How do A/B Testing and secondary data analysis on AIED systems influence future research?

Nidhi Nasiar[1], Ryan S. Baker[1], Jillian Li[1], Weiyi Gong[1]

[1] Graduate School of Education, University of Pennsylvania
nasiar@upenn.edu

Abstract. Recent years have seen a surge in research conducted on intelligent online learning platforms, with a particular expansion of research conducting A/B testing to decide which design to use, and research using secondary platform data in analyses. This scientometric study aims to investigate how scholarship builds on these two different types of research. We collected papers for both categories - A/B testing, and educational data mining (EDM) on log data- in the context of the same learning platform. We then collected a randomized stratified sample of papers citing those A/B and EDM papers, and coded the reason for each citation. On comparing the frequency of citation categories between the two types of papers, we found that A/B test papers were cited more often to provide background and context for a study, whereas the EDM papers were cited to use past specific core ideas, theories, and findings in the field. This paper establishes a method to compare the contribution of different types of research on AIED systems such as interactive learning platforms.

Keywords: Scientometrics, A/B testing, Online Learning, AIED systems.

1 Introduction

1.1 Research on Interactive Learning Platforms

Large-scale platforms for interactive online learning have become a core part of educational practice, a trend that has accelerated due to the pandemic-related shutdowns of educational institutions. There are several benefits of interactive learning platforms for learners. They make learning significantly more accessible [29] for learners unable to travel, learners whose work constraints make class attendance infeasible, and learners at home in quarantine. They are often also beneficial even when learners can attend class in-person, enabling classroom instruction to be enhanced by using data from online activities given as homework or in-class [34, 37]. Research-based platforms such as intelligent tutoring systems tend to lead to substantial learning benefits, an average of 0.76 standard deviations better than traditional curricula [33].

Even beyond these benefits, AIED learning platforms provide opportunities for enhancing learning through research [32] and can support it by iterative refinement through A/B tests and secondary data analysis. A large number of automated experiments have been conducted on these online learning platforms. Initially, it was common for single research groups to use their own platforms for research [2, 25]. In the early 2000s, the Pittsburgh Science of Learning Center (PSLC) built an

infrastructure enabling hundreds of studies to be conducted in classrooms [20], albeit in a relatively resource-intensive fashion where researchers visited individual classrooms. In recent years, the ASSISTments learning platform has developed a research platform that allows automatic deployment of studies across the web. This platform has been used by dozens of external researchers to carry out their studies in thousands of math classrooms [27]. Increased support for A/B studies has also been incorporated into MOOC platforms [30], leading to large-scale studies such as [18], which tested an intervention in over 200 courses with millions of enrolled learners.

There has been an even larger expansion in the use of AIED learning platform data in secondary analyses by educational data mining (EDM) researchers. Initial research within the educational data mining conference was heavily based on data sets from the PSLC [19], with 14% of total analyses using DataShop data [1]. Over time, a range of learning platforms have moved towards sharing their data publicly, increasing the number of research questions that can be investigated by researchers without direct access to a large-scale platform. Specific data sets have become standards for comparing algorithms across papers – for instance, many papers have used a specific public data set from ASSISTments, to study student knowledge modeling [17, 38, 39], and Cognitive Tutor data has been used to compare ways to automatically refine knowledge structures [14, 22].

Both A/B testing infrastructure and secondary data analyses have facilitated and expedited research in the learning sciences, but the full details of how these trends have impacted the field are not fully known. We know there are more papers, but how do these papers influence the field? And do these two innovations influence future research in similar ways or do they have different types of influence? In this paper, we investigate the question of how the research afforded by these learning platforms impacts scientists and projects even beyond the specific papers that are produced. In other words, what is the scientific impact of each type of research, and is there a different impact on the science of learning from A/B tests versus EDM analyses?

1.2 Scientometrics in secondary data analysis

In answering this question, we draw upon methods and past work in scientometrics, the field of scientific study which investigates the properties of scientific publications in order to better understand science more broadly. One of the core and long-standing questions and contributions of scientometrics has been in terms of comparing papers in terms of citation counts [15, 31] and comparing the relative contribution of different scientists [6]. This has been a prominent area of analysis in the learning analytics community. For example, research studies have looked at what learning analytics and EDM papers are most cited [1, 8, 36], and have analyzed the quantity of research output and collaboration in order to rank universities and scholars [11, 36]. This work has been highly useful to researchers in understanding the state and scope of the field of learning analytics. However, it does not answer our current research question around how this field makes progress scientifically.

