Modelling Argument Quality in Technology Mediated Peer Instruction

Sameer Bhatnagar
Polytechnique Montreal

Michel C. Desmarais
Polytechnique Montreal

Amal Zouaq
Polytechnique Montreal

Technology Mediated Peer Instruction (TMPI) is the process whereby students submit explanations to justify their reasoning, and are subsequently prompted to reconsider their own answer by being presented with explanations written by their peers. We frame this as an instance of comparative judgment, as applied to evaluating the quality of natural language arguments, along the dimension of *convincingness*. This study proposes a two-step methodology for modelling data from TMPI: aggregation of pairwise preference data to produce rankings ordered on quality (as judged by peers), followed by a regression task using a rich set of linguistic features as input to supervised learning algorithms. We evaluate this methodology on publicly available datasets from argument mining research, and apply it to data from a TMPI learning environment spanning data from multiple disciplines. We compare feature-rich regression models that favour interpretability, with the current state-of-the-art neural approach, and provide insight as to the features and question types where modelling *convincingess* for pedagogical support is most easily achieved.
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.

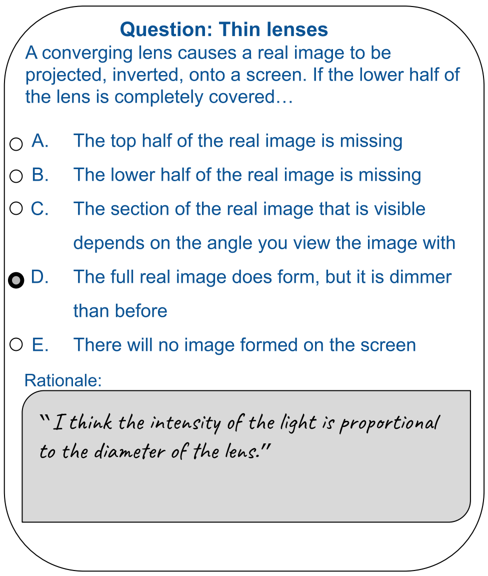
On the second step (figure 1b), students are prompted to revise their answer choice, by taking into consideration a subset of explanations written by their peers.

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.

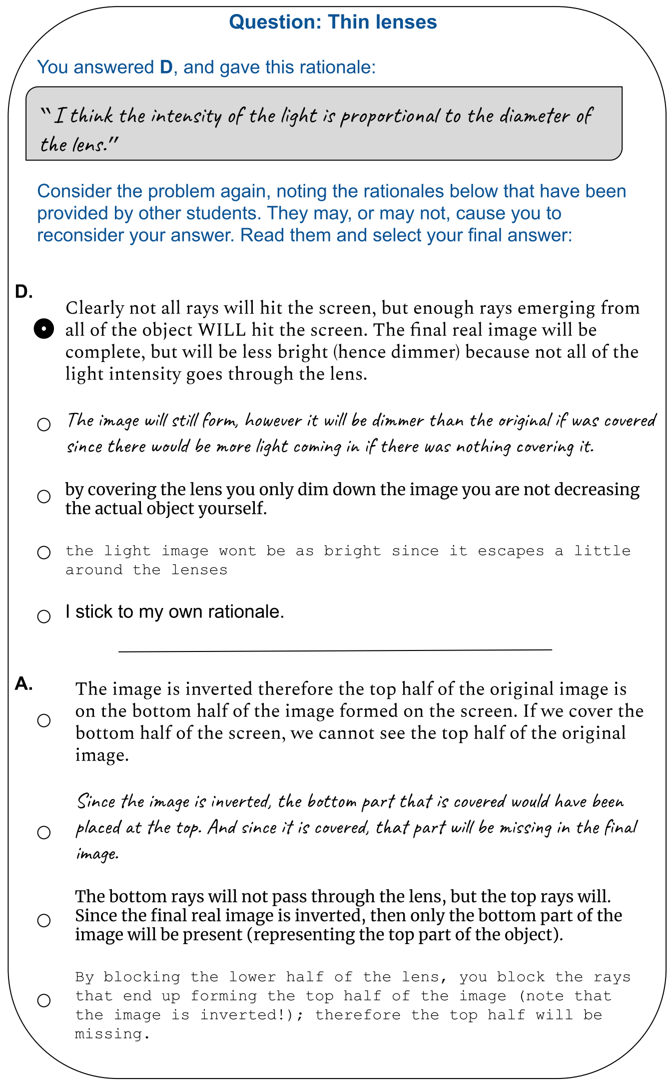
Whenever the student goes with either of the first two scenarios above, we frame this as "casting a vote" for the chosen peer explanation.

Moreover, when one of the answer choices is labelled at "correct", and the other are "incorrect", as is often the case in question items from the STEM disciplines, the three possibilities above can produce one of four *transitions*: Right → Right, Right → Wrong, Wrong → Right, or Wrong → Wrong. The transition possibilities, and the relative proportions present in the the TMPI platform we study, are shown in the Sankey diagram of figure 2.



