A Dataset of Learnersourced Explanations from an Online Peer Instruction Environment

Sameer Bhatanagar
Polytechnique Montreal
sameer.bhatnagar@polymtl.ca

Amal Zouaq
Polytechnique Montréal

Michel C. Desmarais
Polytechnique Montréal

Elizabeth Charles
Dawson College

ABSTRACT

Online *Peer Instruction* has become prevalent in many "flipped classroom" settings, yet little work has been done to examine the content students generate in such a learning environment. This study characterizes a dataset generated by an open-source, web-based homework system that prompts students to first answer questions, and then provide explanations of their reasoning. Of particular interest in this dataset, is that students are also prompted to evaluate a subset of peer explanations based on how convincing they are, as part of the Peer Instruction learning script. Since these student "votes" are then used in the selection of what is shown to future learners, we cast this as an instance of *learnersourcing*, a paradigm that presents new research opportunities for the Learning Analytics community. This study characterizes a dataset from one *Peer Instruction* tool, that includes not only the student generated answers and explanations, but this novel "vote" attribute, which aims to capture how convincing each explanation is to other learners. The dataset includes longitudinal observations of student responses over the course of a semester, following groups from three STEM disciplines. The data is made available to interested researchers[1].

Keywords

datasets, learnersourcing, peer instruction

1. INTRODUCTION

The effectiveness of *Peer Instruction* on learning [4] [25] *in-class*, and the success of Intelligent Tutoring Systems and MOOCs *outside* of class, have in part, led to the development of web-based platforms for *asynchronous* Peer Instruction [6][27]. Recently, other similar learning environments have been developed, centred on having students explain their reasoning, and then evaluate the explanations of their peers[22][5]. The increasing use of this form of online learning exercise implies that a new type of data is being generated, wherein lie opportunities to examine theories of how self-explanation and comparative peer assessment may impact learning.

There are several pragmatic motivations for extending Peer Instruction to out-of class activities. First, when scaling-up *Just-in-Time-Teaching* environments, a web-based platform for asynchronous peer instruction can substantially reduce the time teachers' spend searching the data to identify student misconceptions. Second, when students are asked to compare answers with peers, they receive a form of immediate feedback on their own explanation. Last, when posting threads and sub-threads to large scale on-line discussions, such as MOOCs, an asynchronous Peer Instruction platform offers a more structured alternative and ties student explanations to an answer choice, allowing for more organized aggregation of ideas [2].

These platforms open new research questions and opportunities for the Educational Data Mining community. First, these systems capture new modalities of data, specifically, the written explanations for answer choices, which acts as proxy data: representing the cognitive reflections elicited in conversations students have with peers during small- group in-class Peer Instruction discussions. Second, these environments introduce challenges common to any platform centred on user generated content: quality control and recommendation. The power of having students generating the explanations to different answer choices, and then rating them, enables scaling up of technologies that facilitate flipped teaching practices[14]. However, once these tools *do* scale, sheer volume requires automatic approaches for filtering out low-quality content. Once filtering is complete, recommendation algorithms need to be in place to most effectively help current students navigate the large volume of content generated by past students, with the ultimate objective of optimizing individual learning gains. Further research is needed on *learnersourced* data sets so as to develop best practices that leverage the effectiveness of student written and ranked explanations for adaptive learning experiences, while avoiding the pitfalls that can lead to the valuable data drowning in noise.

2. OBJECTIVES

[1]account required at https://myDALITE.org/signup

Sameer Bhatnagar, Michel Desmarais, Amal Zouaq and Elizabeth Charles "A Dataset of Learnersourced Explanations from an Online Peer Instruction Environment" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 350 - 355

This paper characterizes a dataset generated inside one online platform for asynchronous Peer Instruction, with the aim of identifying the potential research questions and limitations afforded by this novel application. The use of the tool is growing, now reaching over 50 course offerings across at least 5 undergraduate institutions in different science disciplines. The contexts are varied, but the common thread is that instructors are all using the tool as an attempt to increase in-class student engagement with pre-instructional quizzes, and tailoring their lectures based on the free-text explanations students provide for their answers. Thus, the ultimate goal of this study is two-fold:

- introduce a novel source of data to the Educational Data Mining research community, which has the potential to open new lines of inquiry based on the unique "voting" attribute. Students not only write explanations to justify their answer choice to conceptual science questions, but they are asked to choose which of a subset of their peers' explanations are most convincing.
- identify opportunities and challenges related to the design of platforms that rely on *learnersourced* content, such as choosing the most effective content to foster learning; filtering weak or irrelevant student explanations; categorizing and summarizing student explanations for teacher reporting in large classes.

