1 Editor's note

Dear Sameer Bhatnagar, Michel Desmarais, Amal Zouaq:

Your submission to Journal of Educational Data Mining, "Modelling Argument Quality in Technology Mediated Peer Instruction", has been reviewed by three expert reviewers with a high degree of agreement.

Overall the reviewers felt the work had promise but had concerns about the treatment of the convincingness metric and the depth of the analysis, particularly with respect to interpretation.

Accordingly, the editorial decision at this stage is to invite revisions (major revisions). This means that if you choose to submit a revised version, the same reviewers will be asked to review it. Therefore, please address the comments from these reviews carefully in your revision and submit a document describing the list of changes to the discussion section of the submission page.

Andrew M. Olney
University of Memphis
andrew@jedm.educationaldatamining.org

Authors answer

We first would like to sincerely thank the reviewers for their thorough and helpful reviews. Author comments/answers are shown where appropriate.

We include a diff file that shows the changes between versions.

2 Reviewer A

```
------------------------------------------------------
Reviewer A:
Recommendation: Revisions Required
------------------------------------------------------
```

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

The work is quite relevant overall and on topic should be attractive to the JEDM community.

The novelty is partial. What the authors propose to do is to evaluate their metric for ”convincingness” against others and to also evaluate appropriate models for pooling these metrics and making predictions based upon them. I am not aware of any related work beyond that the authors themselves present but I am concerned that their results don’t add a clear novel outcome to the community due to some confounds and lack of analysis that I will discuss below.

Authors answer

Addressed in the general comments section at the end.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

The authors’ scientific contributions come down to their development of the convincingness metric for student arguments, and the comparison of differing algorithms for estimating these scores in the context of a student argumentation context. While the context is clearly novel and their comparisons add information their analysis limits the generality off their results.

The algorithmic aspects of the work are sound as is the authors description of their models.

Do the authors describe the limitations of their approach in a satisfactory manner?

While the authors are clear on some limitations of their approach I have concerns that they are understating the confound due to the nature of their measures.

How significant is the research? Will the paper be likely to have an impact on the community?

I am concerned that due to some issues with their metric and how they compare it to others as well as to their analysis that the significance will be muted.

Does the title of this paper clearly and sufficiently reflect its contents?

Yes.

Are the presentation, organization and length satisfactory?

The presentation is overall fine though there are some persistent grammar and spelling errors which suggest a deep read is needed. Additionally more analysis would be desireable.

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

No

Can you suggest any reductions in the paper, or deletions of parts?

No

Are the illustrations and tables necessary and acceptable?

Yes

Are the key words and abstracts/summary informative?

Yes

Please list any other general comments or specific suggestions below.

I have some strong concerns about this paper which rest on some apparent inconsistencies in how the work is framed and in what the authors are evaluating.

To begin with the authors state that they are working with a metric of ”convincingness” and take effort to compare it to other rating models and metrics of argument persuasiveness. While that is a useful goal the problem is that their metric of ”convincingness” is, as they acknowledge, substantively different. When you have students rate alternative arguments against their own then you are not actually asking them for a bare statement of convincingness but for an assessment of preference and compatibility. Implicitly students’ willingness to adopt a different argument is going to be influenced by (a) the relative confidence that they have in their stated beliefs, and (b) the extent to which they prefer their writing to someone else.

The authors address this implicitly in part by dropping half of their dataset to only include cases where the students did make a change. While that gets them the binary data that they need it means that they have shifted the ground to a higher threshold for the edges and are conflating student confidence with the change.

As such I do not buy the metric as described. While the authors acknowledge this as a difference in their limitations it seems to me that this is more meaningful because it affects the extent to which we can really compare against the scores in their other datasets.

Authors answer (1)

This answer also addresses some of the points raised earlier.

If we understand correctly, the concern is that the student's choice among a set of explanations that includes only peer explanations is of a different nature compared to the case where it includes one own's explanation. The Arg Mining datasets are of the former type, whereas the DALITE/TMPI ones are of the latter.