A second category of scientometric research in EDM has focused on which topics are being studied, and how these EDM topics have shifted over time [8, 9], building on

similar long-standing trends in scientometrics more broadly [4]. Furthermore, researchers have looked at the differences between the topics studied in learning analytics and educational data mining [5, 10], which sub-community's papers are cited more often [5, 10], and the relationships between published topics [36].

A third category of existing scientometric research in learning analytics has investigated equity in the field's practices. Concurrently with an increase in interest within scientometrics more broadly in whether gender, race, and ethnicity influence publication and citation patterns [16], learning analytics researchers have investigated the diversity in the field [5, 24, 36]. Recent work has also studied the degree to which diversity in samples is considered in secondary data analytics research (or even reported) [28]. The results of [28] indicated that most papers in the field do not even mention the background of learners, much less check for algorithmic biases, which makes it challenging to gauge the generalizability and transferability of our findings.

However, despite the considerable interest in scientometrics within communities closely aligned with the AIED community, there has not yet been research on analyzing citations to understand how researchers in these communities build on each others' research or on why papers are cited. In other words, there has been research on who is conducting research in these communities, and what they are researching, but not how they are building upon each others' research. Fortunately, there is considerable work in the scientometrics community that we can build on in analyzing this question for EDM and A/B testing research. Starting with [12], scientometricians have attempted to identify lists of reasons for why a scholar might choose to cite a specific paper. [3] expanded upon a list by Garfield [12] in an extensive review, which [21] then distilled into a manageable coding scheme. In this literature, one of the key steps towards understanding why a citation occurs was developing methods for the qualitative analysis of a citation's context [3, 7]. This literature found that researchers choose to cite a paper for a wide variety of reasons, including both scientific reasons (crediting key past contributions, refuting previously published ideas) and political reasons (citing an important member of the field, citing papers from the venue being submitted to). Political citations can be quite common – for example, a review of citations in computer science education found that few citations actually involved building on the contributions in previous papers [23].

In this paper, we built on this past work to investigate our research question of why researchers cite EDM and A/B testing papers, and what the differences are between the citations to each type of paper. We do so by collecting a corpus of citations of work to each type (citations all to work occurring in the same learning platform, to reduce confounds), qualitatively coding the reasons for each citation, and then statistically comparing the proportion of each reason for citation.

2 Methods

2.1 Research Context

In this paper, we analyze the citations received by papers presenting research conducted in the context of the ASSISTments platform [28]. ASSISTments is an online learning

system with 500K students and 20K teachers currently, primarily used for mathematics. ASSISTments has users in over 20 countries, but the majority of learners are in the United States of America. Randomized controlled studies have demonstrated positive learning gains for students using the platform on an ongoing basis [26]. Learners using ASSISTments complete mathematics problems, and can receive multi-step hints or scaffolding on demand or after making errors. ASSISTments provides support for mastery learning, where learners continue working on a skill until they demonstrate they can answer correctly three times in a row, and offers spiraling practice/review functionality as well.

Among AIED learning systems, ASSISTments offers substantial support for external researchers. Learning analytics and educational data mining researchers are able to download (as of this writing) fourteen publicly available data sets named Open Released Datasets[1]

, which offer extensive interaction log data, combined in some cases with additional data such as field observations of student affect or longitudinal student outcomes. Dozens of external researchers have used data from the ASSISTments system in further analyses.

ASSISTments also offers substantial support for A/B testing research, enabling a researcher to conduct randomized experiments on learners across the United States, using E-Trials, the Ed-Tech Research Infrastructure to Advance Learning Science [41]. A substantial number of external educational psychology and learning sciences researchers have used the ASSISTments system to conduct A/B tests on a wide range of research questions. The large scale of ASSISTments' use in both learning analytics and A/B testing research makes it an appropriate context to compare the scientific impact of these two types of research.

