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:

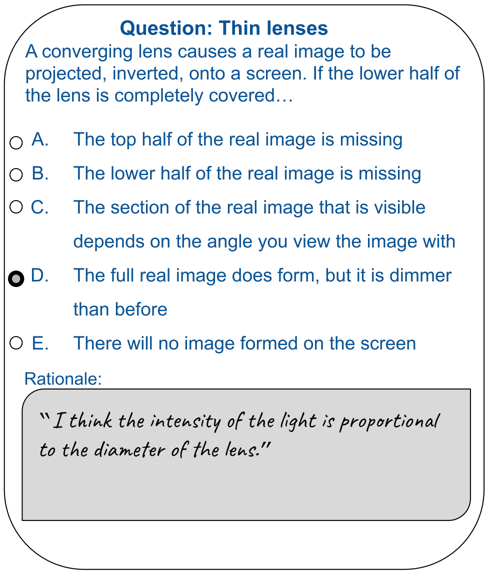
1. Change their answer choice, by indicating which of their peer's explanations for a *different* answer choice was most convincing;

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.

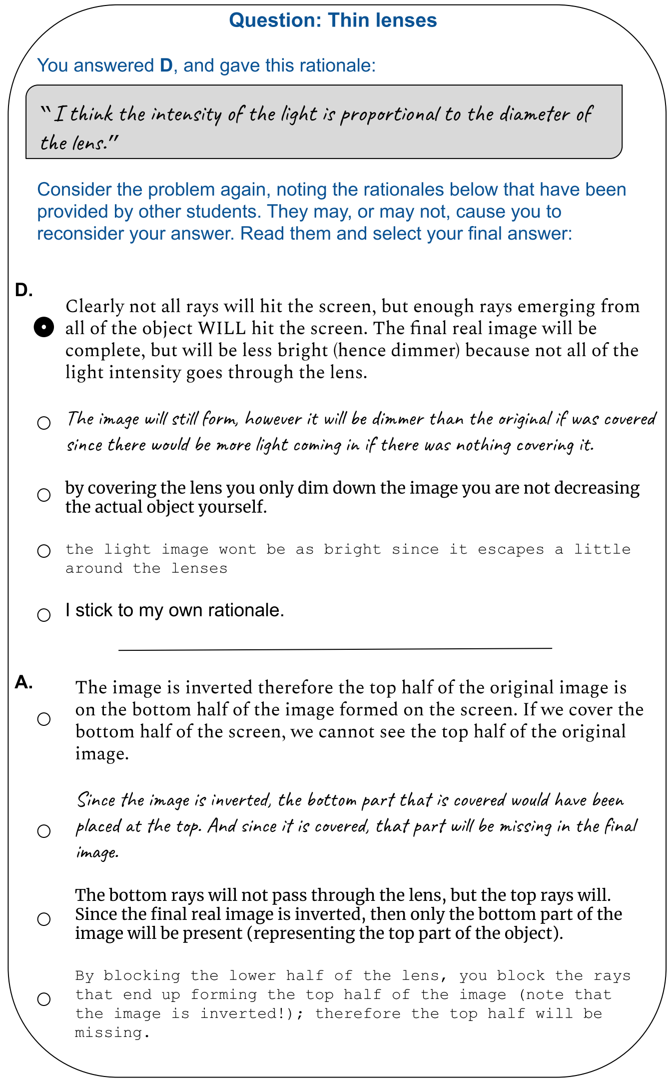
Whenever the student goes with either of the first two scenarios above, we frame this as "casting a vote" for the chosen peer explanation.

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 asked to debate in dyads, and prompted to either find consensus, or instead persuade their peers, there is a relationship between knowledge acquisition, and the quality of arguments the 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 can depend on the audience (Mercier and Sperber, 2011), and so we adopt a more pragmatic measure of argument quality, centred on the premise that the goal of argumentation is persuasion.

In a comprehensive survey of research on the assessment of argument quality, (Wachsmuth et al., 2017) outline a taxonomy of major quality dimensions for natural language, with three principal aspects: logic, rhetoric, and dialect. As students vote on their peer's explanations in TMPI, they may be evaluating the logical cogency (e.g. is this argument sound?), or its rhetorical quality (e.g. is this argument phrased well?). We focus our work on students who choose a peer's explanation *as more convincing than their own*, as there exists a significant bias for the option "I stick to my own".

Therefore, we suggest that the "vote" data collected for each student's explanation in TMPI, 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 already been extended to TMPI in previous work, wherein the objective is to predict which explanations students will choose as more convincing than their own (Bhatnagar et al., 2020).

