Measuring Argument Quality in Technology-Mediated Peer-Instruction

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

ABSTRACT

TO DO

CCS CONCEPTS

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

KEYWORDS

Peer Instruction, Learnersourcing

ACM Reference Format:
Sameer Bhatnagar, Antoine Lefebvre-Brossard, Michel C. Desmarais, and Amal Zouaq. 2021. Measuring Argument Quality in Technology-Mediated Peer-Instruction. In *Proceedings of LAK '21: Learning Analytics & Knowledge (LAK '21).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.000/0000.0000

1 INTRODUCTION

Technology-mediated peer instruction *(TMPI)* platforms [4][19] expand multiple choice items into a two step process. On the first step, students must not only choose an answer choice, but also provide an explanation that justifies their reasoning. On the second step, students are prompted to revise their answer choice, by taking into consideration a subset of explanations written by their peers for another answer choice. In the case that the student wants to keep their original answer choice, but may be unsure of their own explanation, they are also shown peer-explanations for their original answer choice. The student now has three options:

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

LAK '21, April 2021, UCI,CA

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

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

The design and growing popularity of TMPI is inspired by three schools of thought: firstly, prompting students to explain their reasoning is beneficial to their learning [7]. Second, classroom based *Peer Instruction*[8], 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 [5]. 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" [14], addressing student misconceptions they might not otherwise have thought of.

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

Peer-review platforms most often ask students to provide a score based on a rubric, but the difficulty lies in the ability of novices to generate useful feedback before they have gained expertise with the content. A growing number of peer-review platforms address this issue with pairwise *comparative* judgments. Notable examples include ComPAIR[17] and JuxtaPeer[3], both of which present students with a pair of their peers' submissions, and evaluate them with respect to one another. TMPI falls in this category as well, as students apply a comparative judgment of their chosen explanation, relative only to the subset that was shown to them on their review page. However, one of the challenges that arises in these contexts is how to construct the subset of peer-items in the database which will be presented to the current student.

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

- since each student's "vote" in this context represents an incomplete evaluative judgement (the student will not have seen *all* other peer submissions), which *rank aggregation* methods are best suited for TMPI?

- given that it obtaining the *true* ranking is intractable in TMPI settings (as new data comes in, it is prohibitive to have all explanations evaluated again by all students), what methodological approach ensures the greatest reliability for this *unsupervised* rank aggregation task ?

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

The focus of this study is how we *measure* the quality of artifacts created and curated in learnersourceing environments, as it is a necessary pre-cursor to *modelling* their quality (e.g. once we have established a reliable aggregated *convincingness score*, we can set out to predict that score based on linguistic features of the explanation).

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

2 RELATED WORK

2.1 Learnersourcing student explanations

This modality is a specific case of *learnersourcing*[20], wherein students first generate content as part of their own learning process, that is ultimately used to help their peers learn as well. Notable examples include PeerWise [9] and RiPPLE [13], both of which have student generate learning resources, which are subsequently used and evaluated by peers as part of formative assessment activities.

One of the earliest efforts to leverage peer judgments of peer-written explanations specifically is from the AXIS system[21], wherein students solved a problem, provided an explanation for their answer, and evaluated explanations written by their peers. Using a reinforcement-learning approach known as "multi-armed bandits", the system was able to select peer-written explanations that were rated as helpful as those written by an expert. Our research follows from these studies in scaling to multiple domains, and focusing on how the vote data can be used more directly to model argument quality as judged by peers.

2.2 Ranking Arguments for Quality

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

A classical approach for rank aggregation from pairwise preference data is using the Bradley-Terry model, which has been extended to incorporate the quality of contributions of different annotators in a crowdsourced setting when evaluating relative reading level in a pair passages [6].

When evaluating argument convincingness, one of the first approaches proposed is based on constructing an "argument graph", where a weighted edge is drawn from node A to node B for every pair where argument A is labelled as more convincing than argument B. After filtering example pairs that lead to cycles in the graph, PageRank scores are derived from this directed acyclic graph, and the PageRank scores of each argument are used as the gold-standard to rank for convincingness [12].

More recently, a relatively simpler heuristic WinRate score has been shown to be competitive alternative, wherein the rank score of an argument is simply the (normalized) number of times that argument has been chosen as more convincing in a pair, divided by the number of pairs it appears in [16].

Finally, a neural approach based on RankNet has recently yielded state of the art results. By joining two Bidirectional Long-Short-Term Memory Networks in a Siamese architecture, and appending a softmax layer to the output, [11] show that we can jointly model pairwise preferences and overall ranks publicly available datasets.