(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)

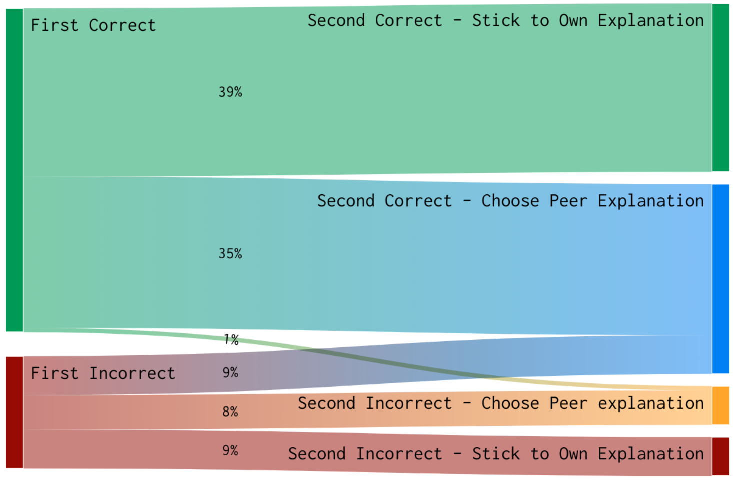


Figure 2: The possible transition types that can occur in TMPI for student answers between their first attempt (when they write their own explanation), and the review step (when they are presented with peer explanations). The relative proportion of each transition type is shown in this Sankey diagram for data from myDALITE.org

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).

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 (AM) 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 3). 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 specifically leveraging peer judgments of peer-written explanations, 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 novel scheme proposed by (Kolhe et al., 2016) also 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, using typically between 10-100 words. Their in-depth historical review of ASAG systems describes 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) measure 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.

We cast this as a task in rank aggregation, with the objective 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. The trade-off, of course is the quadratic scaling in the number of pairs one can generate. 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, and the potential permutations of explanations different students may see is intractably large for a typical question answered by 100+ students.

A classical approach specifically proposed by (Raman and Joachims, 2014) for ordinal peer grading data is the Bradley-Terry (BT) model. The BT model (Bradley and Terry, 1952) for aggregating pairwise preference data into a ranked list, assumes that predicting the winner of a pairwise "match-up" between any two items is associated with the difference in the latent "strength" parameters for those two items, and these parameters can be calculated using maximum likelihood estimation.

The BT method 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).

Specifically in the context of evaluating argument convincingness from pairwise preference data, 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 passage pairs that lead to cycles in the graph, PageRank scores are derived from this directed acyclic graph, and are used as the gold-standard rank for convincingness (Habernal and Gurevych, 2016). (This dataset is included in our study, from now on labelled as **UKP**.)

More recently, a relatively simpler heuristic WinRate score has been shown to be a competitive alternative for the same dataset, 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). (This dataset is also included in our study as a reference, labelled as **IBM_Evi**.)

The key difference between to keep in mind between the above mentionned studies 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, one of its variants (the Elo rating system), and the simple heuristic scoring model, "WinRate". (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. The methods we chose have several readily available implementations in different programming languages, and we err on the side of simplicity when possible in our methodological choices.)

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

Table 1: Examples of argument pairs from each reference argument mining datasets. These examples were selected because they were incorrectly classified by all of our models, and demonstrate the challenging nature of the task. In each case, the argument labelled as more convincing is in *italics*.

(a) A pair of arguments from **UKP**, for the prompt topic: "school uniforms are a good idea".

a1	a2
I take the view that, school uniform is very comfortable. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Every teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	*I think it is bad to wear school uniform because it makes you look unatrel and you cannot express yourself enough so band school uniform OK*

(b) A pair of arguments from **IBM_ArgQ** , for the prompt topic: "We should support information privacy laws".

a1	a2
if a company is not willing to openly say what they are going to do with my data, they shouldn't be allowed to do it.	if you are against information privacy laws, then you should not object to having a publicly accessible microphone in your home that others can use to listen to your private conversations.

Table 2: Examples of argument pairs from Physics and Ethics disciplines, taken from our TMPI environment. These examples were selected because they were incorrectly classified by all of our models, and demonstrate the challenging nature of the task. In each case, the argument labelled as more convincing is in *italics*.

(a) Student explanations from **dalite**, for the question prompt: "Rank the magnitudes of the electric field at point A, B and C shown in the following figure from greatest magnitude to weakest magnitude".

a1	a2
At B, the electric field vectors cancel (E=0). C is further away than A and is therefore weaker.	A is closest, B experiences the least since it is directly in the middle, and C the least since it is most far away.

(b) Student explanations from **TMPI** in an Ethics MOOC, for the question prompt: "Assuming that motorcycle drivers are willing to pay their own medical bills, should they be allowed to ride without a helmet?".

a1	a2
Law should always to make for the good of their people.If wearing helmet help,then it should be enforce.Also,you cannot assure that every motorcycle in the country would want to pay.	I believe that motorcycle helmets should be mandatory for ALL motorcycle drivers . Although they may be willing to pay their own medical bills , you ca n't pay anything if you 're not alive to do so . Motorcycle wrecks can kill the driver , leaving the drivers family with funeral expenses and the like . The family may not be able to afford it . Wearing a helmet increases possibility of surviving a crash that you may not otherwise survive . So yes , motorcycle helmets should be mandatory even if the rider is willing to pay their own medical expenses ..

of these methods in our educational setting, we apply the same methods to publicly available datasets from the AM research community as well, and present the results. These datasets are described in table 3, alongside the TMPI data at the heart of our study.

