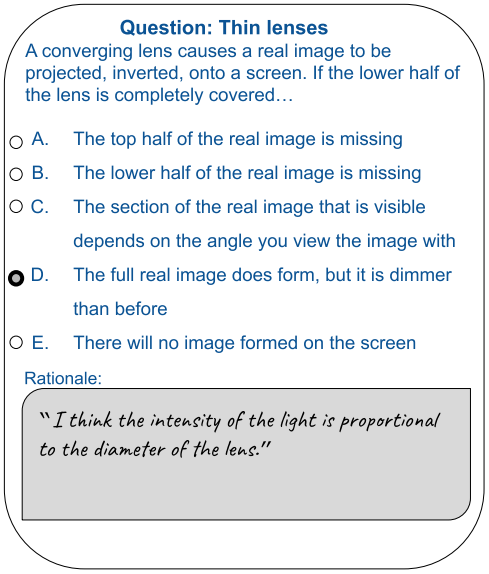
1. Change their answer choice, by indicating which of their peer's explanations for a *different* answer choice was most convincing;

2. keep the *same* answer choice, but indicate which the peer's explanations the student found more convincing than their own;

3. choose "I stick to my own", which indicates that they are keeping to the same answer choice, and that their own explanation is best from among those that are shown.

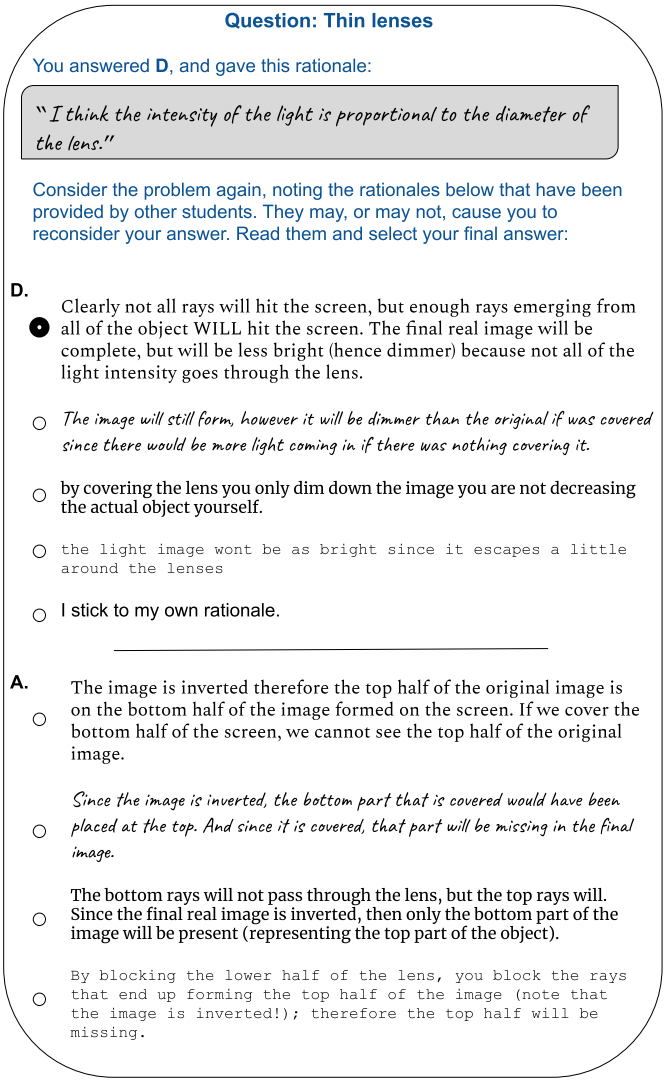
Whenever the student goes with either of the first two scenarios above, we frame this as "casting a vote" for the chosen peer explanation.

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 (Chi et al., 1994). Deliberate practice of argumentation in defence of one's ideas has been shown to improve informal reasoning for science students(Venville and Dawson, 2010). There exists empirical evidence on the positive relationship between constructing formally sound arguments and deep cognitive elaboration, as well as individual acquisition of knowledge(Stegmann et al., 2012).



(a) The first step in TMPI, where a student is presented with a multiple choice item. The student must enter a "rationale", or "explanation" justifying their answer choice.

The panel on the right, figure 1b, shows the second, review step of TMPI. Before any feedback is given on the correctness of their first attempt, the student is prompted to reconsider their answer choice, by reading a subset of explanations written by previous students. A set of peer-explanations is shown for the student's own answer choice, and another set is shown for a different answer choice.



(b)

Figure 1: The two steps in technology-mediated peer instruction (TMPI)

Second, classroom based *Peer Instruction* (Crouch and Mazur, 2001), 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 (Charles et al., 2015). 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" (Nathan et al., 2001), addressing student misconceptions they might not otherwise have thought of.

