Same Learning Platform, Different Types of Research: A National-Level Analysis

Nidhi Nasiar, Ryan S. Baker, Juliana Ma. Alexandra L. Andres, Namrata Srivastava
University of Pennsyvlania
nasiar@upenn.edu, ryanshaunbaker@gmail.com, alexandraandres@gmail.com, namratas@upenn.edu

ABSTRACT
Online learning platforms have facilitated A/B and secondary data analysis (SDA) studies, which contribute to science differently. This paper compares these types of research within the context of 123 studies conducted in ASSISTments, analyzing how these two types of research differ in research topics, the institution location and affiliation of researchers, citations, and whether these studies serve as a first entry to the field for new researchers or a first opportunity to use new methods. We find all A/B studies are from the USA, while the majority of SDA studies come from China, particularly after 2020. Over half of SDA studies involve Knowledge Tracing (KT), especially in China. In contrast, USA SDA studies involve a broader range of topics. Finally, first-time researchers are more likely to publish SDA than A/B studies, and are more likely to publish at EDM than other conferences.

Keywords
Secondary data analysis, A/B studies, Scientometrics, Online learning systems, comparative analysis

1. INTRODUCTION

1.1 Research supported by Online Learning Platforms

The increase in the use of online learning platforms has opened up new opportunities for students to learn, and has provided researchers with better means to study student learning in depth. The adoption of online learning platforms has grown in both classrooms and non-traditional educational settings [28, 5]. The increase in usage has resulted in the collection of extensive digital trace data of student interaction, which has created potential for research [23]. A large user base has also enabled researchers to conduct automated experiments on a much larger scale in authentic learning settings. The increase in learners using educational platforms has created these two major research opportunities, enabling a plethora of scientific studies to investigate student learning and behaviors in specific educational contexts [24]. This work can be categorized into two broad types of studies: a) A/B studies that conduct experiments on online platforms, and b) secondary data analysis (SDA) on large-scale datasets.

In the early days of the field, and still to a large extent today, it was common for individual research groups to use their own platforms for research (such as in [13]). This practice limited replication and restricted the research focus of studies to align with the interests of specific research teams and their funders, giving these groups a dominant influence on the field's direction. The advent of large-scale platforms for educational data sharing, such as the PSLC DataShop [10]), reduced barriers for external researchers to access and analyze large educational datasets. Furthermore, the advent of tools enabling external researchers to conduct automated experiments within digital learning platforms has helped democratize research in technology-enhanced education and learning sciences [19].

Today, the ability of open platforms to support large-scale automated experiments conducted with thousands of students has made it possible to conduct studies that were previously difficult to carry out in traditional classroom settings. The ability to conduct these experiments with bigger sample sizes results in higher statistical power, increasing their likelihood of capturing significant effects (if they exist), and the ability to conduct them across wider samples also increases the likelihood that findings are generalizable across platforms and populations. This scale-up in research helps improve education in general by offering opportunities to test a range of learning theories and hypotheses, and inform learning engineering efforts [25]. Some platforms like ASSISTments' E-TRIALS [18], and Terracotta [14] have created tools to make the process of setting up and running A/B tests easier for external educational researchers [18]. Beyond this, publicly available large educational datasets with rich fine-grained data have opened avenues for secondary post-hoc analyses by researchers to find meaningful insights on learner processes and performance [24].

ASSISTments, an online math learning environment, has taken steps to facilitate both of these two types of research – A/B studies and SDA analyses. The availability and accessibility of ASSISTments has opened up research to a broader community of scientists, facilitating research and making it less expensive to conduct. However, few studies have examined how these two types of research opportunities have been utilized and by whom.

1.2 Research Questions

Recent studies by [17] and [3] have shown that A/B and SDA papers are cited for different reasons, indicating that both contribute to advancing scientific discourse, but in distinct ways [3]. These papers found that A/B papers were cited more often to provide background and context for a study, while SDA papers were cited to use past specific core ideas, theories, and findings in the field. However, this focus on research impact leaves a gap in understanding various other ways that A/B and SDA studies might differ, such as research topics and where and by whom these types of research are conducted. In this paper, we investigate:

RQ1: What topics are commonly studied in A/B and SDA research?
RQ2: How do A/B and SDA studies differ from each other on the following dimensions:

a) Institutional location and affiliation
b) Research impact
c) First-time contributions