Student votes in TMPI 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. However aggregating these votes should be done with care: when a student chooses an explanation as convincing, they are doing so only with respect to the subset that were shown, as well as the one they wrote themselves.

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 related research in learnersourcing of student explanations, automatic short-answer grading, and argument quality ranking (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 and propose evaluation metrics based on practical issues in TMPI environments. 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 5).

2. Related Work

2.1. 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 students 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. The scheme proposed by (Kolhe et al., 2016) applies the potential of learnersourcing to the task of short answer grading: the short answers submitted by students are evaluated by "future" peers who are presented with multiple choice questions, where the answer options are the short answers submitted by their "past" counterparts. 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. Automated Writing Evaluation

A central objective of our work is to evaluate the quality of student explanations in TMPI. Under the hierarchy of automated grading methods proposed by (Burrows et al., 2015), this task falls under the umbrella of automatic short-answer grading (ASAG); students must recall knowledge and express it in their own way, using natural language, usually using between 10-100 words. Their in-depth historical review of ASAG systems describe a shifting focus in methods, from matching patterns derived from answers written by experts, to machine-learning approaches, where n-grams and hand-crafted features are combined as input to supervised learning algorithms, such as decision trees and support vector machines.

For example, (Mohler et al., 2011) measures alignment between dependency parse tree structures of student answers, with those of an expert answer. These alignment features are paired with lexical semantic similarity features that are both knowledge-based (e.g. using WordNet) and corpus-based (e.g. Latent Semantic Analysis), and used as input to support vector machines which learn to automatically grade short answers.

Another similar system proposed by (Sultan et al., 2016) starts with features measuring lexical and contextual alignment between similar word pairs from student answers and a reference answer, as well as semantic vector similarity using "off-the-shelf" word embeddings. They then augment their input with "domain-specific" term-frequency and inverse document-frequency weights, to achieve their best results on several ASAG datasets using various validation schemes.

In addition to similarity features based on answer text, (Zhang et al., 2016) show that question-level (e.g. difficulty, expert-labelled knowledge components) and student-level features (e.g. pre-test scores, Bayesian Knowledge Tracing probability estimates) can improve performance on the ASAG task when input to a deep learning classifier.

While modelling the quality of TMPI explanations has much in common with the ASAG task, and can benefit from the features and methods from the systems mentioned above, a fundamental difference lies in how similarity to an expert explanation may not be the only appropriate reference. The "quality" we are measuring is that which is observed by a group of peers, which may be quite different from how a teacher might explain a concept.

2.3. Ranking Arguments for Quality

Previous work on automated evaluation of long-form persuasive essays (Ghosh et al., 2016), (Klebanov et al., 2016) (Nguyen and Litman, 2018) has focused on modelling the holistic scores given by experts. Our work here does not set out to "grade" student explanations, but provide a ranked list for *convincingness* as judged by a set of peers.

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), which has also been specifically proposed for ordinal peer grading data(Raman and Joachims, 2014). 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. By appending a softmax layer to the output, pairwise preferences and overall ranks were jointly modelled in publicly available datasets (Gleize et al., 2019).

The key difference between to keep in mind between this work in modelling the quality rankings of arguments, and that of TMPI explanations, is that the students are not indifferent crowd-labellers: each student will have just submitted their own explanation justifying their answer choice, and we analyze the aggregate of their choices as they indicate when a peer may have explained something better than themselves.

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 one of its variants (the Elo rating system), and the simple heuristic scoring model. (We omit a neural approach in this study, as we consider the work on interpreting the model results from a neural model for pedagogical purposes, out of the scope of this paper.)

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 2.3. In order to contextualize the performance of these methods in our educational 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 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. There is no simple and reliable way to determine whether students choose this option "genuinely" (because the shown alternatives were not sufficiently convincing), or because they did not want to read their peers' explanations. For this reason, we only include observations where students explicitly change explanations (whether for their own answer choice, or for a different answer choice, regardless of correctness.) There is a strong bias for students to simply choose '*I stick to my own rationale*, and so this reduces our data by approximately 50%.

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

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

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.

3. As a platform for formative assessment, not all instructors provide credit for the explanations students write, and there are invariably 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.

4. after the previous two steps, we only include data 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.

After the filtering steps described above, we take the TMPI observations for each question, and construct explanation pairs, as in figure 2(b).