We situate student explanations from TMPI, in the context of computational argumentation, a sub-field of NLP focused on identifying argumentative components, and in their links to one

another. Modelling argument "quality" is an area of active research, with direct applications in education, such as in automated scoring of persuasive essays written by students (Persing and Ng, 2015) (Nguyen and Litman, 2018). When students are coached asked to debate in dyads, and prompted to either find consensus, or to persuade peers, there is correlation to the quality of arguments students produce, as measured by the presence of formal argumentative structures (e.g. claims, premise, etc.) (Garcia-Mila et al., 2013).

However experiments have also shown that the perceived quality of an argument depends on the audience, and so we follow a more pragmatic measure of argument quality, centred on the premise that the goal of argumentation is persuasion (Mercier and Sperber, 2011). Therefore we suggest that the "vote" data collected for each student's explanation, is a proxy for argument quality, along the dimension of *convincingness*, as judged by peer learners. This is a direct application of the argument mining task originally proposed by (Habernal and Gurevych, 2016): if crowd-workers are presented with a pair of arguments for the same stance of a debatable topic, can we predict which of the two they will choose as more convincing? This task has been extended to TMPI, wherein the objective is to predict which explanations students will choose as more convincing than their own (Bhatnagar et al., 2020).

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. Student explanations can then be ranked along such a score, allowing for instructors to gain insights on the thinking of their students with respect to specific content, and potentially even help students to improve how they communicate ideas within their discipline.

The problem of aggregating the results of evaluative peer-judgments extends beyond TMPI. For example, in response to the difficulty students can have providing a holistic score to their peers' work, there is a growing number of peer-review platforms built on *comparative* judgments. Notable examples include ComPAIR (Potter et al., 2017) and JuxtaPeer (Cambre et al., 2018), both of which present students with a just a pair of their peers' submissions, and prompt the learner to evaluate them with respect to one another. As in TMPI, students apply a comparative judgment to only the subset of peer content that they are shown during the review step. There is a need for a principled approach to aggregating this learnersourced data, in a pedagogically relevant manner, despite the inevitable absence of some "true" ranking.

This sets the stage for our central research questions:

RQ1 since each student's "vote" in this context represents an incomplete evaluative judgement, which rank aggregation methods are best suited for ranking the quality of student explanations in TMPI?

RQ2 once we establish a ranked list of explanations along the dimension of *convinciness*, can we model this construct, and identify the linguistic features of the most effective student explanations, as judged by their peers?

Work on modelling *convincingness* has, in large part, been centred on web discourse data. In the educational setting, previous work in automated scoring of persuasive essays has focused on modelling holistic scores given by *experts* on longer form essays. To our knowledge, we are among the first to aggregate and model student "votes", in order to evaluate student explanations for their *convincingness* as judged by *peers*.

We suggest that the results of our work can inform the design of TMPI platforms. However, in a broader context, we aim to contribute to the growing body of research surrounding

technology-mediated peer-review, specifically where learners do not provide holistic scores, but generate their evaluative judgments in a comparative setting. Such platforms will invariably have to deal with at least three issues, which our work helps to address.

The first issue is about students: providing feedback to learners on the characteristics common to the most convincing arguments in their discipline, promotes learning and the development of critical reasoning skills. The second issue is in providing support to teachers: in such platforms, the amount of data generated scales very quickly. The data associated with each student-item pair includes many relevant variables: correct answer choice on first attempt, student explanation, subset of explanations shown, time spent writing and reading explanations, correct answer on second attempt, and the peer-explanation chosen as most convincing (see figure 2). This amount of information can be overwhelming for instructors who use such tools regularly as part of formative assessment. Automatically identifying the highest, and lowest, quality student explanations, as judged by other students, can support instructors in providing timely feedback. A third related issue is in maintaining the integrity of such platforms: automatic filtering of irrelevant/malicious student explanations is paramount, since they may be shown to future students (Gagnon et al., 2019), a non-trivial task for natural language content, without expensive expert moderation.

This paper begins with an overview of research related to learnersourcing, and argument mining (section 2). We then describe our TMPI dataset, as well as publicly available reference datasets of argument quality, which we use to evaluate our methodology (section 3). Our most important contribution is in proposing a methodology for evaluating the quality of student explanations, along the dimension of *convincingness*, in TMPI environments; we demonstrate this methodology in section 4. We then present our results on choosing the appropriate *measure* of explanation convincingness (section 5), and finally, we describe how we *model* these convincingness "scores" so as to identify the linguistic features of explanations most often associated with high-quality explanations (section 6).