For RQ1, we investigate which research topics are more frequently explored, creating a complement to past studies which have compared the research topics between sub-communities of our field [5]. RQ2a identifies trends in the geographic location (country) of

researchers conducting A/B and/or SDA studies, in order to understand whether the research traditions and practices of a particular country's academic institutions influence the choice of methods. To investigate RQ2b, research impact was measured by the analyses of citation counts received by A/B and SDA studies. For RQ2c, both types of studies were analyzed according to two separate aspects of first-time contribution: first time using the method, and first time publishing in a specific conference. For this analysis, we focused on the Educational Data Mining (EDM), Learning Analytics and Knowledge (LAK), and Artificial Intelligence in Education (AIED) conferences. By investigating the new use of a method and new entry into a conference, we can investigate if these open data sets create opportunities for entry to new scholars (either to the method, the conference, or overall).

The research questions of this study broadly fall within the area of Scientometrics, the study of the properties of scientific publications using statistical and (more recently) data science methods [16], which has been used in EDM/LAK/AIED to assess the progress and development of a research community [1, 6, 12, 22, 27], evaluate contributions to the field [2, 7, 23, 24, 27], and to identify the common topics that are published at conferences and conferences' trajectories of evolution over time [21, 23, 27].

2. CONTEXT: ASSISTMENTS

We conduct this research in the context of papers conducting research using ASSISTments, a platform widely used by external researchers for both A/B tests and SDA analyses. By conducting this study within papers involving a single platform, we control for possible differences due to differences in platforms. ASSISTments is an online math learning platform used for both homework and in-class activities [18] by over half a million students a year worldwide, the majority in the United States. It provides mastery learning, spiraling feedback, and real-time feedback, and offers teachers data on student performance as a formative assessment tool to support future learning. Several randomized controlled trials (RCTs) have been conducted to evaluate the platform's effectiveness [15, 29]. Over the last several years, ASSISTments has been one of the learning platforms most used for research by researchers in EDM and related communities [23].

ASSISTments is an appropriate choice for this scientometric study, as it is among the few widely used platforms that facilitates and supports external research of both types (A/B and SDA), investigating questions about math learning and tutoring. ASSISTments offers large-scale anonymized datasets of student interaction logs, available for analysis by researchers. There are fourteen Open Released Datasets with rich interaction logs. Some datasets also include additional data, such as field observations of learner behavior and affect, or longitudinal student outcomes. These publicly available datasets from ASSISTments have been used in over 100 papers since 2012. ASSISTments' E-Trials platform also supports A/B testing research to run large-scale automated randomized experiments since 2014. E-Trials has been used by over 80 external researchers and collaborators.

2.1 Publications Surveyed

Data collection consisted of exhaustively collecting all published papers that used ASSISTments' open datasets and those conducting A/B studies on the platform until March 2023[1]. First, an exhaustive list of papers using E-Trials and ASSISTments open datasets was obtained from ASSISTments. Searches of the DBLP database and Google Scholar did not obtain other qualifying articles. Google Scholar was used to retrieve each paper's authors, affiliated institutions, location of institution, publication year, abstract, and citation counts as of March 2023. Google Scholar has been used previously in many scientometric studies seeking coverage of conferences [27, 9]. Full access to all papers was obtained using the University's Library Services. Other categories of investigation such as topics, type of affiliation, and first-time contribution across papers were qualitatively coded (see sections below).

The focus of this current study is on examining the utilization of publicly available datasets and the platform's infrastructure for online experiments by external scholars, so studies published before 2012 for SDA and before 2014 for RCTs were not considered. Additionally, studies where a platform founder is the first author or the only author were excluded. Studies conducted solely by scholars at WPI (the university where the ASSISTments team is based) were also excluded. However, publications that involved collaborations between researchers at WPI and other universities were considered, along with entirely non-WPI publications. The final list of papers included full and short papers; poster papers were excluded due to their limited length, leading these papers to have insufficient information for our analyses.

We used this corpus of papers to conduct summary and exploratory analyses of the papers, their topics, their patterns of citation, and what papers are published by which authors. The following sections will describe how we distilled each of these types of information for these papers. After conducting summary and exploratory analyses, we also conducted a set of statistical analyses where we compared the proportions of papers in different categories. Each statistical analysis was a chi-squared test. Due to multiple comparisons, a Benjamini and Hochberg post-hoc correction [4] is applied to all the p-values from all of the chi-square tests in the entire paper together.