We agree with the reviewer that the two types of choices involve different factors (such as (a) and (b), indeed). But while we use more or less the same method to assess what we name convincingness for both contexts, it is not our intent to argue that these types of choices involve the same constructs, nor that the scores are comparable across.

We have rephrased the text to exclude any such suggestions and emphasize that the Arg mining data is another family of data over which the method can be tested on, albeit involving different factors to explain the choices. Furthermore, we echo the differences raised by the reviewer between TMPI and argument data in Section 6 (limitations and future work).

Of course, we could use two different terms in place of *convingness*, one for Arg Mining and one for TMPI. But because it really is the same computed score method, we feel it would make the reading more tedious. See also Authors comments (3).

There is a further challenge here because of the educational context. In essence the prior ”convincingness” metrics focused on abstract ratings of the argument alone from datasets that are more general and, is at least one case, heavily edited. That is not the case here nor do we have a reason to believe that the arguments are correct so the evaluation is again mixing things.

My second critique is over the analysis of the ML results. The authors review a series of ML models and present some discussion of what models and features were better, but not why. In order to generalize based upon this I need to better understand what features were used by the winning models so that I can assess the potential to use this down the road. Simply knowing that random forest won is good but it should be possible to examine whether the features used are really specific to the prompts or are features that can be used across contexts.

Authors answer (2)

We added some details to address this comment, but we do not claim they really allow to assess the individual contribution of the different features across contexts. It would indeed be desirable, but considerably more work. However, we added some text in the future work along these lines.

Finally as a minor quibble, the authors initially describe this work as fitting into the context of argument mining but the argument mining task is about extracting argumentative texts or identifying them. Here the authors are ranking existing texts for quality. Additionally thee work contains some typos which suggest a close read and edit is necessary.

Authors answer

We made some corrections to ensure this work does not fall into the argument mining taks, but solely borrows from this field.

To address these issues I believe that the authors need to make a major revisions to the text to address these concerns and to refine their work.

3 Reviewer B

```
------------------------------------------------------
Reviewer B:
Recommendation: Revisions Required
------------------------------------------------------
```

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

This paper presents a methodology for measuring the convincingness score of peer-submitted explanations in the Technology Mediated Peer Instruction. The proposed approach was validated by comparing with the reference convincingness score using different datasets. In addition, the paper also tried to identify the key features of convincing explanations through building a set of predictive models. The topic of this paper is highly relevant for JEDM.

How novel is the described research? Are the authors aware of related work?

From the methodological perspective, the majority of methodological elements used in this research were established work in relevant domains such as argument mining. However, this research innovatively built a novel framework to evaluate the convincingness of peer-submitted explanations, which was rarely examined in the literature. The authors had good coverage of relevant works in their related work section.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

As said, the major contribution of this work is to introduce a novel methodology for measuring the convincingness of peer submitted explanations in the Technology Mediated Peer Instruction. In general, I found the technical details and results to be well explained. However, it would be more helpful if the practical implications of the proposed approach can be discussed in more detail (please see my comments below).

Is the work technically sound? Are there enough methodological details? Are claims convincingly substantiated, either through theoretical argument or empirical data?

In general, the overall method is well explained. However, there are several improvements the authors might want to consider:

- p. 10, “thus we retain only those student answers that have been presented to at least 5 other students.” why >= 5? Is this an arbitrary decision?

- p.11 “at least 100 remaining student answers.” Any rationales?

- p.13 “Assuming there are explanations, labelled by students, and labelled pairs,” I might be wrong, but it seems that the notations N and SK were not used for the equations in this section.

- p.16 “a univariate feature selection step, wherein the top 768 most discriminating features are retained”. I would suggest adding more details here about the rationale of using univariate feature selection and how it works to select the top 768 features.

- It is not clear how hyperparameters were tuned for the predictive models.

- It would be more helpful to have a table summarizing the major categories of selected features and the number of features for each category.