2. Related Work

2.1. Automated Essay Scoring

TO DO taxonomy of argument quality (Wachsmuth et al., 2017)
(Persing and Ng, 2015)
(Nguyen and Litman, 2018)

2.2. Learnersourcing student explanations

TMPI is a specific case of *learnersourcing*(Weir et al., 2015), wherein students first generate content, and then help curate the content base, all as part of their own learning process. Notable examples include PeerWise (Denny et al., 2008) and RiPPLE (Khosravi et al., 2019), 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(Williams et al., 2016), 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.

2.3. 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 only in relation to the subset of others that are shown.)

A classical approach for aggregating pairwise preference data into a ranked list is using the Bradley-Terry model (Bradley and Terry, 1952). This 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 (Chen et al., 2013).

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 (Habernal and Gurevych, 2016).

More recently, a relatively simpler heuristic WinRate score has been shown to be a 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 (Potash et al., 2019).

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, pairwise preferences and overall ranks were jointly modelled in publicly available datasets (Gleize et al., 2019).

We will explore two of these options as part of our methodology in our rank aggregation step, via several related methods: the probabilistic Bradley-Terry model, as well as two of its variants (CrowdBT and the Elo rating system), and the simple heuristic scoring model. (We omit a neural approach, as model interpretability is key in the educational context,)

3. Data

3.1. Argument Mining Datasets

Much of our methodology is inspired by work on modelling argument quality along the dimension of *convincingness*, as described in section 2. In order to contextualize the performance of these methods in our educationale setting, we apply the same methods to publicly available datasets from the argument mining research community as well, and present the results. These datasets are described in table 1, alongside the TMPI data at the heart of our study. The **UKP** dataset(Habernal and Gurevych, 2016) is the first set of labelled argument pairs to be released publicly. Crowd-workers were presented with pairs of arguments on the same stance of a debate prompt, and were asked to choose which was more convincing. The authors of the **IBM_ArgQ** dataset(Toledo et al., 2019) offer a similarly labelled, but much more tightly curated dataset, with strict controls on argument word count and relative difference in lengths in each pair. This was

source	dataset	topics	args	pairs	args/topic	pairs/topic	pairs/arg	wc
Arg Mining	IBM_ArgQ	22	3474	9125	158 (144)	415 (333)	5 (1)	24 (1)
	IBM_Evi	41	1513	5274	37 (14)	129 (69)	7 (3)	30 (3)
	UKP	32	1052	11650	33 (3)	364 (71)	22 (3)	49 (14)
DALITE	Chemistry	36	4778	38742	133 (29)	1076 (313)	7 (1)	29 (6)
	Ethics	28	20195	159379	721 (492)	5692 (4962)	7 (1)	48 (8)
	Physics	76	10840	96337	143 (42)	1268 (517)	7 (2)	27 (5)

Table 1: Summary statistics for reference datasets from argument mining research community, and DALITE, a TMPI environment used mostly in undergraduate science courses in Canada. In the argument reference datasets *topic* are debate prompts shown to crowdsourcing workers (e.g. *"social media does more good than harm"*), while a *topic* in DALITE is a question item. The explanations given by students are analagous to the "arguments", which are then assembled into pairs based on what was shown, and eventually chosen by each student. *wc* is the average number of tokens in each argument/explanation in each topic. All averaged quantities are followed by a standard deviation in parentheses.

partly in response to the observation that crowd labels could often be predicted simply by choosing the longer text from the pair. The labelled argument pairs in the **IBM_Evi** dataset (Gleize et al., 2019) are actually generated by scraping Wikipedia, and the crowd workers were asked to choose the argument from the pair that provided the more compelling evidence in support of the debate stance.

As described above in our section on related work, these datasets were released not just with the labelled argument pairs, but holistic rank scores for each argument, that were each derived in different ways. We will be comparing our proposed *measures* of convincingness to these rank scores.

3.2. DALITE

The central 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 in the development of active learning pedagogy SALTISE. The data comes from introductory level university science courses, and generally spans different teachers at different colleges and universities in Canada. The *Ethics* dataset comes from a popular MOOC, wherein the TMPI prompts are slightly different from the *Physics* and *Chemistry* prompts, in that there is no "correct" answer choice, and that the goal is to have students choose a side of an argument, and justify their choice. Table 1 gives an overview of the datasets included in this study.

To stay consistent with the argument mining reference dataset terminology, we refer to a question-item as a "topic". Student explanations from DALITE are divided up by the associated question item prompts. The transformation of TMPI student explanations ("args") into "pairs" is described in section 4. The filtering of DALITE data is based on the following three steps:

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

1. Many question items have been completed by several hundreds of students. As such, almost half of student explanations have only been shown by another peer. As such, we retain only those student answers that have been shown to at least 5 other students.

2. As a platform for formative assessment, not all instructors provide credit for the explanations students write. As such, there are some students who do not put much effort into writing good explanations. We filter out only those student answers that have at least 10 words.

3. after the previous two steps, we include data only from those questions that have at least 100 remaining student answers.