3. BACKGROUND AND RELATED WORK

3.1 Peer Instruction

The interactive engagement technique most relevant to our work here is Peer Instruction: a method for promoting classroom discussion that has been shown to enhance learning [8]. In this common classroom practice, teachers

1. poll their students on a multiple-choice item, using some form of Audience Response System (e.g. clickers),
2. collect the distribution of answers, and maybe even share back with the students,
3. without revealing the correct answer, prompt students to explain their reasoning for their answer choice to a peer nearby, ideally with someone with a different perspective
4. re-poll the students after the small group discussions.

The platform at the centre of this study facilitates an *asynchronous* version of the above script.

3.2 Comparison-based peer assessment

There are other systems similar in design to asynchronous peer instruction; they differ in that the items prompt for open-ended responses, as opposed to multiple-choice questions. However these systems still include a similar *review*-step after submission of an answer, where students are asked to compare and evaluate the quality of the explanations submitted by peers who had already answered the item.

For example, in the ComPAIR system [22], students first submit their written answer to a prompt. They are then shown pairs of their peers' answers, prompted first to give feedback to each of the answers in the pair, and then choose one as the better response. The pairwise comparison at the heart of this tool leverages learners' inherent ability to make judgments regarding an answer's quality *relative to another*, to make up for the lack of expertise usually needed to provide useful feedback on content in isolation. JuxtaPeer [5] is a similar system, where the pairwise comparisons are anchored on one object at a time, and have been shown to improve the quality of feedback that peers can provide to one another.

3.3 Explanations Datasets

Two of the most prominent sources of learning analytics datasets are from the ASSIStments platform[15], and PSLC DataShop[19]. They both provide significant contributions to Learning Analytics and Educational Data Mining researchers, by making available a wide variety of data from different online learning tools. They include datasets with free text responses, including math hints generated by students in ASSISTments, and student explanations to science questions inside the Andes project (hosted in Datashop).

If casting student explanations as *short arguments* in favour of their answer choice, we can look to the Argumentation Mining research community for sample datasets. For example, a dataset of persuasive student essays that are fully annotated for argumentative relations was recently released [26]. The International Corpus of Learner English[12] is used extensively to model how students make arguments.

Another sign of the growing interest in analyzing student generated text are the Automated Essay Scoring[16] and Automatic Short Answer Scoring [17] competitions hosted on the data science platform, kaggle.com. These datasets are still freely available as well.

To the best of our knowledge, none of the above data sources include all of the defining characteristics that are generated by online Peer Instruction, such as the student's initial answer choice and explanation, a student's second answer choice after having reviewed peer explanations, and most importantly, the peer explanation the student found most convincing.

3.4 Learnersourcing explanations

Web-based homework systems are effective because students get immediate feedback as to whether they answered correctly. However, as the number of question items grows, as well as associated answer choices, generating high quality explanations that help different types of learners resolve different sets of misconceptions, is impractical for teachers [14]. Moreover, explanations written by content experts may also suffer from the *expert blind spot*, wherein their high level of familiarity with the subject matter actually might actually make their explanations more difficult to understand to novices [20].

The concept of *learnersourcing* is a sub-type of *crowdsourcing*, wherein domain novices contribute to the human computation workflow as part of their learning process [28]. PeerWise [10] is an environment within which students make their own questions, and share them with peers, along with accompanying solutions. RiPPLE is a tool that follows the same model, but adds an adaptive recommendation engine [18].The AXIS system [29] prompts students to provide ex-