Authors answer

The above comments were addressed.

Do the authors describe the limitations of their approach in a satisfactory manner?

The authors discussed several limitations. I would also point out that another limitation of the study is a lack of thorough discussion of the implications of the derived convincingness. A question always in my mind when I read this article is: is the convincingness derived from students’ comparison a reliable measure for the quality of peer-submitted explanations? Although the authors compared it with the reference scores, it seems that not all results are supportive (please see my comments below).

Authors answer

The reference scores are taken from the Arg Mining data sets, where there is no clear cut correct answers, and therefore we cannot necessarily expect a high performance on this predictive task. Having an inter-judge agreement measure would help set expectations on the reliability or predictive performance ceiling we could expect, but it is not available.

We added some comments to that effect.

How significant is the research? Will the paper be likely to have an impact on the community?

The proposed approach is capable of deriving convincingness scores automatically, which is very promising for large-scale scenarios. However, in terms of providing feedback to learners and teachers, the significance of the proposed approach was limited. Unfortunately, not many interpretable features were discovered by the predictive models and thus effective interventions on learning might be limited.

Authors answer

We agree with this assessment, although section 5.3 is an attempt towards this goal. See also authors answer (2).

Does the title of this paper clearly and sufficiently reflect its contents?

Yes

Are the presentation, organization and length satisfactory?

Yes

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

See my replies to other questions.

Can you suggest any reductions in the paper, or deletions of parts?

No

Are the illustrations and tables necessary and acceptable?

In the tables, numbers should be presented with the same two decimal places.

Are the key words and abstracts/summary informative?

Yes

Please list any other general comments or specific suggestions below.

p.18 I agree that the validity of the estimated convincingness is partially justified by the correlation analysis. However, I would say a correlation around 0.6 or 0.7 is not strong enough for reaching a claim that the selected methods capture a "true" ranked list.

Authors answer

We consider this is debatable. As mentioned, to properly assess what is strong enough we would need an inter-judge agreement score for comparison.

p.23 The authors stated that "the higher correlation values for WinRate and BT may well better represent the overall aggregate quality of each argument as judged by the crowd." To me, this claim assumes the features used for building the predictive models are known to be representative of convincingness. As such, their predictions can be used to compare with the convincingness scores. However, this would be contradictory to the purpose of the predictive analysis as it aims to identify the influential features of convincing explanations, according to the second research question. Moreover, this study used the reference convincingness scores to validate the proposed convincingness scores, so the reference convincingness scores should be considered reliable. If so, low correlations might indicate that features used for predictive modeling are not indicative of convincingness.

Authors answer

Paragraphe à réviser. Pas certain de bien comprendre la mesure pour tableau 2. A discuter.

p.23 "The slight decrease in performance of BERT_A may reflect that explanations written by content experts are fundamentally different from student explanations, and do not help model convincingness as judged by peers." Does this mean that convincingness judged by peers is very different from convincingness judged by experts? If so, the convincingness judged by peers might be less trustworthy.

Authors answer

We toned down the assertion, because the results are not different enough warrant this language.

p.28 I would suggest elaborating on how the top features selected by the predictive models contribute to improved teaching and learning. The top features seem to suggest writing skills are very predictive of the rank aggregation scores. Is it possible that convincingness evaluated by student peers, as defined in this study, does not truly represent the true convincingness of the written explanations? For example, a student might choose a "convincing" explanation simply based on whether it is well written. A truly convincing but not very readable explanation might be considered unreliable by students.

Another thought: To validate the proposed methods for measuring convincingness, probably a panel of domain experts can be invited to examine the ranking of estimated convincingness of explanations. I understand that the approach proposed in this study is data-driven.

However, there is no adequate evidence supporting the practical implications of the proposed approach. For example, would using the proposed method to derive convincingness contribute to better learning outcomes than using conventional approaches by human raters? Although this is beyond the scope of this study, it is worth being discussed in the paper.