2.2 Topics in A/B and SDA Studies

To identify the topic of each paper for RQ1, we conducted a thematic analysis for A/B and for SDA studies separately. The process consisted of: i) initial familiarization by going through papers' abstracts and noting preliminary topic categories that emerge across papers; ii) systematically re-reading the abstracts, specifically looking for the research question, objectives, and summary of findings to get further information and developing codes inductively; and iii) reviewing and refining the resulting list of topic categories to ensure it was exhaustive. Afterwards, the topic categories were reviewed iteratively by two experts in the field until a consensus was reached for the final topic categories. Inter-rater reliability (IRR) was then established between the first and third author (kappa = .88) by coding the same sample of papers (n=70). Once reliability was established, the first author labeled the remaining papers individually.

2.3 Institutional location and affiliation

For each publication, the location (country) of each author's affiliated institution was recorded. For papers where all the authors hailed from institutions in the same country, the paper was assigned that country. In 3 cases, authors hailed from institutions from different countries; we assigned the paper to each country, and

[1] The dataset used in this study is publicly available at: https://bit.ly/3vTR20i

treated each assignment as a distinct instance. Each institution was also coded by humans as being an academic or non-academic institution. The academic category included academic institutions and their affiliated labs or centers; non-academic included for-profit corporations, non-profit organizations, and government-funded independent research groups. A near-perfect kappa of 0.97 was achieved for IRR (the one point of disagreement involved a non-academic institution with an ambiguous name). Lastly, the affiliation categories were aggregated for each paper, and the 16 studies where authors came from both academic and non-academic backgrounds were excluded from comparisons of papers coming from academic versus non-academic settings.

2.4 First-time contribution

For RQ2c, the code for the first-time use of a method was binary-coded for each author across all papers. This involved a two-stage process: 1) identifying the method(s) used in the author's paper, and 2) determining whether the author had used that method in any of their previous publications. The first step in the process was to identify the method(s) used in the paper. Two coders conducted a thematic analysis of the abstracts of all the papers similar to the analysis conducted for RQ1. The emerging themes for methods had substantial overlap with the topic categories identified for SDA studies within RQ1. Out of eight categories of topics that are listed in the results section, six of them showed up as method categories. The remaining SDA papers were analyzed qualitatively again, accessing the full text to identify the categories for methods used. Some examples of additional categories were sequence mining and association rule mining. As for the A/B studies, the category itself is based on a method and therefore all A/B studies were labeled as A/B testing/experimental design for their method.

To identify whether this method had been used before by the researcher, the list of publications of each author was filtered for papers before the date of the target paper. The 1st author used a script to automatically scrape the publication list from each author's ORCID ID for the most up-to-date list, and filter out the papers published after the target paper. If ORCID ID was unavailable, the scholar's name was searched on the web for any public record of their publications such as an institution page, personal website, or a ResearchGate or Google Scholar profile. The abstract was extracted for all the papers before the target paper date and searched for the identified method for that author. Authors were considered independently for every paper -- in other words, if an author publishes a paper using method X for the first time in 2019 and again in 2021, then they count as first time in 2019 but not in 2021. Agreement between the two coders applying the code for the first-time use of the method was acceptable, with a kappa of .74.

To analyze whether the support from ASSISTments was helpful to new researchers in joining the community, we also analyzed whether the authors of each paper were publishing in that paper's venue (EDM, AIED, LAK) for the first time. The list of publications for each author, was used to extract the venues of all papers. A script was used to check whether each author on the paper had published at that specific conference before.

3. RESULTS

3.1 Published Papers

The final corpus of papers consisted of 123 papers, with 99 secondary data analysis (SDA) papers and 24 A/B testing papers. A total of 410 unique authors were identified across all publications. Figure 1 shows that (i) there was a substantial increase in SDA studies from 2019-2021, returning to pre-2019 levels in 2022. In contrast, the number of A/B studies fluctuates mildly without any drastic rise. ii) there were more SDA publications than A/B publications each year except for 2018.