2.2 Articles Studied

In this study, we compared the types of scientific impact achieved by two categories of papers, referred from here onwards as the "target" papers. We selected all the papers published up until March 2021 (when we pulled our data set for analysis) that leverage the ASSISTments platform for conducting the two different kinds of research. We filtered out the papers which did not fall into either category.

The first type of papers (referred to as A/B papers) compare the impact of two versions of a learning activity within the ASSISTments system. For the A/B papers, students are experimentally assigned to one condition or the other, to evaluate the impact of intervention on student learning or other outcomes.

The second type of papers (henceforth referred to as secondary data analysis or EDM papers), use interaction log data from the ASSISTments system to investigate a range of research questions, including the impact of different behaviors on student outcomes, the accuracy of different knowledge modeling algorithms, and the linguistic attributes of ASSISTments math problems.

[1] The open released data sets are publicly available at https://www.etrialstestbed.org/resources/featured-studies/dataset-papers

All the target papers for both categories were obtained from the publicly available ASSISTments website, which provides a list of papers that use their Open Released Datasets, as well as a repository of all the published randomized controlled experiments using ASSISTments. This yielded a total of 27 target A/B papers, and 32 target EDM papers. In March 2021, we used Google Scholar to obtain a list of papers citing each of these target articles. An article was considered if the full text could be obtained either openly over the internet, through the University of Pennsylvania library, or through interlibrary loan. Both peer-reviewed and non-peer-reviewed (i.e. dissertations, xArxiv, white papers) documents were included. Only articles in English were considered for the review process. Duplicates were filtered out if a single paper was citing one target paper more than once, however, if a single paper was citing different target papers multiple times, then each citation was considered separately. This gave a total of 2418 citations across all of the target papers (756 total citations for A/B papers, or 28 per paper; 1662 total citations for EDM papers, or 51.9 per paper).

We conducted statistical power analysis in order to determine how many citing papers to sample from this large number of articles for qualitative coding. An initial analysis of the citations of two highly-cited papers was used to choose parameters for the statistical power analysis. Statistical power was calculated using G*Power 3.1.9.4, assuming an effect size where papers in one category were cited 50% more often for one reason than the other paper category, with a baseline of 40% citation for the less common reason (i.e. 40% versus 60%; risk ratio = 1.5), with the allocation ratio set to one (i.e. we will sample approximately the same number of papers of each type), and α set to 0.05, using the Z test of the significance of the difference between two independent proportions (this test is mathematically equivalent to $\chi 2$ with one degree of freedom – they provide the exact same p values). For this test, statistical power of 0.8 would be achieved with samples of 97 and 97. Given this goal number of papers, we conducted stratified random sampling (stratified to equalize the number of citing papers per target paper as much as possible). This resulted in a data set of 174 papers citing A/B papers and 167 papers citing EDM papers for coding, moderately larger than the goal sample size.

2.3 Coding Scheme

We identified all the citations of any target papers within each article that cited one or more of the target papers. In many cases a citing article cited multiple target papers, in most cases all from the same type of paper (A/B or EDM) and in exactly one case from both.

Next, we developed a coding scheme to identify the reasons why a citation might cite an article. Our first step towards developing this coding scheme was to take an extensive list of reasons why people cite published articles [21], which had been distilled from a review of 30 studies on citing behavior [3]. We then eliminated reasons not found in our citing articles or that would not be explicitly stated in the text surrounding a citation. For instance, [21] notes that authors may choose which paper to cite based on the availability of full text for that paper, a reason that would be difficult to identify from how a paper is cited within the text. We then removed or merged

categories that were not clearly differentiated from each other, and categories that did not seem to occur in our papers. This yielded our final coding scheme for citations. As will be noted below, not all of the categories we chose to code were ultimately found in our sample of citing papers. The final coding scheme was:

Publication-Dependent Reasons

Citation due to some attribute of the publication being cited (in the target article)

P1: The target paper was the original publication in which an idea or concept was discussed – a "classic" article.

P2: Using/giving credit to ideas, concepts, theories, methodology, and empirical findings by others.

P3: Earlier work on which current work builds.

P4: Providing background, to give "completeness" to an introduction or discussion.

P5: Empirical findings that justified the author's own statements or assumptions.

P6: Refuting or criticizing the work or ideas of others.