4. 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, whose task it is to 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.

4.1. RANK AGGREGATION

The raw data emerging from a TMPI platform is tabular, in the form of student-item observations. As shown in figure 2(a), 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"), as well as timestamps for the first and second attempt.

It should be noted that there is no credit associated with which explanation is chosen in DALITE (all points are attributed based on the correctness of the answer choice on the first and second steps). After carefully looking at timestamp data, we observe that a large fraction of students who choose to "stick to their own", spend as little as 5 seconds on the review step. For this reason, we exclude these students' data, and build all rank scores only based on students who explicitly choose a peer's explanation over their own.

After this first filtering step, we take the TMPI observations for each question, and construct explanation pairs, as in figure 2(b).

1. **WinRate**, defined as the ratio of times an explanation is chosen to the number of times it was 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

$$P(a > b) = \frac{1}{1 + e^{\beta_b - \beta_a}}$$

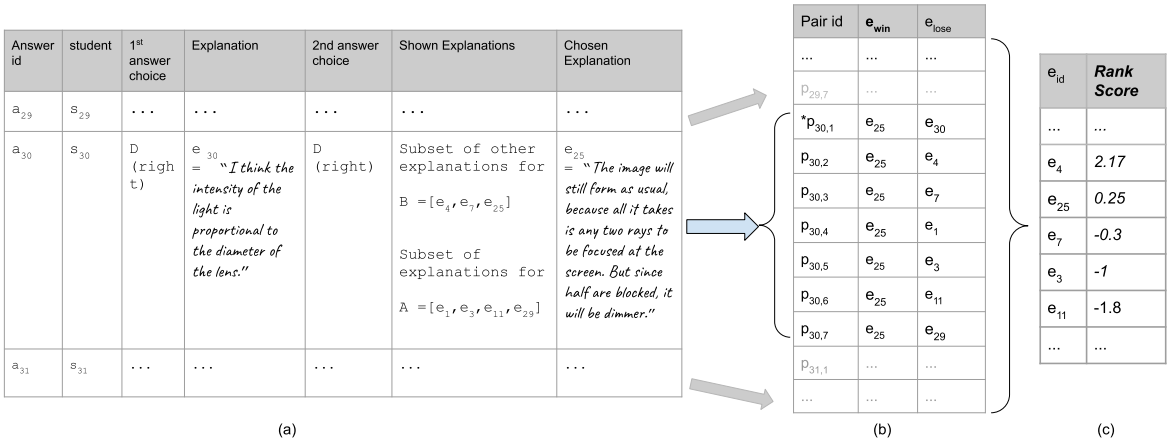


Figure 2: Example of student-item observations from a TMPI environment. This figure follows from figure 1. (a) Student s_{30} chose the correct **D** as the answer on their first attempt, and provided the explanation e_{30} in the dataset for this question. Without providing feedback on whether this is the correct answer, the student is shown a subset of explanations from previous students for **D**, as well as for **A** (the most popular incorrect answer). The student decides to keep the same answer choice **D**, and indicates that the explanation e_{25} is the most convincing. This is referred to as a *Right→Right* transition. (b) This observation is transformed into 7 explanation pairs. The first pair is for the choice of e_{25} over what the student themself, and the other six are for the choice of e_{25} over the other shown explanations. The pairs are labelled as either having the left or right exlanation being more *convincing*. (c) This pairwise preference data is aggregated global ranked list, where each explanation is assigned a Rank Score.

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. Assuming there are N explanations, labelled by K students, and S_K labelled pairs,the latent strength parameters are estimated by maximizing the log-likelihood given by:

$$\ell(\boldsymbol{\beta}) = \sum_{K} \sum_{(i,j)\epsilon S_K} \log \frac{1}{1 + e^{\beta_i - \beta_j}}$$

subject to $\sum_i \beta_i = 0$.

3. The **Elo** rating system(Elo, 1978), which was originally proposed for ranking chess players, has been successfully used in adaptive learning environments (see (Pelánek, 2016) for a review). This rating method can be seen as a heuristic re-parametrization of the **BT** method above, where the probability of argument A being chosen over argument B is given by

$$P(a > b) = P_{ab} = \frac{1}{1 + 10^{(\beta_b - \beta_a)/400}}$$

All arguments are initialized with an initial value of 1500 points, an the rating of any argument is only updated after it appears in a pairwise comparison with another. The rating update rule transfers points from the winner, to the loser, in proportion to the difference in strength:

$$\beta'_a := \beta_a + K(P_{ab} - \beta_a)$$

While the **BT** model can be thought of a *consensus* approach, **Elo** ratings are dynamic and implicitly give more weight to recent data(Aldous, 2017).