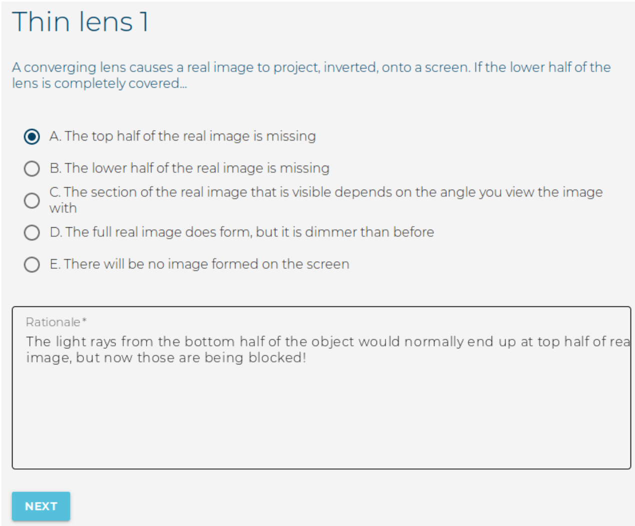

Figure 1: Asynchronous Peer Instruction - Step 1, screenshot of student answering multiple-choice question, and explaining their thinking inside a text box

planations to their answers, rate the explanations of their peers, and then machine learning to curate these to a constantly evolving set of explanations that optimize for promoting student learning. ASSISTments, another widely used learning platform, developed the PeerASSIST plugin [24], which asked students to write explanations to their answer submissions, to be used as hints for future students.

4. MYDALITE: PLATFORM AND DATA

4.1 The Platform

dalite-ng is an open-source project [23] that has been in active development since 2013, and has been used in MOOCs as well as on campus course offerings. myDALITE.org is one instance of this code-base, offered as a hosted service that is free to all teachers and students. It is maintained by a network of learning science researchers and practitioners, whose mission is to promote the uptake of student-centred active learning pedagogical practices. Teachers sign up, author their own questions, and distribute to their students at their discretion. The script for the student completing a question item in *dalite-ng* is:

1. **Question start**: student is presented with a multiple-choice question. They are asked to choose an answer choice and enter a free text response to explain their reasoning.
2. **Question review**: without indicating whether the student chose the correct answer, the tool reflects back to the student their own choice, and the explanation they just entered. They are then prompted to reconsider their answer, by reading the explanations of other students. In the top half of the page, they are shown up to 4 other explanations by students who chose the same answer choice. In the second half of the page, they are then shown up to 4 more explanations to a different answer choice. Students must indicate which is their second answer choice in this review step, by selecting one of these explanations. They also have the option of selecting their own explanation as the most convincing. There are several factors that go into the selection of what the students are shown here:
   - if the student answered incorrectly on the first step, the explanations in the second half of the page will be for the correct choice
   - if the student did in fact answer correctly on the first step, the explanations in the second half of the page will be for the most popular incorrect answer.
   - There are two different heuristics for the selection of explanations for each answer choice:
     – Random, which is useful for when a question is newly introduced to the database, and not enough students have answered to reliably estimate which answers are most *convincing*
     – preferentially selecting from explanations that have already been chosen as convincing
3. **Question summary** The entire flow of information is reflected back to the student for review: their first answer, their own explanation, their second answer