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

3 METHODOLOGY

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

3.1 Rank Aggregation

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

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

After this first filtering step, we take the TMPI observations for each question, and construct explanation pairs, as in figure 1.

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

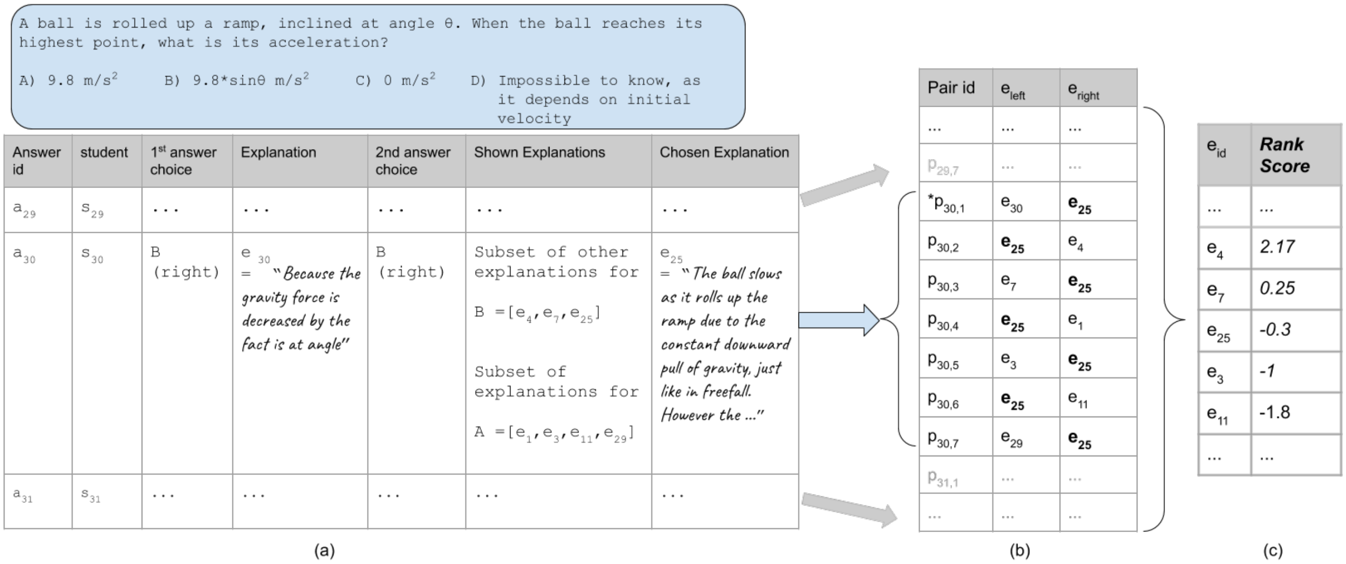

Figure 1: Example of student-item observations from a TMPI environment. (a) Student s_{30} chose the correct B as the answer on their first attempt, and provided the explanation e_{30} in the dataset for this question. Without providing feedback on whether this is the correct answer, the student is shown a subset of explanations from previous students for B, as well as for A (the most popular incorrect answer). The student decides to keep the same answer choice B, and indcates that the explanation $e_2 5$ is the most convincing. This is referred to as a $Right- > Right$ transition. (b) This observation is transformed into 7 explanation pairs. The first pair is for the choice of e_{25} over what the student themself, and the other six are for the choice of e_{25} over the other shown explanations. The pairs are labelled as either having the left or right exlanation being more *convincing*. (c) This pairwise preference data is aggregated global ranked list, where each explanation is assigned a Rank Score.

(2) **BT** score, which is the argument "quality" parameter estimated for each explanation, according to the *Bradley-Terry* model, where the probability of argument A being chosen over argument B is given by

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

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

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

subject to $\sum_i \beta_i = 0$.

(3) The **Elo** rating system[10], which was originally proposed for ranking chess players, has been successfully used in adaptive learning environments (see [15] for a review). This rating method can be seen as a heuristic re-parametrization of the **BT** method above, where the probability of argument A being chosen over argument B is given by

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

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

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

While the **BT** model can be thought of a *consensus* approach, **Elo** ratings are dynamic and implicitly give more weight to recent data[1].