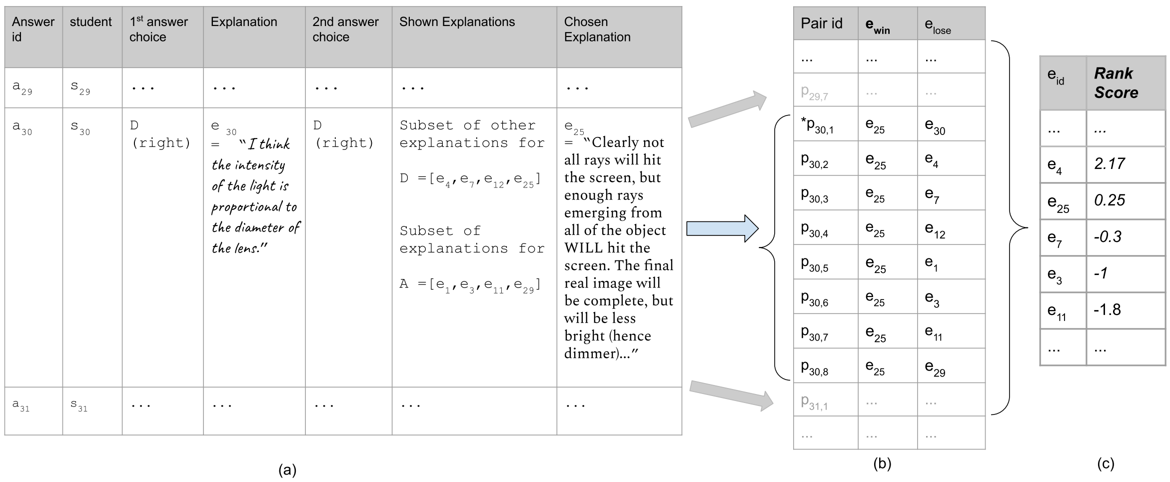


Figure 2: Example of student-item observations from a TMPI environment. This figure follows from figure 1. (a) Student s_{30} chose the correct **D** as the answer on their first attempt, and provided the explanation e_{30} in the dataset for this question. The student is shown a subset of explanations 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 8 explanation pairs. The first pair is for the choice of e_{25} over what the student wrote themselves, and the other seven are for the choice of e_{25} over the other shown explanations. The pairs are labelled as such that e_{25} is the more convincing of the pair. (c) This pairwise preference data is aggregated global ranked list of student explanations for this question, where each explanation is assigned a real-valued rank score (using the methods described in section 4.1).

Using these explanation pairs, we apply the following rank aggregation techniques in order to derive a real valued *convincingness* rank score, as in figure 2(c).

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

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

   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

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

where δ is a constant. All arguments are initialized with an initial strength of β_0, and the rating of any argument is only updated after it appears in a pairwise comparison with another. The rating update rule transfers "strength" 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 (all rank scores are recalculated after each pair is seen), **Elo** ratings are dynamic and implicitly give more weight to recent data(Aldous, 2017).

How we evaluate the fit of these ran aggregation methods to our data is described in section 4.3

4.2. Modelling rank scores

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.

Recent experimental results posted state-of-the-art results for this same regression task on a large argument mining dataset, using a neural embeddings in a bidirectional encoder representations from transformers (BERT) (Gretz et al., 2019). However we favour a feature-rich approach and simpler learning algorithms, keeping in mind downstream priorities such as interpretability for teachers in their reporting tools.

In order to evaluate the generalizability of our models to new question items, we employ a "cross-topic" cross-validation scheme, wherein we hold out all of the answers on one question item train models on all of the answers for all question items except one 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 the context writing analytics that are not included here are cohesion, sentiment, and psycholinguistic features, as they do not seem pertinent for shorter responses that deal with STEM disciplines.

4.3. Evaluation of methodology

In order to evaluate our choice of rank aggregation scores, and address our research question RQ1, we perform several validation tests.

Firstly, the reference argument mining datasets that we use for this study, each include some their own derived aggregated rank score (described in 2.3) for each argument. We evaluate the soundness of our choice of simpler rank aggregation methods by measuring the correlation between our ranking scores and the reference scores. For each topic in the different AM datasets, we calculate the Pearson correlation between the "reference" score each argument and our simpler scores. The distribution of Pearson correlation coefficients across the different topics are shown in the boxplots in figure 3,