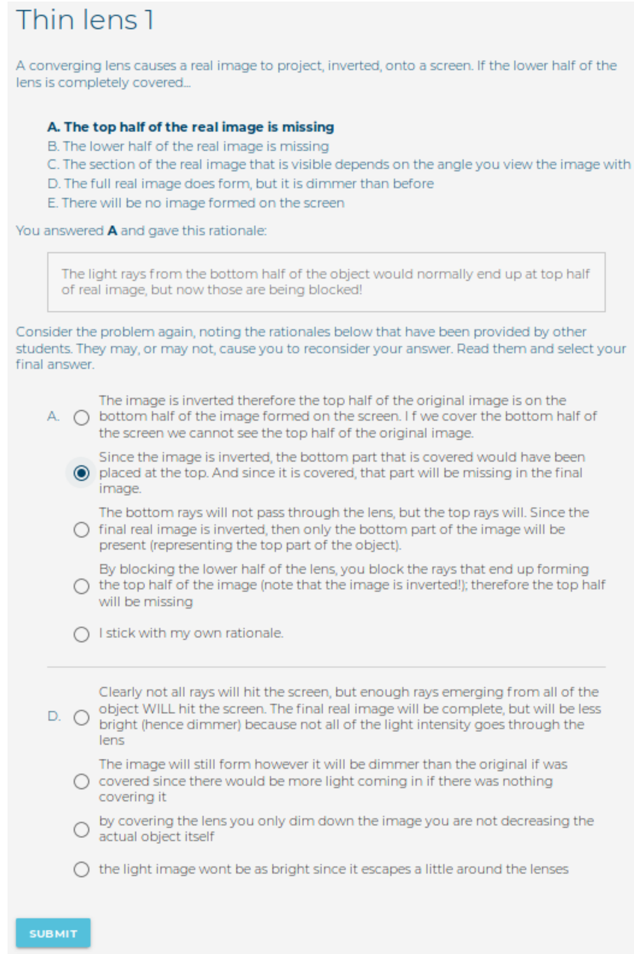

Figure 2: Asynchronous Peer Instruction - Step 2 , screenshot of student choosing a peer's explanation of a different answerchoice

choice and the associated explanation that they chose as most convincing. The correct answer is also finally revealed.

4.2 Data Collection

The data in this study comes from the 2018-2019 academic year, wherein the platform was more heavily used than ever before, due to additional on-boarding support offered to teachers by the host network. All teachers who make question items on the platform must release their content under Creative Commons licenses, and are made aware that the learning-data generated by the students in their groups may be used for academic research. Students are advised upon signing in, that their learning traces, in anonymized form, may be used for research, and that if they do not wish to share their data, they can revoke their consent at any time, without any impact on academic standing in their courses. The data gathered for this study spans three STEM disciplines where there happened to be the most activity: Biology, Chemistry, and Physics. There are many different groups of students in Physics and Chemistry, each with a different teacher (although all in undergraduate level courses), while all of the data in Biology comes from one large freshman group that used the tool very heavily. In the case of a few groups, the items were assigned by teachers as optional, not-for-credit items, meant to provide extra practice study exercises (this information is provided in the meta-data file of the dataset). For those cases when myDALITE was used for credit, students received 0.5 marks for choosing the correct answer on their first attempt, and 0.5 marks for choosing the correct answer choice after the review step. No credit was ever assigned based on a formal expert evaluation of the student explanations.

4.3 Dataset

Each record in the dataset is comprised of the following fields:

- anonymized student identifier
- anonymized group/course identifier (with meta-data on whether the activities were assigned for credit of not)
- question prompt text (and any associated images)
- student's first answer choice
- student's explanation for their first answer choice
- peer explanations shown to student on second step
- student's second answer choice
- the peer explanation they selected as most convincing for their second answer choice
- timestamps associated with
  - when the student first opened the problem
  - when the student entered their first answer choice, and associated explanation
  - when the student entered their second answer choice, and associated peer explanation

Certain filters were applied for the purposes of data extraction for this study. The only groups that were retained were those having 10 students or more, each of whom having answered at least 10 questions. The only disciplines that were included in the current dataset were ones with over 10,000 student responses.

Table 1: Size of dataset across disciplines

	N_g	N	N_q	N_a	$\overline{N_a}$
Biology	1	346	232	19653	57
Physics	16	1250	572	50286	40
Chemistry	16	1055	532	28319	27

Table 2: Relative number of answer transitions, from step 1 to step 2

	$1^{st}C$	Δ			Δe
		$\sum$	r→w	w→r	
Biology	0.70	0.10	0.01	0.09	0.40
Physics	0.79	0.09	0.01	0.07	0.44
Chemistry	0.69	0.12	0.01	0.11	0.38

5. DESCRIPTIVE STATISTICS

As can be seen in Table 1, there are relatively similar numbers of responses across the three disciplines.

N_g : number of groups
N : number of unique students
N_q : number of items
N_a : total number of answers by all students
$\overline{N_a}$: average number of items completed by each student

This table demonstrates the valuable longitudinal nature of the dataset, in that across the disciplines, there are, on average, more than 25 observations per student, which could help building a more robust learner models.

In Table 2, we see the proportion of times students changed their answers on the answer review step.