P7: Mentions of other work ("see also", "see for example", "cf", "e.g.", "i.e.") without further discussion.

P8: Used target paper's dataset for secondary analysis

Author-Dependent Reasons

Citation due to some attribute of the author being cited (in the target article).

A1: Paying homage to a pioneer in the research area/giving general credit for related work.

A2: Ceremonial citation, the author of the cited publication is regarded as "authoritative".

A3: Self-citation: one of the authors was also an author on the target article

Note that this coding scheme is not exhaustive; some citations may not be coded as representing any of these categories (for instance, articles cited as a part of the systematic review of studies) for both types of paper.

Initially, a subset of citations for each target paper was coded[2] in terms of this coding scheme by two coders (the first and third authors), to establish inter-rater reliability, and then the first author coded all the papers. If a coder judged that a paper was cited for multiple reasons – for instance, if it was cited in different parts of the paper – multiple codes were given. However, if a citing paper cited the same target paper multiple times for the same reason, it was counted a single instance – i.e., if the citing paper cited the target paper for reason P2 in four different places, it was treated as a single citation because of reason P2.

The proportion of each citation category found across citing papers was compared using the chi-squared test, between the two types of target papers (i.e. A/B versus EDM). Both Bonferroni and Benjamini & Hochberg corrections were applied (separately).

Inter-rater reliability (Cohen's Kappa) was calculated for each coding category, treating each category as independent (i.e. a set of binary codes) since coding was non-

[2] The data set created is publicly available at https://osf.io/rmswe/?view_only=d496417aef1e4046907d2271b8a86cbb

exclusive. The average Kappa across categories was 0.77 for A/B and 0.72 for EDM, 0.75 overall. Kappa was above 0.6 for every category. Categories P1, A1, and A2 were never coded for any citation by either of the two coders. For A1 and A2, this might be due to difficulty in identifying an author-dependent reason for citation from the text of the paper; much of the research on author-dependent reasons for citation has involved self-report rather than content analysis ([36, see review in [3]). The lack of application of P1 may similarly be due to the difficulty of identifying it from the paper text. Although the original reason for citing a paper may be its classic status, the practice of academic writing may result in a paper being discussed in terms of a different reason.

3 Analysis and Results

After inter-rater reliability was established, the first coder coded every citation in every paper. We next analyzed the prevalence of each citation category for each type of paper, and whether the prevalence of any citation category was statistically significantly different between the two types of papers. As mentioned above, within analysis we considered each citing paper/reason combination only once for each target paper, even if a target paper was cited for the same reason more than once in the same citing paper.

Table 1 shows that the most common citation category, for both papers, was P2, using/giving credit to specific ideas, concepts, theories, methodology, and empirical findings by others. It was seen in around more than half of the citations (averaged at the level of citing papers) for target EDM papers, and 35.6% for A/B papers. P4 appeared in a substantial 32.2% of citations for A/B papers, and about half of that in EDM papers. Two categories were seen between 15% and 25% of the time for both types of papers: P3, Earlier work on which current work builds, and A3, Self-citations. The remaining three categories were seen less than 10% of the time for both papers.

Table 1. The prevalence of different Citation Categories for each of the two paper types

Reason for Citation	Average Prevalence (paper AB)	Average Prevalence (paper EDM)	p value
P2: Using/giving credit to specific ideas, concepts, theories, methodology, and empirical findings by others.	**35.6%**	**58.1%**	**0.00003**
P3: Earlier work on which current work builds.	18.4%	15.6%	0.49
P4: Providing background, to give "completeness" to an introduction or discussion.	**32.2%**	**16.2%**	**0.00057**
P5: Empirical findings that justified the author's own statements or assumptions.	9.8%	6.0%	0.20

P6: Refuting or criticizing the work or ideas of others	1.2%	3.6%	0.14
P7: Mentions of other work (“see also”, “see for example”, “cf”, “e.g.”, “i.e.”) without further discussion.	8.0%	9.0%	0.76
P8: Used target paper’s dataset for secondary analysis.	4.0%	1.8%	0.22
A3: Self-citation	19.5%	24.6%	0.35

Statistically significant differences between the two paper types are given in boldface.