The **UKP** dataset (Habernal and Gurevych, 2016) is one of 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 dataset that is similarly labelled, but much more tightly curated, 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 in section 4.4.

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, SALTISE, focused on supporting teachers in the development of active learning pedagogy. 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 3 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:

- approximately 1 in 10 students decide that they want their explanations to be shared with only with their instructor, and not seen by other students, nor used for the purposes of research. The answers of these students are removed from the dataset.
- 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%.

[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 3: 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.

- 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.
- 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 include only those student answers that have at least 10 words.
- after the previous two steps, we only include data from those questions that have at least 100 remaining student answers.
- we remove any duplicate pairs before the rank aggregation step that have the same "winning" label, as explanations that appear earlier on in the lifetime of a new question are bound to be shown more often to future students.

We see in table 3 that the resulting datasets from the different disciplines in our TMPI dataset are comparable to the reference AM datasets (just proportionately larger). The division of the TMPI data into multiple disciplines, despite the source the same platform, is because we assume that in modelling the quality of explanations, different features will be important for each.

4. Methodology

We borrow our methodological approach from research in argument mining, 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. The 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 3(a), the fields include the item prompt, the student's *first* answer choice, their accompanying explanation, the peer explanations shown on the review step (as in figure 1b), 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 3(b).

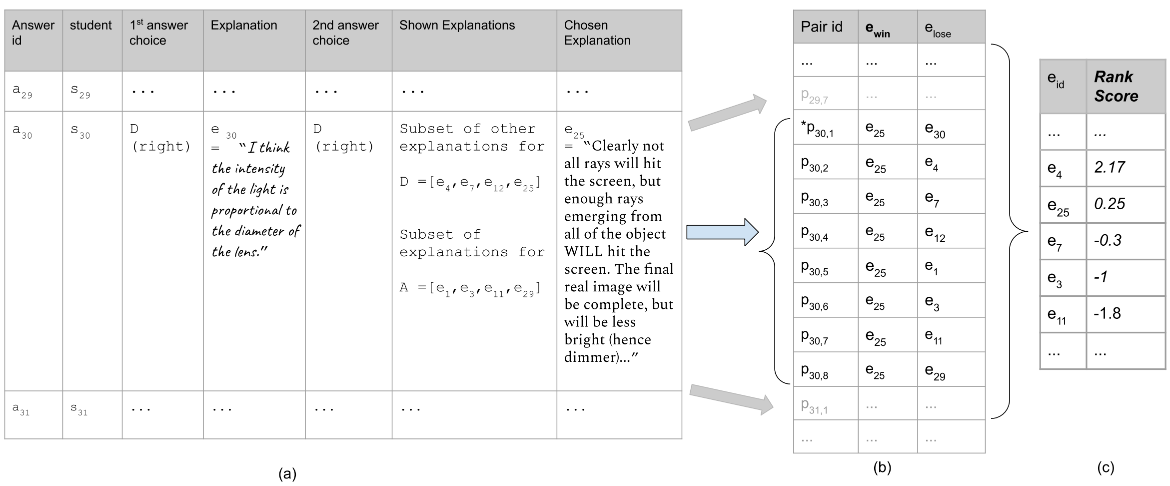


Figure 3: Example of student-item observations from a TMPI environment. This figure follows from figure 1. (a) Student s_{30} chose the correct **D** as the answer on their first attempt, and provided the explanation e_{30} in the dataset for this question. The student is shown a subset of explanations 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 3(c).

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

times it was shown. This method does not take into account *which* peer explanations were shown to each student, neglecting the potential impact of comparative judgment.

2. **WinRate**: as described in (Potash et al., 2019), this measure argument quality is defined as the number of times it is chosen as more convincing in a pairwise comparison, normalized for the number pairs in which it appears. In the context of TMPI, when we calculate the *WinRate* of a student eplanation after the data transformation depicted in figure 3a and figure 3b, we take a step towards including the effect of the comparative judgement, as pairs are specifically constructed for each observation from the explanation that was chosen, and the ones that were shown.

3. **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 $\sum_i \beta_i = 0$.

4. 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 latent "strength" rating points from the loser, to the winner, 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).

5. **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 η_k 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

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

How we evaluate the fit of these rank aggregation methods to our data is described in section 4.4

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.

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: sentence count, max/mean word length, max/mean sentence length;
- Lexical: uni-grams, type-token ratio, number of keywords (defined by open-source discipline specific text-book), number of equations (captured by a regular expression);
- Syntactic: POS n-grams (e.g. *nouns, prepositions, verbs,conjunctions,negation, adjectives, adverbs, punctuation*), modal verbs (e.g. *must, should, can, might*), dependency tree depth;
- Semantic:
  - Using pre-trained GloVe (Pennington et al., 2014) vectors, we calculate similarity metrics to i) all other explanations, ii) the question item text, and, when available, iii) a teacher provided "expert" explanation.

– we derive our own discipline specific embedding vectors, trained on corresponding open-source textbooks[2]. We experiment with a word-based vector space model, Latent Semantic Indexing (LSI) (Deerwester et al., 1990), due to its prevalence in text analytics in educational data mining literature, as well as Doc2Vec (Le and Mikolov, 2014), which directly models the compositionality of all the words in a sentence[3]. We take the text of the question prompt, as well as an "expert explanation" provided by teachers for each question, and determine the 10 most relevant sub-sections of the textbook. For each student explanation, we then calculate the mininum, maximum, and mean cosine similarity to these 10 discipline specific "reference texts".