4. **Crowd-BT** (Chen et al., 2013) is an extension of the **BT** model, tailored to settings where different annotators may have assigned opposite labels to the same pairs, and the reliability of each annotator may vary significantly. A reliability parameter is estimated for each student,

$$\eta_k \equiv P(a >_k b | a > b)$$

where $\eta_k \approx 1$ if the student k agrees with most other students, and $\eta_k \approx 0$ if the student is in opposition to their peers. This changes the model of argument a being chosen over b by student k to

$$P(a >_k b) = \eta_k \frac{e^{\beta_a}}{e^{\beta_a} + e^{\beta_b}} + (1 - \eta_k) \frac{e^{\beta_b}}{e^{\beta_a} + e^{\beta_b}}$$

and the log-likelilood maximized for estimation to

$$\ell(\boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{K} \sum_{(i,j)\epsilon S_K} \log \left[\eta_k \frac{e^{\beta_a}}{e^{\beta_a} + e^{\beta_b}} + (1 - \eta_k) \frac{e^{\beta_b}}{e^{\beta_a} + e^{\beta_b}} \right]$$

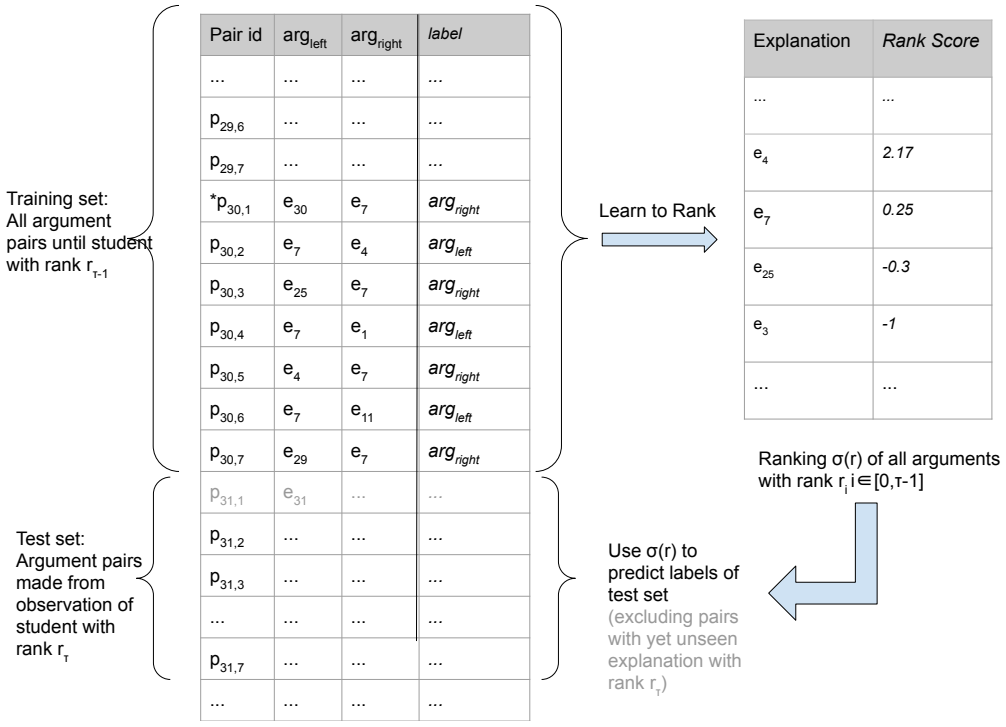


Figure 3: All of the pairs constructed for answers by students $i <= 30$ are used to learn rank scores σ, which are evaluated in their ability to predict the label in the held out test set: argument pairs constructed from the answer of student $i = 31$. We must exclude those pairs which include the 31st student, as this explanation is unseen in the training set, and does not have a rank score

In order to evaluate these rank aggregation different scores, and address our research question, 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: which arguments will be chosen as more convincing from the pairs constructed for the current student?

5. Measuring Argument Quality

TO DO

6. Modelling Argument Quality Scores

TO DO

Cross-topic validation scheme; control for word-count

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 those explanations

We address **RQ2** with a regression task of predicting the argument *convincingness* scores via a feature-rich document vector.