We then compared the prevalence of each citation category between paper A/B and paper EDM using a chi-squared test. This test assumes that paper A/B and paper EDM are cited by different sets of papers. In practice, only 1 paper in our sample cited both of these two categories of papers (out of a total of 341 papers), so this seemed like a safe assumption rather than a situation where a significantly more complex method tailored to partial overlap of data sets would be warranted. The statistically significant categories are P2 and P4. Category P2 stands for using/giving credit to specific ideas, concepts, theories, methodology, and empirical findings by others, which was cited 35.6% of the time by A/B papers, and 58.1% by the EDM papers, $\chi 2$ (df =1, N=341)=17.26, p=0.00003. Category P4 represents providing background, to give “completeness” to an introduction or discussion, and it was about twice as commonly cited in A/B papers (32.2%) than in the EDM papers (16.2%), $\chi 2$ (df =1, N=341) = 11.87, p=0.0005. The full pattern of statistical evidence is given in Table 1.

There is an inflated risk of Type I error since we ran eight statistical tests. To address this risk, we applied Benjamini & Hochberg and Bonferroni post-hoc controls. No significant tests became non-significant after the post-hoc test. Categories P2 & P4 were found to have $p < 0.001$, so they remain significant after post-hoc control. All other tests were non-significant, even without a post-hoc correction.

Conclusions and Discussions

In this study, we have investigated the reasons behind why scientists cited two types of papers using AIED systems for research. One category of papers used the platform to conduct automated A/B tests, the other category of papers used the platform’s data to do secondary learning analytics (EDM) research.

We distilled a list of eleven reasons on why a paper is cited from prior literature on scientometrics, and then applied this list of reasons (as citation categories) to a sample of papers that cited one of the two types of target papers, within the same learning platform, with two coders who established inter-rater reliability for each code. Within

this learning platform, the EDM papers were cited almost twice as much as the A/B papers, which may reflect several factors, including the relative contribution of each type of work, the ease in building on work of each type, or the size of the large and flourishing learning analytics research community.

In our findings, both types of papers were cited primarily for publication-based reasons rather than author-based reasons (except for self-citation). However, this may simply be due to the difficulty in identifying author-based reasons for citation. For example, a paper may have been cited because of its author's political power, but that citation may then be justified within the paper in terms of some scientific aspect of the paper, such as category P7 (citations to a paper as an example of some more general category, without further discussion). As such, determining if a citation is author-based probably depends on other forms of data collection such as anonymous surveys [32].

In comparing the two types of articles, two statistically significant differences were found: the EDM type of papers were cited for reason P2 (Using/giving credit to specific ideas, concepts, theories, methodology, and empirical findings by others) over 50% of the time, which was 1.6 times more than A/B papers cited for that reason. This finding suggests that EDM papers are more prevalent in generating ideas, concepts, and empirical findings that other researchers in the field find useful. This type of research directly contributes to the field moving forward.

On the other hand, category P4 (Providing background, to give "completeness" to an introduction or discussion) was cited as a reason twice as many times by the A/B papers than the EDM papers. These citations were primarily found in the 'Introduction' or the 'Literature Review' section of the papers. The findings might indicate that A/B papers are being cited for related work, and to cover the breadth of the research related to that topic, instead of directly building on previous work.

Overall, these findings seem to highlight the different types of contributions the two types of papers make – EDM type of papers seem to have a larger impact on subsequent research than A/B papers. A/B research studies seem to be carried out more independently from prior work. One possible explanation for this pattern can be because the range of potential design modifications is large and varies based on the original design of the system being studied, whereas EDM algorithms tend to either compare algorithms (directly using previous work) or develop an analysis across papers (like work on defining wheel-spinning and studying it). It is also possible that as the community of learning platform A/B researchers develops, they will converge to a smaller set of designs and begin to use P2 citations more often.