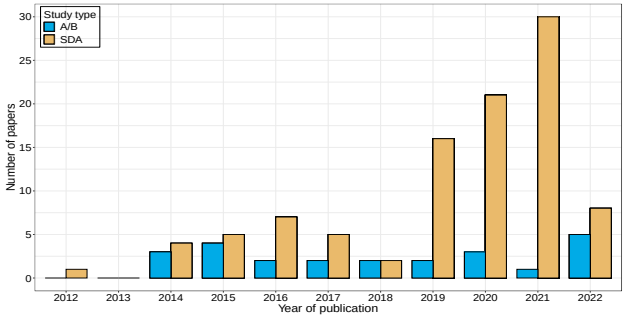

Figure 1: The number of A/B and SDA studies across years

3.2 Topics Studied

There were 12 total categories of topics identified by coders, 4 for A/B studies and 8 for SDA studies. Full descriptions of each topic are given in Appendix 1. A/B studies compare two conditions/interventions. The most common A/B topic (37.5% of A/B studies) was learning transfer and strategies, followed by 29% of papers on feedback types, hints, scaffolds, and worked examples for improving performance; 25% papers on language modification of content and format of problems; and 8% of studies testing spacing and scheduling in math problems. More than half (67%) of SDA studies involved Knowledge Tracing, followed by papers on methods of success prediction other than Knowledge Tracing (11%), and behavior detectors (7%). Reinforcement learning and NLP (natural language processing) techniques each represent 6% and knowledge structures, correlation mining, and clustering each comprised 1% of studies using ASSISTments open datasets.

Figure 2 shows that Knowledge Tracing (KT) has seen a major rise from the year 2018 with n = 1 to peaking in 2021 with n = 24. Reinforcement learning and NLP have seen a gradual increase across years, whereas success prediction and behavior detectors have had consistent popularity throughout.

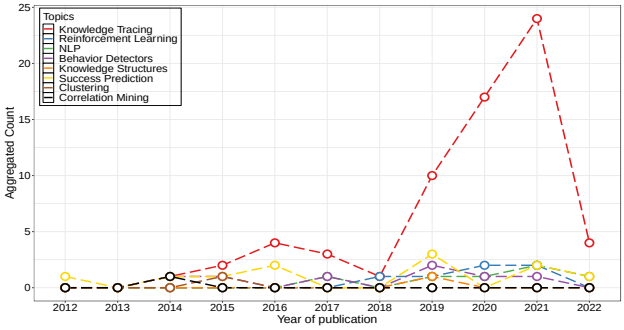

Figure 2: Number of studies with different topics across years

3.3 Differences in Institutional Locations and Affiliations

All 24 A/B studies in the sample (100%) were conducted by researchers in the USA (see Figure 3), with no authors in other countries. By contrast, secondary data analysis (SDA) studies using the ASSISTments open datasets were carried out in the USA, China, Australia, Japan, India, France, South Korea, Italy, Canada, and Scotland. The largest number of SDA studies are conducted in China (44%), followed by the USA (36%). After this, a range of countries each represented 2 to 5% of the data set: Australia, India,

Japan, France, Canada, South Korea, Italy, and Scotland. The relationship between geographic location (across countries) and the type of study conducted (A/B versus SDA) had a significant difference in proportions, $\chi 2(9, N=126) = 30.6$, $p = .0004$, adjusted $\alpha = .013$, remaining significant after applying a Benjamini & Hochberg post-hoc correction collectively for all chi-square tests in this paper. These results indicate that there was a higher proportion of A/B tests in the USA than other countries (the only difference in the data set, as all other countries had a proportion of 0% A/B).

Changes over time are shown in Figure 4. The USA held a leading position in the publication of SDA studies until 2020 when China surpassed it. In the year 2021, China saw a marked increase in its publication count of SDA studies from n=10 to n=20, whereas the USA experienced a downturn from n=7 to n=4.