- $1^{st}C$: fraction of responses where students chose the correct answer choice on their first attempt
- Δ : fraction of responses where students switch their answer choice on review step
  - $\sum$: total fraction of answers where students changed their answer choice from step 1 to step 2
  - $r \rightarrow w$: fraction of responses where students switch their answer choice on review step, going from right to wrong
  - $w \rightarrow r$: fraction of responses where students switch their answer choice on review step, going from wrong to right, presumably after reading their peers' explanations
- Δe : fraction of responses when students do not change their answer choice on review step, but choose an explanation other than their own as most convincing

Across the disciplines, the items in this dataset are easy enough for students to choose the correct answer choice on their first attempt almost three out of four times. The explanations of their peers are almost never able these convince students to switch from the right answer choice to a wrong one. However of the students who choose the wrong answer on their first attempt, after having access to the explanation of their peers, these students switch to the correct answer choice at the review step almost one out of three times.These

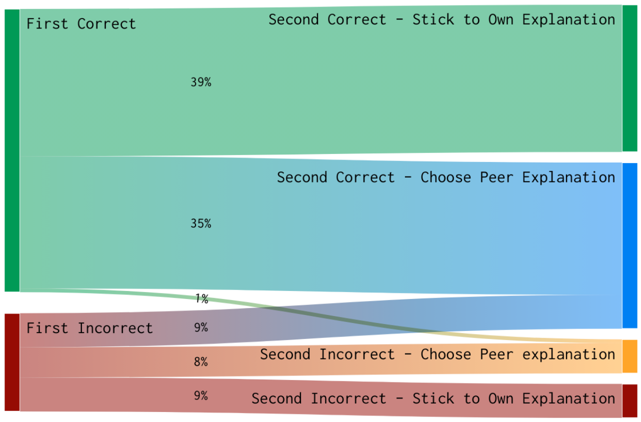

Figure 3: Visualizing student transitions in asynchronous peer instruction with a Sankey diagram

Table 3: Descriptive statistics on explanation word counts

	$\overline{WC}$	$WC < 5$	$\Delta e_{\text{longest}}$
Biology	7	0.63	0.19
Physics	15	0.42	0.23
Chemistry	19	0.19	0.18

relative transitions are visualized in the Sankey diagram in Figure 3.

In table 3,

- $\overline{WC}$ is the average word count of the explanations
- $WC < 5$ is the proportion of explanations that have less than 5 words
- $\Delta e_{\text{longest}}$ the proportion of times students selected the peer explanation that had the most words amongst those that they were shown.

Here in Table 3, we see that many students write explanations that are too short to form a sentence in the Biology subset, and that even in the other disciplines, the explanations are not long-form persuasive essays, but likely closer to short answers. However, students seem to show a preference for explanations that are longer in length when "voting" for the most convincing explanation on the review step.

6. DISCUSSION

Learnersourcing shows immense potential for scaling up online Peer Instruction, but also presents new challenges common to contexts centred on user-generated content.

A quick sampling of the large number of explanations with less than 5 words likely indicates that students do not see the value of writing explanations, unless they will receive course credit for the task. Work from the argument mining community may be useful here to automatically assess the quality of explanations. Under study is the impact of web-based reputation systems on increasing student engagement, which have been shown to increase engagement in learning environments by offering virtual achievement rewards, such as badges and leaderboards[9]. Another open research question is in automatic quality control, given that the first few students who complete a question, and submit an explanation, will have their work shown to many subsequent students. Work that has been done on automatic filtering [11] of explanations based on unsupervised clustering could prove beneficial here.

The value and uniqueness of this dataset remains in the "voting" data: modelling what linguistic properties and conceptual constructs students find convincing, in the language of their peers, is fertile ground for research. The longitudinal data also allows for modelling the evolution of how students start integrating domain specific concepts into their explanations across a semester, as well as "voting" for them in the peer-explanations they find most convincing.

6.1 Future Work

Work must now be done on better understanding how to optimize the heuristics that select what peer explanations are shown to students in order to enhance learning. This will require building student models of ability and models of item difficulty. The linguistic properties are also of key interest: can this mode of comparative peer assessment data be used to inform our models of whether students have attained domain literacy? Finally, how do such environments promote student engagement in *flipped classroom* contexts? We look forward to collaborating with the community through this novel source of data to along these lines of research.

Many of the design/implementation decisions for these platforms are made with pragmatic motivations in mind and need to be better informed by learning analytics theory. The platform at the center of this study is a model to examine more closely also because it is an open-source project, developed as part of Research Practice Partnership [7], where learning analytics researchers are actively working with instructors using the tool to better align teaching practices with sound pedagogical design.