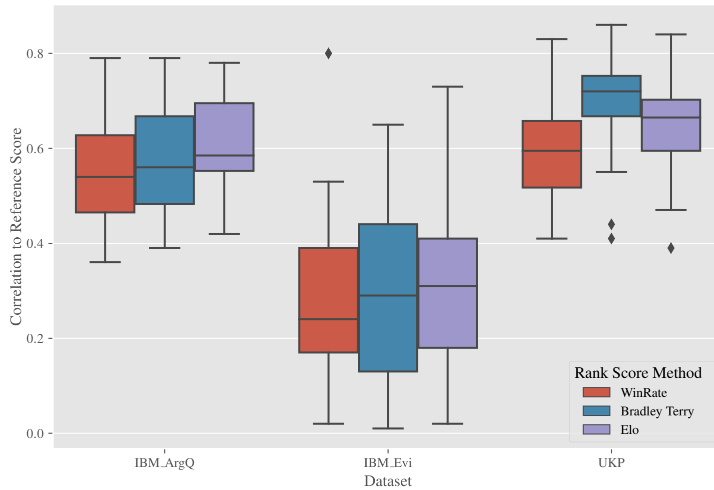


Figure 3: Distribution of Pearson correlation coefficients measured between "reference" rank scores, and the rank aggregation methods (WinRate, BT, Elo) used in our proposed methodology, across the different topics of the reference argument mining datasets.

In order to evaluate the reliability of these rankings we employ a validation scheme similiar to one proposed by (Jones and Wheadon, 2015). Students are randomly split into two batches, and their answers ...

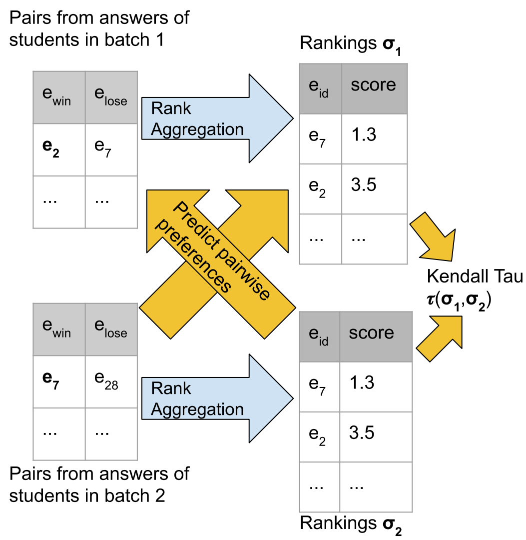


Figure 4: Evaluating of reliability of rank scores: for each question, student answers are divided into two batches, yielding two batches of corresponding pairs, and two aggregated rankings. Two measures of reliability of the derived rankings are shown with the yellow arrows: i) the rank scores of each batch can be used to predict the pairwise preferences of the other batch of students, and ii) the Kendall tau correlation coefficient can be calculated between the two independantly derived ranked lists for each batch of students.

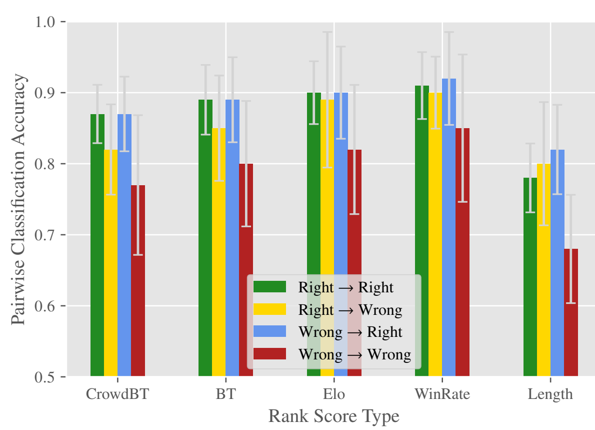


Figure 5: Comparing the classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair. Rank scores are calculated with the vote data of half the students, and tested on the pairs generated by the other half. Data is averaged across all questions, dis-aggregated by different TMPI transition types.

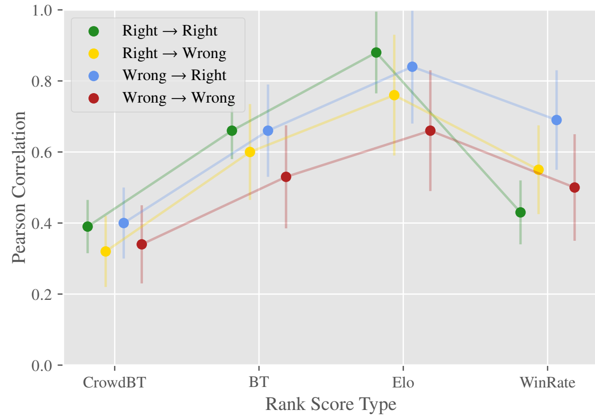


Figure 6: Pearson correlation coefficient between different rank score types, derived from two independent groups of students, averaged over all questions, dis-aggregated by different TMPI transition types.