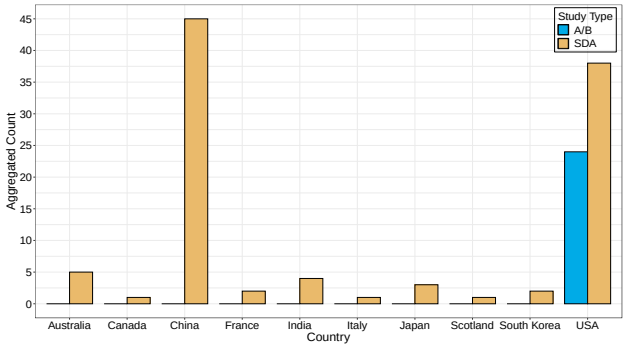

Figure 3: Number of A/B and SDA studies across countries

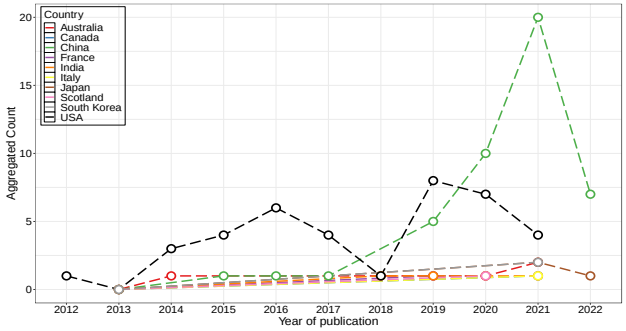

Figure 4: Number of SDA studies across countries over years

As Figure 5 shows, in almost every country, the most popular topic was KT research, but countries varied in their research otherwise. All Japan, France, South Korea, and Canada SDA studies involved KT. 82% of research in China involved KT, and 80% of research in Australia involved KT. In the USA and India, about half of SDA research involved KT, but other methods were also seen. Italy and Scotland, each with two papers, saw 100% use of NLP methods. The results indicate that all A/B studies were conducted by researchers at academic institutions only. The majority of SDA studies were carried out by researchers affiliated with academic institutions (90.5%). Figures 6 and 7 show that the number of publications in A/B and SDA studies has changed over the years for different groups of researchers. The sharp rise in SDA studies from 2019 to 2021 appears to have occurred almost entirely within academic institutions.

Conversely, the frequency of A/B studies conducted by academic affiliations shows an overall pattern of decrease from 2015 to 2021, with a rise only from 2021 to 2022, and with no A/B studies

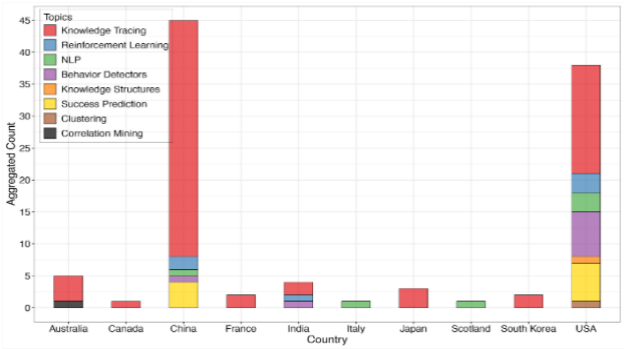

Figure 5: Number of SDA studies on topics across countries

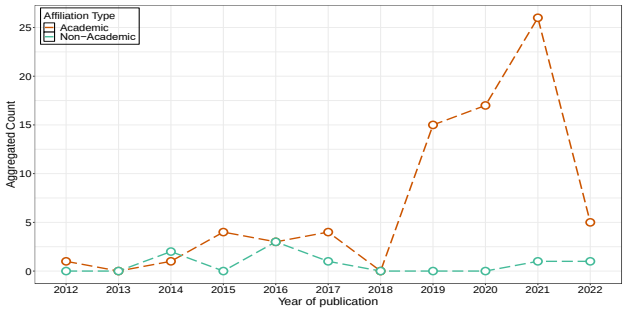

Figure 6: Number of SDA Studies by affiliations over years

published by non-academic institutions on ASSISTments. This absence of non-academic studies could be because industrial researchers are more inclined to study features and designs internally on their own platforms rather than utilizing external platforms such as ASSISTments. However, the association between academic versus non-academic affiliations and the type of study (A/B or SDA) was not significant, $\chi 2(1, N = 107) = 2.37$, $p = .12$; there is not clear evidence that different types of institutional affiliations produce different kinds of papers, despite the fact that no A/B tests were conducted solely by researchers with non-academic affiliations.

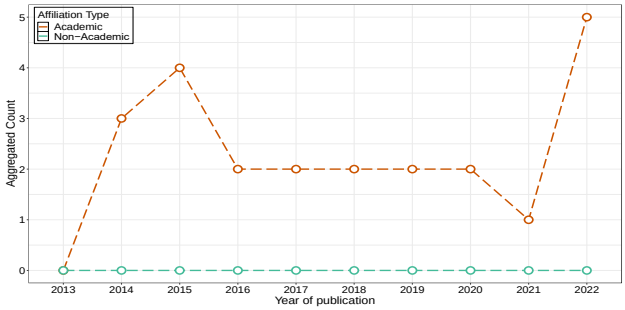

Figure 7: Number of A/B Studies by affiliations over years

3.4 Differences in Research Impact

The dataset included a total of 99 papers for SDA, which had a total combined 4380 citations (average = 44, stdev = 128), and 24 A/B studies which had a total combined 410 citations (average = 17, stdev = 14), both as of March 2023.