- Readability: Fleish-Kincaid reading ease and grade level, Coleman-Liau, automated readability index, spelling errors

Features typical to NLP analyses in the context writing analytics that are not included here are cohesion, sentiment, and psycho-linguistic features, as they do not seem pertinent for shorter responses that deal with STEM disciplines.

4.3. Regression Models

The machine learning models we explore for the regression task are also inspired from writing analytics literature, as well as the design objective, of maximizing interpretability: the ability to explain predictions of which students explanations are most *convincing* is paramount in providing pedagogical support to students an teachers.

As has been described in related work, argument *length* is a difficult baseline to beat when modelling *convincingness* in pariwise preference data. The greater the amount of words, the greater the opportunity to construct a convincing argument, and as such, we set explanation `Length` (the number of whitespace seperated tokens) as our regression baseline as well. The models we include in this study are `Linear` regression, `Decision Tree` regression, and `Random Forest` regressors. So as to provide context with the current state of the art in point-wise prediction of argument *convincingness*, we also fine-tune a pre-trained bi-directional neural transformer model, with argument mining reference datasets, as well as the TMPI data from our three disciplines. In line with the best performing model in (Gretz et al., 2019), the question prompt is combined with the student explanation, separated the model-specific `[SEP]` token, and input as a pair of text sequences during fine-tuning and inference. In what will henceforth be referred to as `BERT_Q`, the contextual embedding of the model-specific `[CLS]` token in the last layer of the fine-tuned transformer is fed as input into a fully dense regression layer, so as to output a predicted *convincingness* score.

4.4. Evaluation of methodology

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

The reference argument mining datasets that we use for this study, along with annotated pairwise preference data, each include their own derived aggregated rank score for each argument (described in 2.3). We begin our evaluation of the soundness of our choice of simpler

[2]https://openstax.org

[3]model implementations from https://radimrehurek.com/gensim/index.html

rank aggregation methods, by measuring the correlation between our ranking scores, and the reference scores, on the AM datasets. For each topic in the different AM datasets, we calculate the Pearson correlation between the "reference" score of each argument, and the simpler scores we choose to include in our methodology (*WinRate*,*BT*, *Elo*. We cannot include *CrowdBT* here, as the AM datasets do not include an identifier for "annotator"). The distribution of Pearson correlation coefficients across the different topics are shown in the box plots in figure 4.

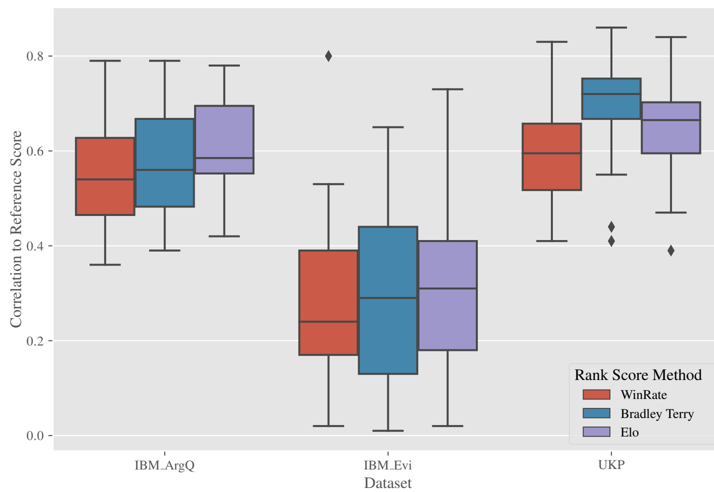


Figure 4: Distribution of Pearson correlation coefficients measured between "reference" rank scores, and the rank aggregation methods (WinRate, BT, Elo) used in our proposed methodology, across the different topics of the reference argument mining datasets.

While the variance across topics of the correlation coefficients between the "out-of-the-box" reference scores and our simpler rank-aggregation scores is quite large, the median lies between 0.5 and 0.7 for the **UKP** and **IBM_ArgQ** datasets. These are significantly higher than for **IBM_Evi**, likely because the reference scores for this set are dependant on a specific Bi-LSTM architecture. The relative alignment between our chosen rank aggregation techniques (*WinRate*, *Bradley-Terry*, and *Elo*), and the modified PageRank score provided with **UKP**, indicates that all capture approximately the same information about overall *convincingness*. Also of note is the correlation between the **IBM_ArgQ** reference rank score, and the methods we include in our methodology. The reference score here was actively collected by the authors of dataset, first by presenting crowd workers with individual arguments, and prompting them to give a binary score of 1/0, based on whether "they found the passage suitable for use in a debate", and then averaging the score over all labellers. The correlation between *WinRate*, *Bradley-Terry*, and *Elo*, and this actively collected reference score, would indicate that these methods capture a "true" ranked list.

In order to evaluate a measure of *reliability* of these rankings, we employ a validation scheme similar to one proposed by (Jones and Wheadon, 2015). Students are randomly split into two batches, and their answers are used to derive two independent sets of rank scores, as shown in figure 5.