A limitation of this study remains that it investigated the citation reasons for two types of research on a single learning platform. It is possible that some aspect of the design of ASSISTments facilitated conducting work that would receive citations for specific ideas more in EDM research than A/B research (although ASSISTments is one of the learning platforms currently most committed to supporting external A/B

researchers). It is also possible that the learning domain (of mathematics) influenced the contributions made by the work, or that the community of researchers drawing upon mathematics education research influenced this paper's results. To draw more substantial conclusions, this work must be replicated within a wide variety of learning platforms (also varying by subject matter). However, there are currently only a small number of learning platforms used at scale both for A/B testing and learning analytics research, though this number is increasing. In future work, we recommend that researchers focus analysis on single platforms, as in this paper. Comparing between different platforms raises confounds not present in single-platform analysis.

Other factors in the field may of course also impact how studies are cited. For example, differences in the expectations of reviewers in venues that see more A/B studies versus EDM studies may impact how authors cite papers when submitting to these venues. The time it takes to conduct A/B studies may also explain the lower total quantity of citation for A/B studies, although not why the type of citation differed.

Another limitation to the study was a possible lack of statistical power. Although a power analysis was conducted prior to research, some rare categories had seeming differences that were not statistically significant (i.e. 1.2% versus 3.6% for category P6). Unfortunately, this limitation was unavoidable for the overall data set, even if we had coded every example in the data (an arduous task). Power of 0.8 would only have been achieved by category P6 if we had been able to code 878 examples of both A/B and EDM, larger than the total current population for A/B, even if we had skipped the necessary step of conducting a post-hoc correction. P5, the next closest category to significance, would have required 1088 examples of each category. Thus, investigating differences in categories this rare would require a substantially larger data set. It is possible that this paper's work can eventually contribute to such a goal, by developing a categorization scheme and building a corpus of codes that can be used as a training set for an eventual NLP approach that can automatically detect why one paper cites another [13]. Ultimately, the work presented here suggests that EDM papers and A/B testing papers are cited for different reasons. More comprehensively investigating this topic – and investigating subcategories within these broader categories of research – may help us to understand how scientific progress occurs, in our field and more broadly.

References

1. Baker, R. S., Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1(1), 3-17 (2009).
2. Beck, J. E., Arroyo, I., Woolf, B. P., Beal, C. An ablative evaluation. In Proceedings of the 9th International Conference on Artificial Intelligence in Education, 611-613 (1999).
3. Bornmann, L., Daniel, H. What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64, 45-80 (2009).

4. Cambrosio A., Limoges, C., Courtial, J., Laville, F. Historical scientometrics? Mapping over 70 years of biological safety research with coword analysis. Scientometrics 27(2),119-143 (1993).
5. Chen, G., Rolim, V., Mello, R. F., Gašević, G. Let's shine together! a comparative study between learning analytics and educational data mining. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 544-553 (2020).
6. Cole, J. R., Cole, S. The Ortega hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. Science 178, 4059, 368-375 (1972).
7. Cronin, B. The citation process. The role and significance of citations in scientific communication, 103 (1984).
8. Romero, C., & Ventura, S. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), (2020).
9. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), 1432-1462 (2014).
10. Dormezil, S., Khoshgoftaar, T., Robinson-Bryant, F. Differentiating between Educational Data Mining and Learning Analytics: A Bibliometric Approach. In Proceedings of the Workshops of the International Conference on Educational Data Mining, 17-22 (2019).
11. Fazeli, S., Drachsler, H., Sloep, P. Socio-semantic Networks of Research Publications in the Learning Analytics Community. In Proceedings of the LAK Data Challenge (2019).
12. Garfield, E. Can citation indexing be automated. Symposium Proceedings of the statistical association methods for mechanized documentation, 189-192 (1965).
13. Garzone, M., Mercer, R.E. Towards an automated citation classifier. In Conference of the Canadian society for computational studies of intelligence, 337-346 (2000).
14. Goel G., Lallé, S., Luengo, V. Fuzzy logic representation for student modelling. Proceedings of the International Conference on Intelligent Tutoring Systems, 428-433 (2012).
15. Gross, P. L. K., Gross, E.M.K. College libraries and chemical education. Science 66 (1713), 385-389 (1927).
16. Hopkins, A. L., Jawitz, J. W., McCarty C., Goldman, A., Basu, N. Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors. Scientometrics, 96(2), 515-534 (2013).
17. Khajah, M., Lindsey, R. V., Mozer, M. C. How deep is knowledge tracing? In Proceedings of the International Conference on Educational Data Mining (2016).
18. Kizilcec, R., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., Tingley, D. Scaling up behavioral science interventions in online education. In Proceedings of the National Academy of Sciences, 117 (26), 14900-14905 (2020).
19. Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining, 43-56 (2010).
20. Koedinger, K. R., Corbett, A. T., Perfetti, C. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36 (5), 757-798.
21. Lindgren, L. If Robert Merton said it, it must be true: A citation analysis in the field of performance measurement. Evaluation 17 (1), 7–19 (2011).
22. Liu, R., Koedinger, K. R. Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains. Journal of Educational Data Mining 9(1), 25-41 (2017).
23. Malmi, L., Sheard, J., Kinnunen, P., Sinclair, S., Sinclair, J. Theories and models of emotions, attitudes, and self-efficacy in the context of programming education. Proceedings