3.5 Differences in First-time Contribution

24% of the authors of SDA papers used the method identified in the paper for the first time, whereas only 8% of the authors of A/B studies used the method for the first time, a significant difference in proportions, $\chi 2\ (1, N = 402) = 12.07$, $p = .0005$, adjusted $\alpha = .025$. Within SDA papers, the USA and China have comparable

proportions of researchers using a method for the first time (27% for USA, 26% for China).

36% of authors conducting A/B studies in AIED using the ASSISTments platform had never published there before. 23% of authors of A/B studies were publishing in EDM for the first time, and 22% of authors conducting A/B studies were publishing in LAK for the first time. 54.5% of authors conducting SDA studies using the ASSISTments platform published in LAK were publishing in LAK for the first time. 43% of authors conducting SDA studies in EDM were first-time authors there. 31% of authors of SDA studies publishing in AIED did so for the first time. A Chi-square test indicated that after post-hoc correction, there was a marginally significant difference in these proportions, $\chi 2$ (2, N = 61) = 6.23, p = .0443, adjusted α= 0.038.

As Figure 8 shows, the majority of first-time authors of SDA papers at EDM and LAK were based in the USA. By contrast, the majority of first-time authors of SDA papers at AIED were based in China.

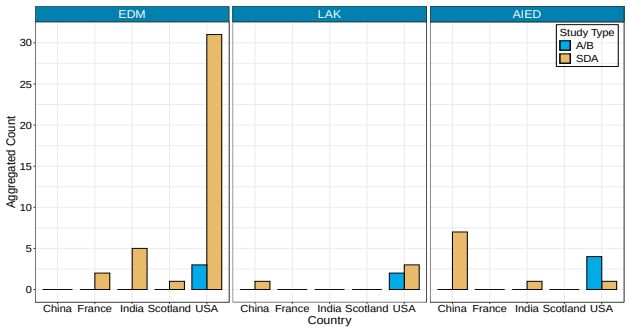

Figure 8: Number of authors publishing for the first-time at EDM, LAK, or AIED conferences across countries

4. CONCLUSION AND DISCUSSION

Online learning platforms have facilitated two types of research that are key to our field -- A/B tests and secondary data analyses (SDA). These two types of studies are both important, as our findings show, but contribute differently to the field. We investigated how these two types of research differ in terms of research topics, the institution location and affiliation of researchers conducting these studies, research impact (i.e. citations), and the degree to which these studies serve as a first entry to the field for new researchers or as a first opportunity to use new methods.

Our findings show that A/B in ASSISTments studies were exclusively conducted in the USA, while SDA studies were more globally distributed, led by China (accounting for 44% of SDA studies). The USA's early lead in SDA publications involving ASSISTments was overtaken by China in 2020, which saw a significant increase in output in that year and the following year. The following drop in 2022 suggests that this spike in SDA publications may have been related to the pandemic, though other factors could have played a role. The absence of international studies conducting automated experiments on ASSISTments could be due to differences in ethics requirements. ASSISTments require Institutional Review Board (IRB) approval (or exemption) from researchers looking to use their platform for A/B studies, which creates barriers for researchers in countries that lack IRBs or other ethics review processes accepted by WPI's IRB. Another possible factor may be that funding bodies and promotion processes at Schools of Education in many countries may not value studies conducted in the ASSISTments system's predominantly US-based populations of learners. Computer Science departments, by contrast, may be more receptive to SDA papers conducted on US data. While open datasets have encouraged international scholarly contributions, participation remains low outside of China, with less than 5% of SDA papers coming from any individual country other than China and the USA. This result underscores that the potential of open datasets and open platforms to enable worldwide research are yet to be fully realized. Establishing a standardized and simplified approach for researchers worldwide to use A/B testing could be a valuable step in this direction. In terms of secondary data analysis, it is worth noting that other open data sets have seen broader international usage, such as the OULAD data set [11], suggesting that differences between data sets (and the types of research they enable) may explain some of the reasons why ASSISTments SDA research was concentrated in two countries.