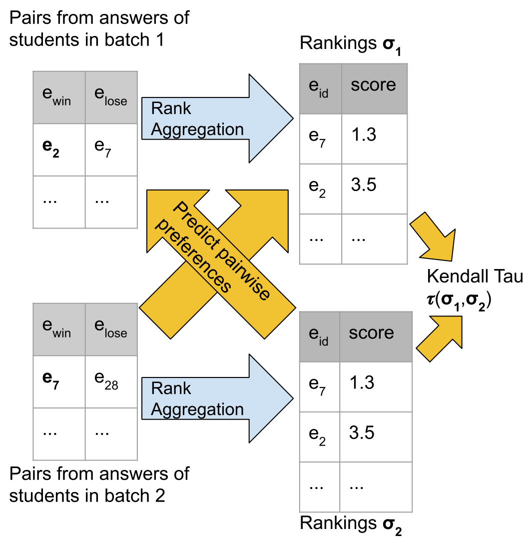


Figure 5: Evaluating of *reliability* of rank scores: for each question, student answers are divided into two batches, yielding two batches of corresponding pairs, and two aggregated rankings. Two measures of reliability of the derived rankings are shown with the yellow arrows: i) the rank scores of each batch of students can be used to predict the pairwise preferences of the other batch, and ii) the Kendall tau correlation coefficient can be calculated between the two independently derived ranked lists for each batch of students.

We apply these evaluations of reliability on the derived rank scores from the pairwise preference data from `dalite`, and dis-aggregate the results by possible TMPI transition types (figures 6 and 7.

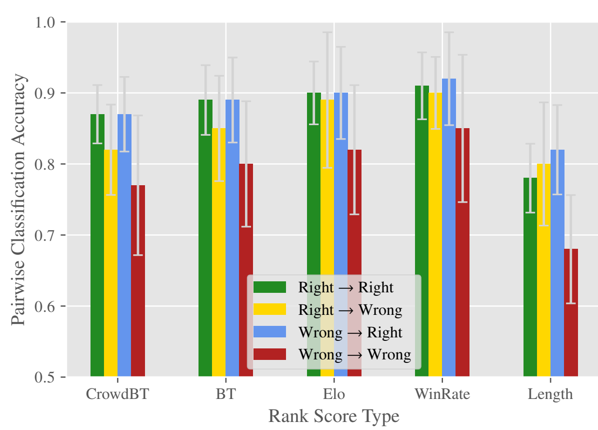


Figure 6: Comparing the average pairwise classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair. Rank scores are calculated with the vote data of half the students, and tested on the pairs generated by the other half. Data is averaged across all questions, dis-aggregated by different TMPI transition types.

It should be noted that, as shown in the relative proportions of the Sankey diagram (figure 2, the vast majority of the data is represented in the Right→Right transition (the rarest transition is Right→Wrong). When we consider using the rankings derived from one batch of students, and use them to predict the pairwise preferences of the other batch, the classification accuracies are roughly equivalent across the different rank score methods (figure 6). All of the methods outperform a baseline "Length" method, which is where the pairwise preference is chosen by simply choosing the explanation with the most words.

However there seems to be a slight advantage with the *Elo* method in evaluating the reliability of the rankings across independent batches of students if we consider the alignment between the rankings themselves (figure 7).

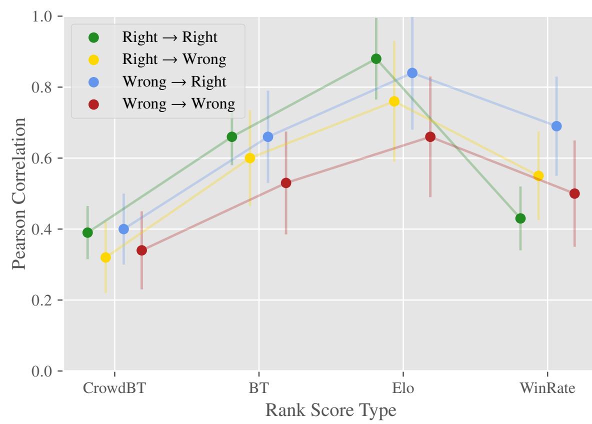


Figure 7: Mean of Pearson correlation coefficients between independent rank scores, derived from two independent batches of students, averaged over all questions, dis-aggregated by different TMPI transition types.

In practice, after choosing the most reliable rank-aggregation scoring method, the second step of our proposed methodology is to address our second research question, **RQ2**, and build feature-rich supervised regression models to predict the individual argument scores. We choose our feature sets based on relevant related research, as described in section 4.2, and use Pearson and Spearman correlation coefficients to performance, as is standard practice in the litterature on pointwise prediction of argument quality along the dimension of *convincingess*.

In order to estimate the generalizability of these 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 as the test set, training models on all of the answers for all other question items in the same discipline. This approach is meant to capture discipline specific linguistic patterns, while addressing a the "cold-start" problem for new items before vote data can be collected.