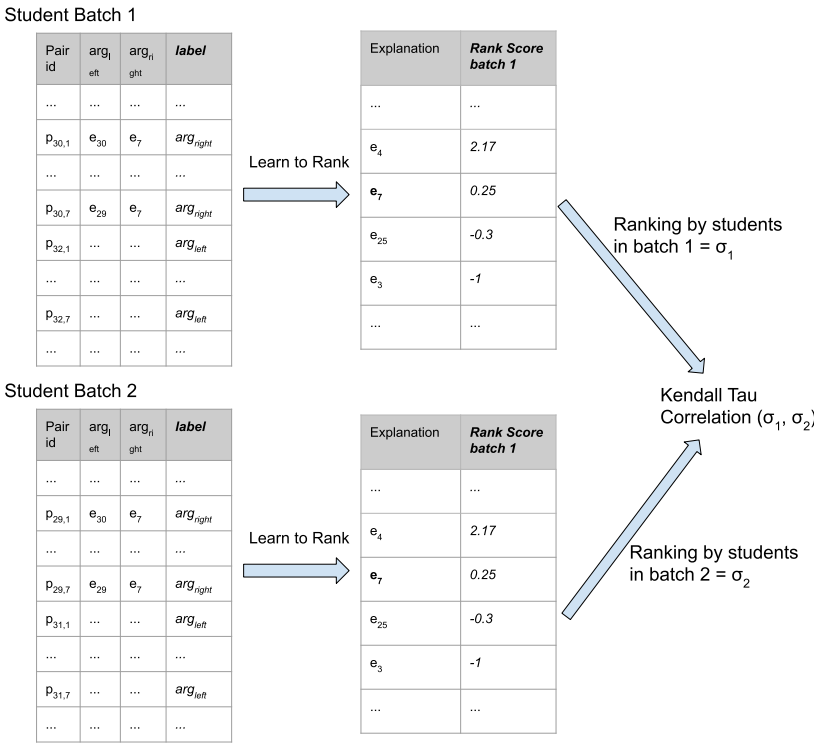


Figure 4: Validation scheme for evaluating reliability of rankings. At each time step τ we take all students $i <= \tau$, and split student into two batches (chosen at alternating time steps) We learn rankings for each of these batches, and evaluate the Kendall Tau rank correlation as estimate of reliability of the rankings.

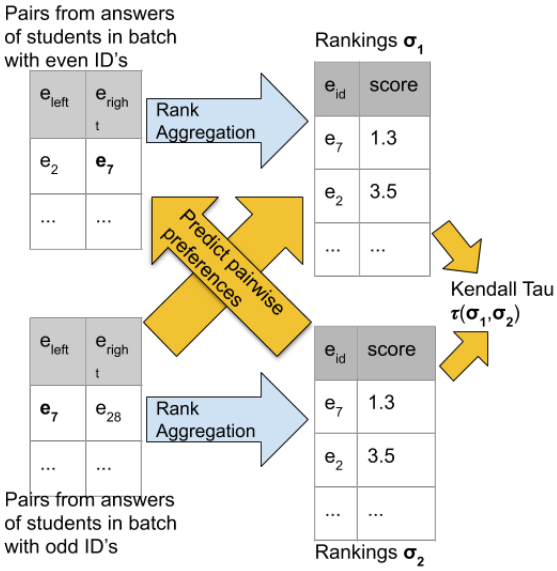


Figure 5: Methodology for evaluation of rank scores

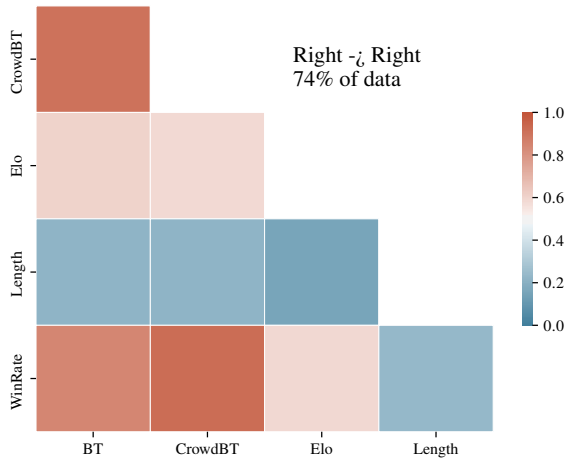


Figure 6: Correlation between different Ranking Scores for each explanation, disaggregated by transition type

Figure 7: Comparing the classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair. Rank scores are calculated with the vote data of half the students, and tested on the pairs generated by the other half. Data is averaged across all questions, dis-aggregated by different TMPI transition types.

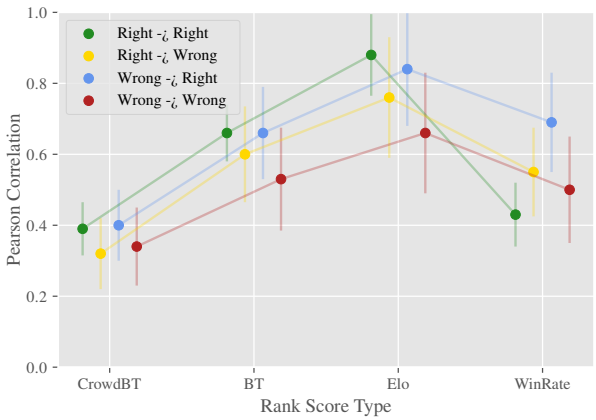


Figure 8: Pearson correlation coefficient between different rank score types, derived from two independent groups of students, averaged over all questions, dis-aggregated by different TMPI transition types.