Second, the results show that Knowledge Tracing (KT) is the most common topic for SDA studies, accounting for 67% of all SDA studies across the countries. This focus aligns with the longstanding academic interest in this topic. From 2018, there was an explosion in papers investigating variants of Deep Knowledge Tracing (DKT), with ASSISTments becoming used as a common benchmark data set across papers [8, 20, 26]. This result shows the contribution of ASSISTments to this development but also indicates that there may be room within the field for data sets tailored to other types of secondary analysis.

A third substantial finding involves scholars using a method for the first time. These studies are substantially more likely to be SDA studies than A/B studies, and are more likely to be published at EDM than other conferences. This may suggest that there are lower barriers to entry for publishing a secondary analysis of a dataset than for conducting an A/B study. Beyond the international factors discussed above, seeing through an A/B test requires a broader range of skills than a secondary data analysis. It therefore may be valuable to hold summer schools (as seen in the Simon Initiative Summer School) that scaffold junior researchers in designing, planning, implementing, and analyzing their studies.

In considering these results, it is important to remember that our investigation within this paper focused on the trends in these two types of research within a single online learning environment. While this choice avoids confounds between different platforms, the characteristics of datasets from ASSISTments and the nature of A/B studies feasible with this platform may influence the differences observed between the two types of studies. Thus, future comparisons should be conducted across a broader array of learning platforms. However, few scaled online learning platforms currently support external researchers in conducting either A/B testing or SDA research, much less both. Most of the platforms that do offer these types of research support have not made this functionality available for nearly as long as ASSISTments, temporarily reducing the ability to draw clear conclusions about differences and trends.

In the longer term, a move towards more platforms offering open functionality for experimentation and sharing their data will enable a wider range of studies in different cultures and contexts. Our current study aims to enhance understanding of how publicly available datasets and research platforms are being utilized to conduct A/B and SDA studies. These insights provide a view of where things are today, and suggest directions for better supporting researchers in ASSISTments and other platforms. Future studies, by better understanding how these trends are playing out in other platforms, and studying a broader range of research questions around who is researching and what they are researching, will help us move from a field where most research remains based on personal connections and affiliations to a field where research is open, public, and communal.

5. REFERENCES

[1] Baek, C., & Doleck, T. (2020). A bibliometric analysis of the papers published in the journal of artificial intelligence in education from 2015-2019. International Journal of Learning Analytics and Artificial Intelligence for Education, 2 (1).

[2] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of educational data mining, 1(1), 3-17.

[3] Baker, R. S., Nasiar, N., Gong, W., & Porter, C. (2022). The impacts of learning analytics and A/B testing research: a case study in differential scientometrics. International Journal of STEM Education, 9(1), 1-10.

[4] Benjamini, Y.,Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Author(s). Journal of the Royal Statistical Society, Series B, 58, 1 (1995), London, UK, 289-300.

[5] Chen, G., Rolim, V., Mello, R. F., & Gašević, D. 2020. 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 (pp. 544-553).

[6] Dietze, S., Taibi, D., & d'Aquin, M. (2017). Facilitating scientometrics in learning analytics and educational data mining–the LAK dataset. Semantic web, 8(3), 395-403.

[7] Fazeli, S., Drachsler, H., & Sloep, P. (2013). Socio-semantic networks of research publications in the learning analytics community. In Proceedings of the Learning Analytics & Knowledge (LAK) data challenge.

[8] Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing?. Journal of Educational Data Mining, 12(3), 31-54.

[9] Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and information technologies, 26, 205-240.

[10] 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).

[11] Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. Scientific data, 4(1), 1-8.

[12] Mirriahi, N., Gasevic, D., Dawson, S., & Long, P. D. (2014). Scientometrics as an Important Tool for the Growth of the Field of Learning Analytics. Journal of Learning Analytics, 1(2), 1-4.

[13] Mostow, J., Beck, J. E., & Valeri, J. (2003). 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), (pp. 61–64).

[14] Motz, B. A., Üner, Ö., Jankowski, H. E., Christie, M. A., Burgas, K., del Blanco Orobitg, D., & McDaniel, M. A. (2023). Terracotta: A tool for conducting experimental research on student learning. Behavior Research Methods.