Minor issues:

• Bradley-Terry or BT: I would suggest making the use of terminologies more consistent throughout the paper.

• I would suggest using past tense in the methodology and results sections.

The authors should proofread the manuscript more carefully. Some typos:

• P.4 In “help eliminate the “cold-start” associated when”: is it “cold-start” associated problems?

• p.4 “present students with a a pair of” → : a pair of

• p.7 “submission, Student” → submission. Student

• p.10 “regardless of correctness.)” → regardless of correctness).

• p.10 ‘I stick to my own rationale, → ‘I stick to my own rationale’,

• P.15 “explanation. 5.” → drop a period

• P.15 “average height of syntactic parse tree for each sentence;”: semicolon → period

• p.17 “We refer to this approach as BERT_A, .”: remove comma

• p.21 “cases where A is more convincing B, B is more convincing than C, but C is more convincing that A” → cases where A is more convincing than B, B is more convincing than C, but C is more convincing than A

Authors answer

The above comments are addressed.

4 Reviewer C

```
------------------------------------------------------------
Reviewer C:
Recommendation: Revisions Required
------------------------------------------------------------
```

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

The submission is quite relevant to the scope of JEDM because it is concerned with "processes or methodologies followed to analyse educational data." (1) It proposes and evaluates a new methodology for evaluating the quality of student explanations based on "convincingness".

[1] https://jedm.educationaldatamining.org/index.php/JEDM/about

How novel is the described research? Are the authors aware of related work?

The research described in this paper is novel but potentially incremental. The authors note – "To the best of our knowledge, we are the first to propose, evaluate, and subsequently model a common set of rank aggregation methods for the calculation of point-wise argument quality scores from pairwise preference data." The related work section is sufficient, and the authors clearly distinguish between their work and previous work and state: "We build on this previous work, and move from the pairwise prediction task, to a pointwise regression." My minor concern with regards to the incremental nature of this work can be addressed by including a few sentences on the significance of and the justification for the move from "the pairwise prediction task, to a pointwise regression." Perhaps going through related work such as this case study [1] can help.

[1] Melnikov, V., Gupta, P., Frick, B., Kaimann, D., Hüllermeier, E. (2016). Pairwise versus pointwise ranking: A case study. Schedae Informaticae, 25, 73-83.

Authors answer

Thank you for the suggestion and we took the comments into account.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

The scientific contributions of this submission are clearly explained. It contributes (i) a methodology for evaluating the quality of student explanations in using "convincingness", (ii) an evaluation of the proposed methodology using "data from a real, live TMPI environment", and (iii) a comparison of feature-rich linguistic regression models, with neural transformer-based models, for the prediction of real-valued convincingness scores. With regards to advancing EDM and related fields, the authors "refine work from argument mining research, and propose the use of consistent rank aggregation methods independent of model architecture."More broadly, they "inform pedagogical practice, in how teachers design conceptual questions meant to promote the elaboration of rich self-explanations, which in turn can lead to deep reflection among peers in TMPI."

Is the work technically sound? Are there enough methodological details? Are claims convincingly substantiated, either through theoretical argument or empirical data?

The work is technically sound. In section 4, they provide sufficient methodological details to give the reader a clear picture of what they did both in terms of what their proposed methodology is (4.1-4.3 and how they evaluated it (4.4).

That being said, one big concern I have is that the authors may not have adequately substantiated their fundamental metric of concern – the notion of “convincingness”. There are over 57 mentions of the term, and it appears to be both central to the author’s proposed methodology, and a core concept of interest in the paper. This concern can be addressed by adding more details that elaborate on questions like – How do they define “convincingness”? Why did they choose “convincingness” as opposed to other metrics? Perhaps a crisp one-line description of “convincingness” the first time it is used, along with a few sentences in related work that elaborate on how this metric relates to other metrics will be helpful.

Authors answer (3)

This comment relates to Authors answer (1). We elaborate on the definition of convingness in the context of TMPI and what other measures we could envision. See section “1.1 On convincingness and TMPI”.