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 (Habernal and Gurevych, 2016)(Persing and Ng, 2016)on student essays, automatic short answer scoring (Mohler and Mihalcea, 2009)

- 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 text-book), 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 (Heylighen and Dewaele, 2002), 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 (Persing and Ng, 2016): 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.

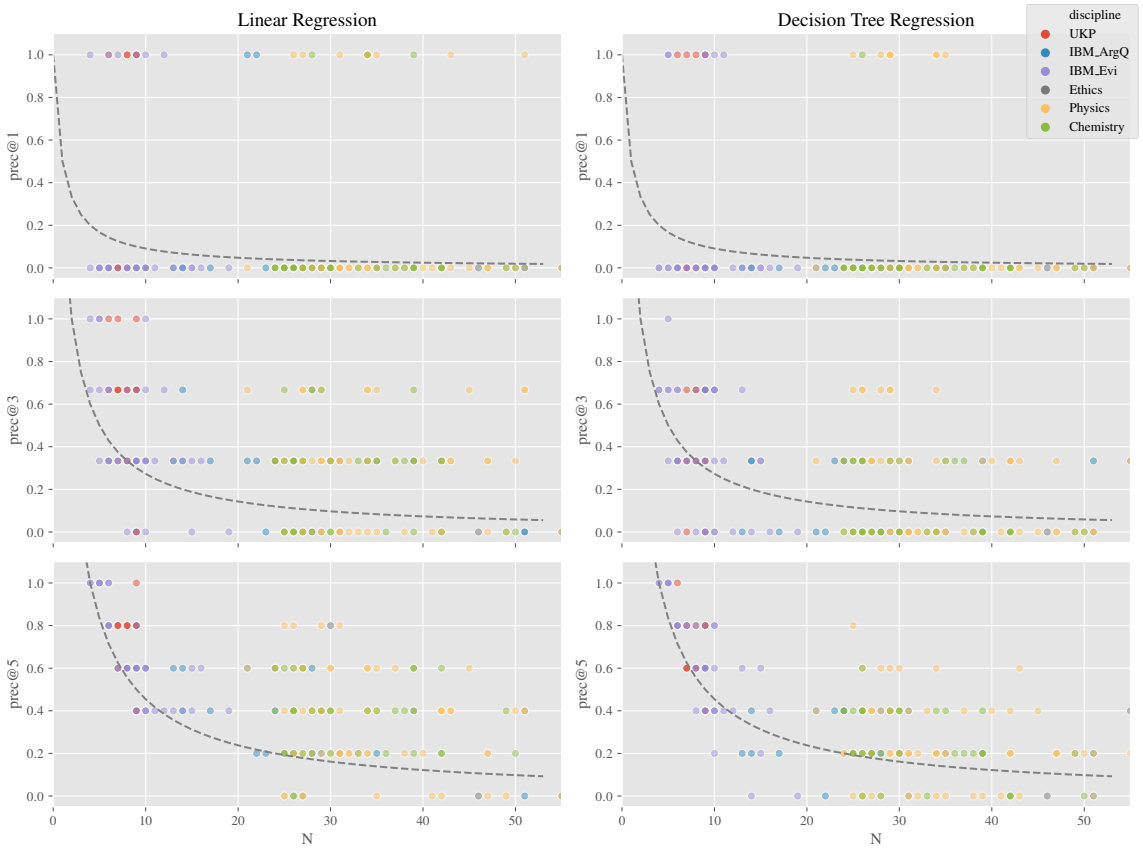


Figure 9: Evaluating of regression models tasked with predicting *convincingness* score of arguments/TMPI-explanations based on linguistic features. Evaluation metric is **precision @ K**, where we verify how many of the predicted top-K ranked explanations are in the measured top-K list (using "winrate" as a *measure* of convincingness). Precision is plotted against the size of the test-set on the horizontal axis, under a cross-topic validation scheme. The dashed line is a (harsh) approximation of the probability of choosing the top-K explantions purely by chance (K/N, which is much greater than "N choose K"). Each dot represents the performance on one held-out topic/TMPI-question-prompt, color-coded based on which dataset/discipline it originates from. We compare Linear Regression with a Decision Tree Regressor in the two columns.

REFERENCES

ALDOUS, D. 2017. Elo ratings and the sports model: A neglected topic in applied probability? *Statistical science 32,* 4, 616–629. Publisher: Institute of Mathematical Statistics.

BHATNAGAR, S., ZOUAQ, A., DESMARAIS, M. C., AND CHARLES, E. 2020. Learnersourcing Quality Assessment of Explanations for Peer Instruction. In *Addressing Global Challenges and Quality Education*, C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, and S. M. Dennerlein, Eds. Springer International Publishing, Cham, 144–157.