5. RESULTS

TO DO

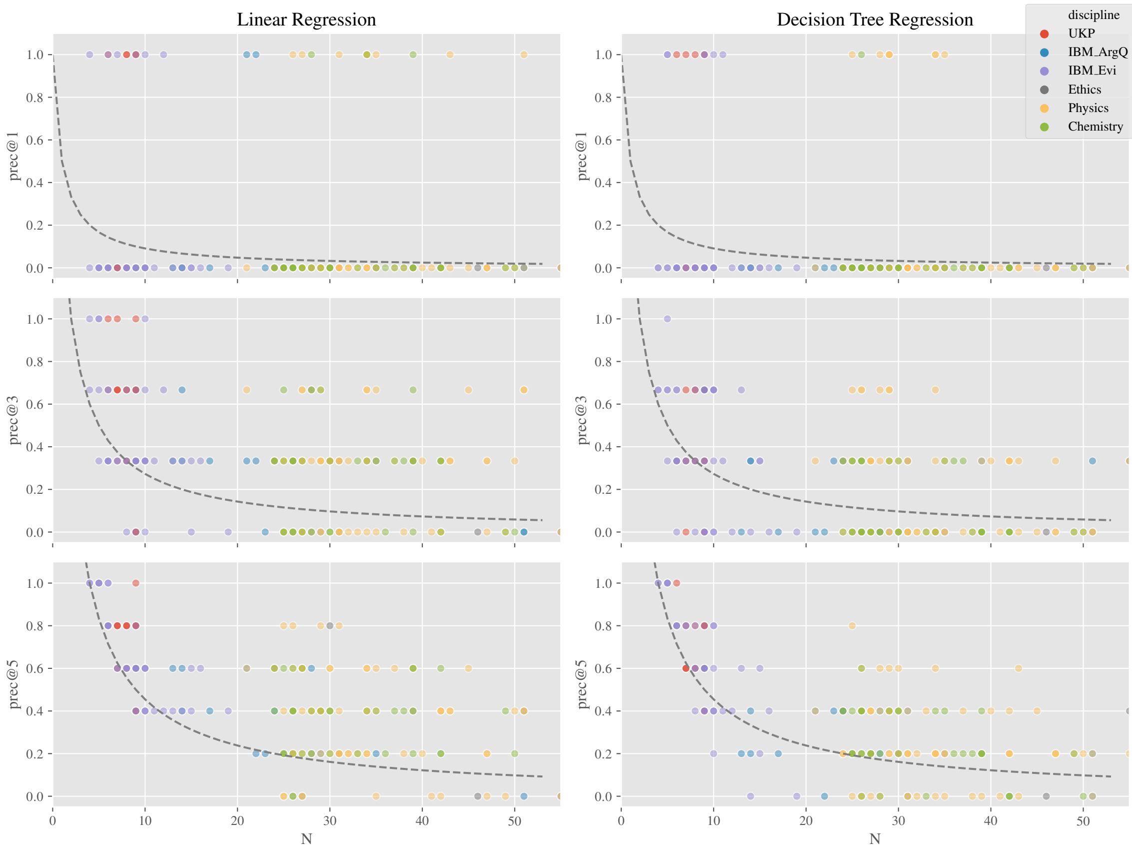


Figure 7: Evaluating of regression models tasked with predicting *convincingness* score of arguments/TMPI-explanations based on linguistic features. Evaluation metric is **precision @ K**, where we verify how many of the predicted top-K ranked explanations are in the measured top-K list (using "winrate" 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.

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.

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.

BURROWS, S., GUREVYCH, I., AND STEIN, B. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education 25,* 1, 60–117. Publisher: Springer.

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.

GHOSH, D., KHANAM, A., HAN, Y., AND MURESAN, S. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 549–554.

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*.

GRETZ, S., FRIEDMAN, R., COHEN-KARLIK, E., TOLEDO, A., LAHAV, D., AHARONOV, R., AND SLONIM, N. 2019. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis. *arXiv preprint arXiv:1911.11408*.

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.

JONES, I. AND WHEADON, C. 2015. Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation 47*, 93–101. Publisher: Elsevier.

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

KLEBANOV, B. B., STAB, C., BURSTEIN, J., SONG, Y., GYAWALI, B., AND GUREVYCH, I. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. 70–75.

KOLHE, P., LITTMAN, M. L., AND ISBELL, C. L. 2016. Peer Reviewing Short Answers using Comparative Judgement. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 241–244.

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

MOHLER, M., BUNESCU, R., AND MIHALCEA, R. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 752–762.

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.

RAMAN, K. AND JOACHIMS, T. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1037–1046.

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.

SULTAN, M. A., SALAZAR, C., AND SUMNER, T. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1070–1075.

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.

ZHANG, Y., SHAH, R., AND CHI, M. 2016. Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading. *International Educational Data Mining Society*. Publisher: ERIC.