Once feature-rich models are trained and tested under this validation scheme, we inspect these using *permutation importance*, based on *feature importance* introduced by (Breiman, 2001) for random forests, and generalized to be model agnostic by (Fisher et al., 2019): each feature is randomly permuted for a set number of repetitions, and the importance of that feature is measured by the average decrease in performance of the model on the un-perturbed dataset.

y_reference model	UKP	IBM_ArgQ	IBM_Evi
Length	0.33	0.14	0.15
Linear	0.23	0.14	0.28
DTree	-	0.17	0.25
RF	0.34	0.25	0.33
SotA	0.49	0.42	-
BERT	0.23	0.37	0.5
BERT_Q	0.26	0.39	0.56

(a) Pearson correlation

y_reference model	UKP	IBM_ArgQ	IBM_Evi
Length	0.59	0.14	0.15
Linear	0.33	0.17	0.35
DTree	-	0.20	0.24
RF	0.46	0.23	0.32
SotA	0.67	0.41	-
BERT	0.36	0.34	0.5
BERT_Q	0.37	0.37	0.55

(b) Spearman correlation

Table 4: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the different "ground truth" reference score accompanying different argument mining datasets

5. Results & Discussion

One of the contributions of this study is to propose a methodology for the analysis and leveraging of learnersourced explanation quality labels inside TMPI learning environments, or more broadly speaking, any setting where there is an ordinal/comparative peer grading task for natural language student submissions.

We begin by applying this methodology on publicly available argument-mining datasets in table 4, wherein we use the linguistic features described in section 4.2 (excluding those linked to any specific disciplinary textbook). We train our different models to predict the real-valued *convincingness* score provided by these datasets, and report the state-of-the-art (*SotA*) Pearson and Spearman correlation coefficients.

It should be noted that the real-valued *reference* scores that are provided with the argument mining datasets, and are the target variables for the model training in table 4, are each calculated in different ways. For example, in the **UKP** dataset, the ground truth *convincingess* score of each argument is derived by constructing an argument graph, and calculating a variant of the PageRank score, after removing cycles induced by the pairwise data (e.g. cases where *A* is more convincing *B*, *B* is more convincing than *C*, but *C* is more convincing that *A*). For **IBM_ArgQ**, the real-valued score is the average of multiple binary "relevance judgments" explicitly collected from a set of crowd-labellers. Finally, the real-valued score accompanying arguments in **IBM_Evi** are the output of a regression layer that is appended to a two-armed siamese Bi-LSTM model, wherein only one arm is provided with the a GloVe embedding of the input argument.

So as to be able to more consistently compare the impact of our methodological choices across datasets, in tables 5 and 6, we train our models to predict a target variable that can be calculated from any pairwise preference dataset, namely **WinRate** and **BT**. To the best of our knowledge, we are the first to propose and evaluate common rank aggregation methods for the calculation of pointwise argument quality scores from pairwise preference data.

While our best performing feature-rich model (`Random Forests`) beat the `Length` baseline, the fine-tuned neural transformer model `BERT_Q` dramatically outperforms across the three

y_winrate model	UKP	IBM_ArgQ	IBM_Evi
Length	0.58	0.13	0.13
Linear	0.18	0.23	0.17
DTree	0.46	0.21	0.21
RF	0.60	0.28	0.30
BERT	0.71	0.52	0.46
BERT_Q	0.72	0.51	0.48

(a) Pearson correlation

y_winrate model	UKP	IBM_ArgQ	IBM_Evi
Length	0.61	0.12	0.11
Linear	0.22	0.24	0.20
DTree	0.45	0.21	0.19
RF	0.59	0.28	0.30
BERT	0.70	0.51	0.44
BERT_Q	0.70	0.51	0.46

(b) Spearman correlation

Table 5: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *winrate* across pairwise preference data, for different argument mining datasets

y_BT model	UKP	IBM_ArgQ	IBM_Evi
Length	0.54	0.16	0.11
Linear	0.20	0.18	0.2
DTree	0.42	0.21	-
RF	0.58	0.31	0.3
BERT	0.60	0.55	0.5
BERT_Q	0.64	0.55	0.52

(a) Pearson correlation

y_BT model	UKP	IBM_ArgQ	IBM_Evi
Length	0.59	0.15	0.11
Linear	0.30	0.25	0.24
DTree	0.41	0.22	-
RF	0.57	0.29	0.29
BERT	0.65	0.55	0.48
BERT_Q	0.69	0.54	0.49

(b) Spearman correlation

Table 6: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *Bradley-Terry* score across pairwise preference data, for different argument mining datasets

y_winrate model	Ethics	Physics	Chemistry
Length	0.16	0.36	0.32
Linear	0.17	0.19	0.14
DTree	0.22	0.27	0.25
RF	0.26	0.35	0.31
BERT	0.29	0.38	0.34
BERT_Q	0.29	0.39	0.34
BERT_A	-	0.37	0.31

(a) Pearson correlation

y_winrate model	Ethics	Physics	Chemistry
Length	0.24	0.34	0.32
Linear	0.21	0.25	0.19
DTree	0.23	0.27	0.27
RF	0.26	0.33	0.31
BERT	0.29	0.35	0.32
BERT_Q	0.29	0.36	0.33
BERT_A	-	0.36	0.30

(b) Spearman correlation

Table 7: Average correlation (under cross-topic validation scheme)between convincingness score predicted by different models, and the convincingness score as given by the *winrate* across pairwise preference data, for different disciplinary datasets from TMPI environment

reference datasets, when all trained for the same task, under the same cross-topic validation scheme, confirming results from the literature.

When we move to extend this methodology to data from our three TMPI discipline-specific datasets in tables 7 and 8, we observe that while `BERT_Q` is also clearly the best performing model for the **Ethics** dataset, in **Physics** and **Chemistry**, the `Length` baseline is most correlated with *convincingness* scores, whether calculated via *WinRate* or *Bradley-Terry*.