[15] 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).

[16] Nalimov, V. V., & Mulchenko, B. M. (1969). Scientometrics. Studies of science as a process of information. Moscow, Russia: Science.

[17] Nasiar, N., Baker, R. S., Li, J., & Gong, W. (2022, July). How do A/B Testing and secondary data analysis on AIED systems influence future research?. In International Conference on Artificial Intelligence in Education (pp. 115-126). Springer International Publishing.

[18] Ostrow, K. S., & Heffernan, N. T. (2019). Advancing the state of online learning: Stay integrated, stay accessible, stay curious. Learning Science: Theory, Research, and Practice, (pp. 201–228).

[19] Pardos, Z. A., Tang, M., Anastasopoulos, I., Sheel, S. K., & Zhang, E. (2023). OATutor: An Open-source Adaptive Tutoring System and Curated Content Library for Learning Sciences Research. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

[20] Pavlik Jr, P. I., & Eglington, L. G. (2023). Automated Search Improves Logistic Knowledge Tracing, Surpassing Deep Learning in Accuracy and Explainability. Journal of Educational Data Mining, 15(3), 58-86.

[21] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), 1432-1462.

[22] Prahani, B. K., Rizki, I. A., Jatmiko, B., Suprapto, N., & Amelia, T. (2022). Artificial Intelligence in Education Research During the Last Ten Years: A Review and Bibliometric Study. International Journal of Emerging Technologies in Learning, 17(8).

[23] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355.

[24] Stamper, J., Lomas, D., Ching, D., Ritter, S., Koedinger, K.R. & Steinhart, J. (2012). The Rise of the Super Experiment. In Proceedings of the 5th International Conference on Educational Data Mining

[25] Uncapher, M.R., 2019. From the science of learning (and development) to learning engineering. Applied Developmental Science, 23(4), pp. 349-352.

[26] Vie, J. J., & Kashima, H. (2019). Knowledge tracing machines: Factorization machines for knowledge tracing. In Proceedings of the AAAI Conference on artificial intelligence (Vol. 33, No. 01, pp. 750-757).

[27] Waheed, H., Hassan, S. U., Aljohani, N. R., & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. Behaviour & Information Technology, 37(10-11), 941-957.

[28] Watson, J., Pape, L., Murin, A., Gemin, B., & Vashaw, L. (2014). Keeping pace with K-12 digital learning: An annual review of policy and practice. Evergreen Education Group.

[29] Xiong, X., Wang, Y., & Beck, J. B. (2015, March). Improving students' long-term retention performance: a study on personalized retention schedules. In Proceedings of the 5th

International Conference on Learning Analytics and Knowledge (pp. 325-329).

APPENDICES.

Appendix 1. Categories of Research Topics for A/B and SDA studies

Type	Topic	Description	% of papers (within category)
A/B	Learning Transfer and Strategies	Papers focused on improving long-term retention, transferability of knowledge, and learning strategies.	37.5%
A/B	Feedback and Learning Support	Papers focusing on feedback types, hints, scaffolds, and worked examples affecting performance.	29%
A/B	Content Presentation	Papers focusing on how the content was presented in terms of language and format	25%
A/B	Spacing	Papers testing spacing effects and scheduling in math	8%
SDA	Knowledge Tracing	Papers on improvements to KT algorithms, introducing newer Deep Knowledge Tracing (DKT) versions, and comparing performance across KT algorithms.	67%
SDA	Success Prediction (not KT)	Papers focusing on predicting student success other than immediate correctness knowledge prediction, like prediction of STEM careers, standardized test scores, etc.	11%
SDA	Behavior Detectors	Papers building behavior detectors or an early-prediction model of a behavior, including behaviors such as carelessness, wheel spinning, and productive persistence.	7%
SDA	Natural Language Processing (NLP)	Papers using NLP techniques including but not limited to bag of words,TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, Universal Sentence Encoder (USE), BERT to analyze ASSISTments data.	6%
SDA	Reinforcement Learning (RL)	Papers involving RL simulation studies, including multi-armed bandit (MAB).	6%
SDA	Knowledge Structures	Papers utilizing Q-Matrices to map items and the underlying skills they assess	1%
SDA	Clustering	Papers using clustering as a technique	1%
SDA	Correlation Mining	Papers using correlations systematically to understand or identify patterns in data; does not include papers with one or two correlations.	1%