of the 2020 ACM Conference on International Computing Education Research, 36-47 (2020).
24. Maturana, R., A., Alvarado, M. E., López-Sola, S., Ibáñez, M. J., Elósegui, L. R. Linked data based applications for learning analytics research: Faceted searches, enriched contexts, graph browsing and dynamic graphic visualisation of data. Proceedings of the LAK Data Challenge (2013).
25. Mostow, J., Beck, J. E., Valeri, J. Can automated emotional scaffolding affect student persistence? A Baseline Experiment. In Proceedings of the Workshop on "Assessing and Adapting to User Attitudes and Affect: Why, When and How?" at the 9th International Conference on User Modeling (UM'03), 61-64 (2003).
26. Murphy, R., Roschelle, J., Feng, M., Mason, C. A. Investigating efficacy, moderators and mediators for an online mathematics homework intervention. Journal of Research on Educational Effectiveness, 13(2), 235-270 (2020).
27. Ostrow, K., Heffernan, N., Williams, J. J. Tomorrow's edtech today: establishing a learning platform as a collaborative research tool for sound science. Teachers College Record 119, 3, 300-306 (2017).
28. Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., Baker, R. S. Who's Learning? Using Demographics in EDM Research. Journal of Educational Data Mining, 12 (3), 1-30 (2020).
29. Park, J., Choi, H. J. Factors influencing adult learners' decision to drop out or persist in online learning. Journal of Educational Technology & Society 12 (4), 207-217 (2009).
30. Reich, J. Rebooting MOOC research. Science 347 (6217), 34-35 (2015).
31. Shockley, W. On the statistics of individual variations of productivity in research laboratories. In Proceedings of the IRE 45 (3), 279-290 (1957).
32. Stamper, J. C., Derek Lomas, Dixie Ching, Steve Ritter, Kenneth R. Koedinger, and Jonathan Steinhart. The Rise of the Super Experiment. Proceedings of the International Conference on Educational Data Mining Society (2012).
33. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46 (4), 197-221 (2011).
34. Verbert, K., Duval, E., Klerkx, J., Govaerts, S., Santos, J. L. Learning analytics dashboard applications. American Behavioral Scientist, 57(10), 1500-1509 (2013).
35. Vinkler, P. A quasi-quantitative citation model. Scientometrics, 12(1-2), 47-72 (1987).
36. Waheed, H., Hassan, S., Aljohani, N. R., Wasif, M. A bibliometric perspective of learning analytics research landscape. Behaviour & Information Technology 37(10-11), (2018).
37. Wise, A. F., Jung, Y. Teaching with analytics: Towards a situated model of instructional decision-making. Journal of Learning Analytics 6 (2), 53-69, (2019).
38. Yeung, C., Yeung, D. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In Proc. of ACM Conference on Learning at Scale, 1-10 (2018).
39. Zhang, J., Shi, X., King, I., Yeung, D. Dynamic key-value memory networks for knowledge tracing. In Proc. of the 26th international conference on World Wide Web, 765-774 (2017).
40. Zouaq, A., Joksimovic, S., Gasevic, D. Ontology Learning to Analyze Research Trends in Learning Analytics Publications. In Proceedings of the LAK Data Challenge (2013).
41. Krichevsky, N., Spinelli, K., Heffernan, N., Ostrow, K., & Emberling, M. R. (2020). *E-TRIALS* (Doctoral dissertation, Worcester Polytechnic Institute).