This may be best explained by the inherent similarities between the **Ethics** TMPI data, and the argument mining datasets: the topic prompts are subjective and personal, and the available answer options students are to choose from are also limited to two opposing stances of an argument, as can be seen in the sample data in tables 1 and 2. As for the the **Physics** and **Chemistry** disciplines, no models significantly outperform the `Length` baseline, which seems to indicate that, in general, students are simply more generally convinced by longer explanations. This is true also when our rank aggregated score jointly estimates the student's agreement with their peer's, as in the **CrowdBT** score (results in table 9).

It should be noted that under our cross-topic validation scheme, different question-level folds witness significantly better agreement between model predictions and rank aggregated *convincingness* scores.

In both **Physics** and **Chemistry**, the question-level folds where our models performed *worst* were with question prompts which ask students to choose one true statement from among a selection (e.g. *Which of the following statements about the force of gravity is false? a) ..., b) ...,*). We posit that the language students use to formulate their explanations in such a multiple choice question item, many describing their internal process of elimination to find the correct answer choice, include patterns our models are not able to learn in the training data.

5.1. Controlling for explanation length

Our contributions are centred on our research questions stated at the beginning of this study. In terms of **RQ1**, we present a methodology grounded in argument mining research and empirically demonstrate its validity in the context of TMPI. We present the result of different approaches to

y_BT model	Ethics	Physics	Chemistry
Length	0.17	0.37	0.36
Linear	0.21	0.27	0.21
DTree	0.21	0.31	0.26
RF	0.25	0.35	0.35
BERT	0.32	0.40	0.37
BERT_Q	0.31	0.40	0.37
BERT_A	-	0.38	0.36

(a) Pearson correlation

y_BT model	Ethics	Physics	Chemistry
Length	0.26	0.34	0.33
Linear	0.26	0.30	0.24
DTree	0.22	0.29	0.25
RF	0.26	0.33	0.32
BERT	0.31	0.36	0.33
BERT_Q	0.31	0.37	0.33
BERT_A	-	0.35	0.32

(b) Spearman correlation

Table 8: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *Bradley-Terry* score across pairwise preference data, for different disciplinary datasets from TMPI environment

y_crowdBT model	Ethics	Physics	Chemistry
Length	0.17	0.38	0.36
Linear	0.19	0.25	0.20
DTree	0.24	0.30	0.27
RF	0.28	0.37	0.35
BERT	0.32	0.41	0.36
BERT_Q	0.32	0.41	0.37
BERT_A	-	0.39	0.35

(a) Pearson correlation

y_crowdBT model	Ethics	Physics	Chemistry
Length	0.25	0.34	0.33
Linear	0.23	0.29	0.23
DTree	0.23	0.28	0.27
RF	0.27	0.33	0.32
BERT	0.31	0.36	0.32
BERT_Q	0.3	0.36	0.33
BERT_A	-	0.35	0.32

(b) Spearman correlation

Table 9: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *Crowd-BT* scores across pairwise preference data, for different disciplinary datasets from TMPI environment

y_winrate_nopairs model	Ethics	Physics	Chemistry
Length	0.13	0.22	0.22
Linear	0.07	0.09	0.09
DTree	0.14	0.18	0.12
RF	0.18	0.19	0.13
BERT	0.21	0.23	0.20
BERT_Q	0.2	0.23	0.20
BERT_A	-	0.20	0.18

(a) Pearson correlation

y_winrate_nopairs model	Ethics	Physics	Chemistry
Length	0.26	0.24	0.25
Linear	0.19	0.18	0.14
DTree	0.2	0.19	0.14
RF	0.24	0.22	0.18
BERT	0.27	0.25	0.24
BERT_Q	0.27	0.25	0.24
BERT_A	-	0.23	0.21

(b) Spearman correlation

Table 10: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *raw winrate* across pairwise preference data, for different disciplinary datasets from TMPI environment

y_winrate_longest model	UKP	IBM_ArgQ	IBM_Evi
Length	0.38	0.21	0.28
Linear	0.34	0.14	0.31
DTree	-	0.22	-
RF	0.32	0.30	0.35
BERT_Q	0.23	0.44	0.31

(a) Pearson correlation

y_winrate_longest model	UKP	IBM_ArgQ	IBM_Evi
Length	0.34	0.21	0.25
Linear	0.3	0.16	0.29
DTree	-	0.22	-
RF	0.31	0.29	0.31
BERT_Q	0.23	0.44	0.3

(b) Spearman correlation

Table 11: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *WinRate* across pairwise preference data, for different disciplinary datasets from reference argument mining datasets, on subset of data which includes only the explanations within the top quartile of word counts.

y_BT_longest model	UKP	IBM_ArgQ	IBM_Evi
Length	0.41	0.22	0.26
Linear	0.3	0.25	0.3
DTree	-	0.18	-
RF	0.35	0.35	0.32
BERT_Q	0.26	0.51	0.3

(a) Pearson correlation

y_BT_longest model	UKP	IBM_ArgQ	IBM_Evi
Length	0.38	0.23	0.27
Linear	0.26	0.22	0.26
DTree	-	0.20	-
RF	0.38	0.33	0.32
BERT_Q	0.26	0.48	0.3