Do the authors describe the limitations of their approach in a satisfactory manner?

Yes, the final line in the abstract and section 6 of the paper provide sufficient details on limitations. Three minor changes that would improve their description of limitations include

(i) in the abstract, when the authors note that “While the neural approach is generally the best performing, results show that success on this task is highly dependent on the type of question, which is itself domain dependent”, the replacing “best performing” with actual values that give a sense of the extent of the difference would be helpful.

Authors answer

We would rather not get into specific numbers here because there really are many factors to consider. For one, the performance is a correlation among variables that would need an explanation. And the existence of a correct answer or not, as well as the comparison with the "length" baseline are also factors to be discussed when presenting numbers.

(ii) in section 6, they state "It remains to be shown that providing feedback to students and their instructors, on the relative convincingness of different student explanations has a beneficial impact on learning." Since it appears that learning gain is yet to be observed, a few sentences on why their results so far are still noteworthy would be helpful.

Authors answer

This section was rewritten for the most part.

(iii) In section 2, the authors note – "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." There appears to be a slight issue here because the abstract notes – "results show that success of this task is highly dependent on the type of question, which is itself domain dependent." Therefore, adding a statement here clarifying whether this paper does in fact, make progress on the challenge of "scaling into multiple domains" would be helpful.

Authors answer

These sections were re-rewitten.

How significant is the research? Will the paper be likely to have an impact on the community?

The work described in this paper is significant and is likely to have an impact on the community because it can “inform the design of TMPI platforms”, and more broadly, “contribute to the growing body of research surrounding technology-mediated peer-review.” The authors clearly state their research objectives (provide feedback to learners, provide support to teachers through automation, and maintain the integrity of platforms through filtering), and focus on an important and non-trivial problem of filtering out low-quality explanations in platforms that rely on learner-sourcing.

Does the title of this paper clearly and sufficiently reflect its contents?

Yes, it reflects their core focus on predicting argument quality. Perhaps they could replace ”Quality” with ”Convincingness” because they mention in the abstract that their ”curation task is confined to the prediction of argument convincingness”.

Are the presentation, organization and length satisfactory?

Yes, the paper is well-written and organized, and the length is sufficient. There are some typos and minor grammatical issues:

[ ] Page 1, Abstract: “...the curation process ought to automated” (“...ought to be automated”?)

[ ] Page 4, Figure 2 (a): “Student submits answer choice, and explanation” (don’t use an oxford comma in a two item list)

[ ] Page 8, paragraph 2: “...only in relation to the subset those that are shown” (“subset of those”?)

[ ] Page 8, paragraph 4: “...for evaluating relative reading level in a pair passages (remove “a”)

[ ] Page 15 paragraph 1: “...to our data is described in section 4.4” (add period “section 4.4.”)

[ ] Page 17 paragraph 4: "...convincingness stems ." (move period to end of previous line)

[ ] Page 29 paragraph 5: "...previous study has show that how instructors. . . " (remove "that")

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

Clearly define and justify the use of "convincingness" as a key metric.

Authors answer

All comments above addressed.

Can you suggest any reductions in the paper, or deletions of parts?

Perhaps the examples of argument pairs from each reference argument mining data in Table 5 could be reduced or moved to an appendix (although I do think having these concrete examples are helpful). The description for figure 1b and the caption for figure 4 could be shortened.

Authors answer

Captions shortened a little.

Are the illustrations and tables necessary and acceptable?

Yes. Apart from the minor changes I suggested in the previous question, I found the illustrations and tables to be helpful in terms of making the text more skimmable and easier to understand.

Are the key words and abstracts/summary informative?

Yes. Some minor revisions – consider Including Technology Mediated Peer Instruction (TMPI) to keywords, and actual values for the last sentence in the abstract that talks about results.

Authors answer

Addressed.

Please list any other general comments or specific suggestions below.

Please refer to the attached PDF to see my annotated comments.