(4) **Crowd-BT** [6] is an extension of the **BT** model tailored to settings where different annotators may have assigned opposite labels to the same pairs, and the reliability of each annotator may vary significantly. A reliability parameter is estimated for each student,

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

where $\eta_k \approx 1$ if the k^{th} student 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

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

and the log-likelilood maximized for estimation to

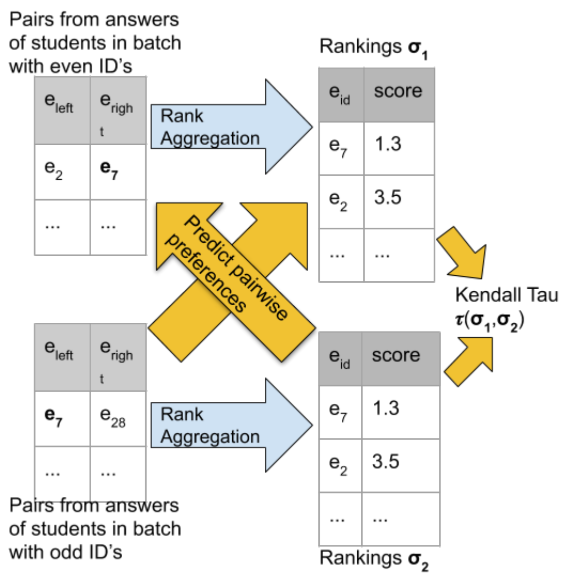

Figure 2: Methodology for evaluation of rank scores

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

(5) **Length**, a method used purely as a baseline, where for each pair, we simply predict that the explanation with more words is the more convincing. This is a commonly used baseline for the pairwise classification task of predicting argument quality [18] has been shown to be competitive for data from learning environments [2]. (Since we only use a basic white-space tokenizer, we round the token-counts of each explanation down to the nearest multiple of five, as it is unlikely that a student could discern which is longer if the difference in lengths is less than this.)

In order to evaluate these rank aggregation different scores, and address our research question, we employ a time-series based cross-validation scheme: at each timestep, we calculate the aggregated argument *convincingness* scores from past students, and set out to predict: which arguments will be chosen as more convincing from the pairs constructed for the current student?

4 DATA

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

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

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

transition	N	N_{pairs}	$wc_{med}(IQR)$
Right -> Right	79816	308509	21 (12)
Right -> Wrong	1340	9587	16 (14)
Wrong -> Right	8373	59541	17 (10)
Wrong -> Wrong	16606	65826	18 (10)

Table 1: Summary statistics of data, aggregated by transition type. N **is the number of student answers,** N_{pairs} **is the number of pairs generated from those answers, and** $wc_{med}(IQR)$ **is the median word count for student explanations, with the inter-quartile range as a measure of dispersion.**

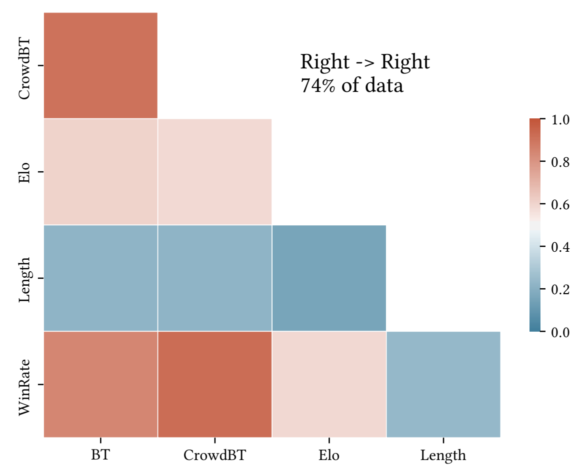

Figure 3: Correlation between different Ranking Scores for each explanation, disaggregated by transition type

5 RESULTS

TO DO

6 DISCUSSION

TO DO

7 LIMITATIONS AND FUTURE WORK

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

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

TO DO

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

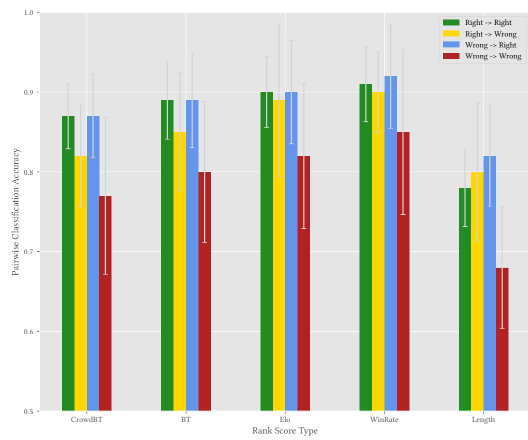

Figure 4: Comparing the classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair. Rank scores are calculated with the vote data of half the students, and tested on the pairs generated by the other half. Data is averaged across all questions, dis-aggregated by different TMPI transition types.