(b) Spearman correlation

Table 12: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *BradleyTerry* across pairwise preference data, for different disciplinary datasets from reference argument mining datasets, on subset of data which includes only the explanations within the top quartile of word counts.

y_winrate_longest model	Ethics	Physics	Chemistry
Length	0.08	0.24	0.21
Linear	0.06	0.14	0.16
DTree	0.07	0.17	0.11
RF	0.11	0.21	0.17
BERT_Q	0.15	0.23	0.15

(a) Pearson correlation

y_winrate_longest model	Ethics	Physics	Chemistry
Length	0.12	0.25	0.19
Linear	0.06	0.16	0.17
DTree	0.07	0.17	0.11
RF	0.12	0.21	0.17
BERT_Q	0.15	0.23	0.15

(b) Spearman correlation

Table 13: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *WinRate* across pairwise preference data, for different disciplinary datasets from TMPI environment, on subset of data which includes only the explanations within the top quartile of word counts

y_BT_longest model	Ethics	Physics	Chemistry
Length	0.08	0.26	0.23
Linear	0.05	0.18	0.17
DTree	-	0.19	0.15
RF	0.08	0.24	0.20
BERT_Q	0.17	0.28	0.13

(a) Pearson correlation

y_BT_longest model	Ethics	Physics	Chemistry
Length	0.13	0.25	0.21
Linear	0.08	0.19	0.17
DTree	-	0.20	0.15
RF	0.07	0.25	0.21
BERT_Q	0.17	0.26	0.13

(b) Spearman correlation

Table 14: Average correlation (under cross-topic validation scheme) between convincingness score predicted by different models, and the convincingness score as given by the *WinRate* across pairwise preference data, for different disciplinary datasets from TMPI environment, on subset of data which includes only the explanations within the top quartile of word counts

rank aggregation from pairwise preference data so as to calculate a *convincingness* score for each student explanation. The pairwise transformation of TMPI data, into a format similar to research from argument research (as described in figure 3) allows for a comparison to related work. The modelling results when we train out models to predict the *raw winrate* are significantly worse (table 10) than any of the other rank aggregation methods. This confirms the findings of (Potash et al., 2019), who first proposed that the heuristic, pairwise *winrate* as a more reliable regression target. With such simple methods as *winrate* and the *Bradley-Terry* scores to measure and rank student explanations in a TMPI environment, instructors reports can focus attention of the points where students may have gaps in their knowledge, based on their reading/evaluating of their peers' explanations.

In an effort to provide insight into our **RQ2**, we look at out best performing folds of our regression task, and note the features with the highest permutation importance:

- Type Token Ratio
- Dale-Chall Readability score
- Number Equations
- Vector Similarity to others (GloVe)
- Vector Similiarity to related portion of textbook using LSI
- Number of spelling errors
- Fleich Kincaid reading ease score
- Mathematical expressions used as a VERB e.g. $F = ma$ *tells us that ...*

Another approach we explore to determine which linguistic features are most associated with explanations that are deemed *convincing* by students, is taking our best performing neural transformer model, `BERT_Q`, and finding the features most correlated with its predicted rankings.

We find that the same features which are list above, are also those most highly correlated with the predicted *convincingness* score

It is these types of features that can provide pedagogical insight to instructors when parsing through data generated inside TMPI based activities. These features are predictive of what the students find most convincing in their peer's explanations, and hence offer a much needed lens into how students operate when at the upper levels of Bloom's taxonomy, evalautiong one anothers words in a comparative setting.

6. Limitations & Future Work

Two of the most important differences between TMPI data, and datasets from argument mining research in *convincingness*, are centred on the "student as labeller".

First, in a traditional crowdsourcing setting, the people who choose the most convincing explanations are not the ones who wrote them. In TMPI, the student will be comparing their peers' explanations with each other, but against the explanation they just submitted as well. The effect of this can be seen in the 1 in 2 chance that students decide to "stick to their own" explanation in this TMPI platform.

Second, typical crowdsourcing tasks include filtering questions which are meant to ensure that the workers are qualified and taking the task seriously. While the *crowdBT* rank aggregation method jointly estimates the student's "seriousness" at the labelling task, the almost null improvement of regression results in table 9 seem to indicate that this may not be the best approach: the measure of how much a student "agrees" with the rest of the crowd may not be a useful piece of information in estimating the overall convincingness. The next steps in our research will be to include student-level and question-level features into our analysis (e.g. student strength, question difficulty).

7. Acknowledgements

Funding for the development of myDALITE.org is made possible by *Entente-Canada-Quebec*, and the *Ministère de l'Éducation et Enseignment Supérieure du Québec*. 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 SALTISE/S4 network of researcher practitioners, and the students using myDALITE.org who consented to share their learning traces with the research community.

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.

BREIMAN, L. 2001. Random forests. *Machine learning 45,* 1, 5–32. Publisher: Springer.

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.

DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W., AND HARSHMAN, R. A. 1990. Indexing by latent semantic analysis. *JASIS 41,* 6, 391–407.

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.

FISHER, A., RUDIN, C., AND DOMINICI, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research 20,* 177, 1–81.

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.

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.

LE, Q. AND MIKOLOV, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

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.

PENNINGTON, J., SOCHER, R., AND MANNING, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. Vol. 14. 1532–1543.

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.