7. ACKNOWLEDGMENTS

Funding for the development of myDALITE.org is made possible by *Entente-Canada-Quebec* program, 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. Special thanks to Dr. Jonathon Sumner for help with data visualization. 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

[1] S. Bhatnagar, M. Desmarais, C. Whittaker, N. Lasry, M. Dugdale, and E. S. Charles. An analysis of peer-submitted and peer-reviewed answer rationales, in an asynchronous Peer Instruction based learning environment. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, editors, *Proceedings of the 8th International Conference on Educational Data Mining*, 2015.

[2] S. Bhatnagar, N. Lasry, M. Desmarais, and E. Charles. DALITE: Asynchronous Peer Instruction for MOOCs. In *European Conference on Technology Enhanced Learning*, pages 505–508. Springer, 2016.

[3] J. L. Bishop, M. A. Verleger, and others. The flipped classroom: A survey of the research. In *ASEE national conference proceedings, Atlanta, GA*, volume 30, pages 1–18, 2013.

[4] J. E. Caldwell. Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, 6(1):9–20, 2007.

[5] J. Cambre, S. Klemmer, and C. Kulkarni. 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*, pages 1–13, Montreal QC, Canada, 2018. ACM Press.

[6] E. S. Charles, N. Lasry, C. Whittaker, M. Dugdale, K. Lenton, S. Bhatnagar, and J. Guillemette. Beyond and Within Classroom Walls: Designing Principled Pedagogical Tools for Student and Faculty Uptake. International Society of the Learning Sciences, Inc.[ISLS]., 2015.

[7] C. E. Coburn and W. R. Penuel. Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, 45(1):48–54, 2016.

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

[9] P. Denny. The Effect of Virtual Achievements on Student Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 763–772, New York, NY, USA, 2013. ACM. event-place: Paris, France.

[10] P. Denny, J. Hamer, A. Luxton-Reilly, and H. Purchase. PeerWise: Students Sharing Their Multiple Choice Questions. In *Proceedings of the Fourth International Workshop on Computing Education Research*, ICER '08, pages 51–58, New York, NY, USA, 2008. ACM. event-place: Sydney, Australia.

[11] V. Gagnon, A. Labrie, M. Desmarais, and S. Bhatnagar. Filtering non-relevant short answers in peer learning applications. In *Proc. Conference on Educational Data Mining (EDM)*, 2019.

[12] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot. *International corpus of learner English*. 2009.

[13] R. R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, 66(1):64–74, 1998.

[14] N. Heffernan, K. Ostrow, K. Kelly, D. Selent, E. Inwegen, X. Xiong, and J. Williams. The Future of Adaptive Learning: Does the Crowd Hold the Key? *International Journal of Artificial Intelligence in Education*, 26, 2016.

[15] N. T. Heffernan and C. L. Heffernan. The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[16] Kaggle. The Hewlett Foundation: Automated Essay Scoring, 2012.

[17] Kaggle. The Hewlett Foundation: Short Answer Scoring, 2013.

[18] H. Khosravi. Recommendation in personalised peer-learning environments. *arXiv preprint arXiv:1712.03077*, 2017.

[19] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, 43:43–56, 2010.

[20] M. J. Nathan, K. R. Koedinger, M. W. Alibali, and others. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the third international conference on cognitive science*, volume 644648, 2001.

[21] G. M. Novak. Just-in-time teaching. *New directions for teaching and learning*, 2011(128):63–73, 2011.

[22] T. Potter, L. Englund, J. Charbonneau, M. T. MacLean, J. Newell, I. Roll, and others. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry*, 5(2):89–113, 2017.

[23] SALTISE. SALTISES4/dalite-ng, June 2019. original-date: 2018-01-11T16:36:55Z.

[24] D. Selent. *Creating Systems and Applying Large-Scale Methods to Improve Student Remediation in Online Tutoring Systems in Real-time and at Scale*. PhD thesis, Worcester Polytechnic Institute Polytechnique Institute, June 2017.

[25] M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su. Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910):122–124, 2009.

[26] C. Stab and I. Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In *COLING*, pages 1501–1510, 2014.

[27] T. . L. T. Univeristy of British Columbia. ubc/ubcpi, Aug. 2019. original-date: 2015-02-17T21:37:02Z.

[28] S. Weir, J. Kim, K. Z. Gajos, and R. C. Miller. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 405–416. ACM, 2015.

[29] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*, pages 379–388, Edinburgh, Scotland, UK, 2016. ACM Press.