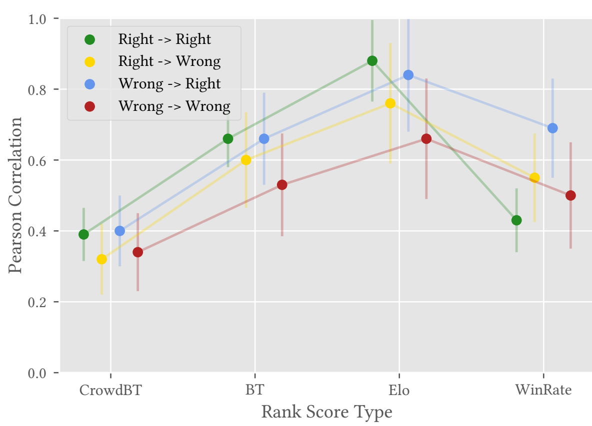

Figure 5: Pearson correlation coefficient between different rank score types, derived from two independent groups of students, averaged over all questions, dis-aggregated by different TMPI transition types.

ACKNOWLEDGMENTS

Funding for this research was made possible by the support of the Canadian Social Sciences and Humanities Research Council *Insight* Grant. This project would not have been possible without the SALTISE/S4 network of researcher practitioners, and the students using myDALITE.org who consented to share their learning traces with the research community.

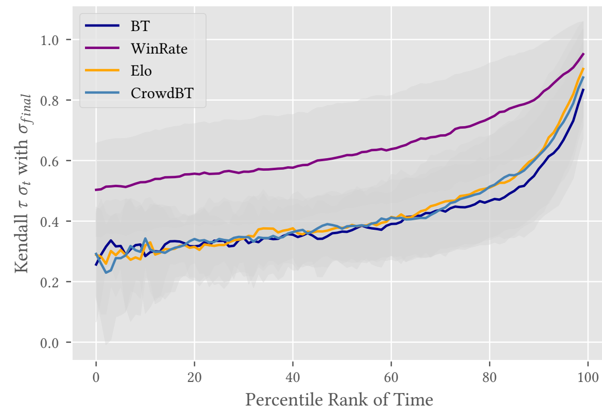

Figure 6: Kendall Tau correlation (τ) between explanation rankings at each normalized time step σ_t, and the final ranking after all students have answered σ_{final}, for different rank score types

REFERENCES

[1] David Aldous. 2017. Elo ratings and the sports model: A neglected topic in applied probability? *Statistical science* 32, 4 (2017), 616–629. Publisher: Institute of Mathematical Statistics.

[2] Sameer Bhatnagar, Amal Zouaq, Michel C. Desmarais, and Elizabeth Charles. 2020. Learnersourcing Quality Assessment of Explanations for Peer Instruction. In *Addressing Global Challenges and Quality Education*, Carlos Alario-Hoyos, María Jesús Rodríguez-Triana, Maren Scheffel, Inmaculada Arnedillo-Sánchez, and Sebastian Maximilian Dennerlein (Eds.). Springer International Publishing, Cham, 144–157.

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

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

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

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

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

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

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

[10] Arpad E Elo. 1978. *The rating of chessplayers, past and present.* Arco Pub.

[11] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. *arXiv preprint arXiv:1907.08971* (2019). https://www.aclweb.org/anthology/P19-1093/

[12] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational

Linguistics, Berlin, Germany, 1589–1599. https://doi.org/10.18653/v1/P16-1150
[13] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. 2019. Ripple: A crowd-sourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522* (2019).
[14] Mitchell J Nathan, Kenneth R Koedinger, Martha W Alibali, and others. 2001. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the third international conference on cognitive science*, Vol. 644648.
[15] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education* 98 (2016), 169–179. Publisher: Elsevier.
[16] Peter Potash, Adam Ferguson, and Timothy J Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*. 146–155.
[17] Tiffany Potter, Letitia Englund, James Charbonneau, Mark Thomson MacLean, Jonathan Newell, Ido Roll, and others. 2017. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry* 5, 2 (2017), 89–113.
[18] Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 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. https://www.aclweb.org/anthology/D19-1564.pdf
[19] Teaching & Learning Technologies Univeristy of British Columbia. 2019. ubc/ubcpi. https://github.com/ubc/ubcpi original-date: 2015-02-17T21:37:02Z.
[20] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 405–416.
[21] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, Edinburgh, Scotland, UK, 379–388. https://doi.org/10.1145/2876034.2876042