BRADLEY, R. A. AND TERRY, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika 39,* 3/4, 324–345. Publisher: JSTOR.

CAMBRE, J., KLEMMER, S., AND KULKARNI, C. 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.

CHARLES, E. S., LASRY, N., BHATNAGAR, S., ADAMS, R., LENTON, K., BROUILLETTE, Y., DUGDALE, M., WHITTAKER, C., AND JACKSON, P. 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.

CHARLES, E. S., LASRY, N., WHITTAKER, C., DUGDALE, M., LENTON, K., BHATNAGAR, S., AND GUILLEMETTE, J. 2015. Beyond and Within Classroom Walls: Designing Principled Pedagogical Tools for Student and Faculty Uptake. International Society of the Learning Sciences, Inc.[ISLS].

CHEN, X., BENNETT, P. N., COLLINS-THOMPSON, K., AND HORVITZ, E. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

CHI, M. T., LEEUW, N., CHIU, M.-H., AND LAVANCHER, C. 1994. Eliciting self-explanations improves understanding. *Cognitive science 18,* 3, 439–477.

CROUCH, C. H. AND MAZUR, E. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics 69,* 9, 970–977.

DENNY, P., HAMER, J., LUXTON-REILLY, A., AND PURCHASE, H. 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. event-place: Sydney, Australia.

ELO, A. E. 1978. *The rating of chessplayers, past and present*. Arco Pub.

GAGNON, V., LABRIE, A., DESMARAIS, M., AND BHATNAGAR, S. 2019. Filtering non-relevant short answers in peer learning applications. In *Proc. Conference on Educational Data Mining (EDM)*.

GARCIA-MILA, M., GILABERT, S., ERDURAN, S., AND FELTON, M. 2013. The effect of argumentative task goal on the quality of argumentative discourse. *Science Education 97,* 4, 497–523. Publisher: Wiley Online Library.

GLEIZE, M., SHNARCH, E., CHOSHEN, L., DANKIN, L., MOSHKOWICH, G., AHARONOV, R., AND SLONIM, N. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. *arXiv preprint arXiv:1907.08971*.

HABERNAL, I. AND GUREVYCH, I. 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.

HEYLIGHEN, F. AND DEWAELE, J.-M. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of science 7,* 3, 293–340. Publisher: Springer.

KHOSRAVI, H., KITTO, K., AND WILLIAMS, J. J. 2019. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522*.

MERCIER, H. AND SPERBER, D. 2011. Why do humans reason? Arguments for an argumentative theory.

MOHLER, M. AND MIHALCEA, R. 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. event-place: Athens, Greece.

NATHAN, M. J., KOEDINGER, K. R., ALIBALI, M. W., 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.

NGUYEN, H. V. AND LITMAN, D. J. 2018. Argument Mining for Improving the Automated Scoring of Persuasive Essays. In *AAAI*. Vol. 18. 5892–5899.

PELÁNEK, R. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education 98*, 169–179. Publisher: Elsevier.

PERSING, I. AND NG, V. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 543–552.

PERSING, I. AND NG, V. 2016. End-to-End Argumentation Mining in Student Essays. In *HLT-NAACL*. 1384–1394.

POTASH, P., FERGUSON, A., AND HAZEN, T. J. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*. 146–155.

POTTER, T., ENGLUND, L., CHARBONNEAU, J., MACLEAN, M. T., NEWELL, J., ROLL, I., AND OTHERS. 2017. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry 5,* 2, 89–113.

STEGMANN, K., WECKER, C., WEINBERGER, A., AND FISCHER, F. 2012. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science 40,* 2 (Mar.), 297–323.

TOLEDO, A., GRETZ, S., COHEN-KARLIK, E., FRIEDMAN, R., VENEZIAN, E., LAHAV, D., JACOVI, M., AHARONOV, R., AND SLONIM, N. 2019. Automatic Argument Quality Assessment-New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5629–5639.

UNIVERISTY OF BRITISH COLUMBIA, T. . L. T. 2019. ubc/ubcpi. original-date: 2015-02-17T21:37:02Z.

VENVILLE, G. J. AND DAWSON, V. M. 2010. The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. *Journal of Research in Science Teaching 47,* 8, 952–977. Publisher: Wiley Online Library.

WACHSMUTH, H., NADERI, N., HOU, Y., BILU, Y., PRABHAKARAN, V., THIJM, T. A., HIRST, G., AND STEIN, B. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 176–187.

WEIR, S., KIM, J., GAJOS, K. Z., AND MILLER, R. C. 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.

WILLIAMS, J. J., KIM, J., RAFFERTY, A., MALDONADO, S., GAJOS, K. Z., LASECKI, W. S., AND HEFFERNAN, N. 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.