How to Open Science: Developing and Testing Reproducibility Metrics on the Educational Data Mining Conference

Anonymous
Anonymous Institution
anonymous@anonymous.edu

ABSTRACT

Despite increased efforts to assess the adoption rates of open science and robustness of reproducibility in sub-disciplines of education technology, there is a lack of understanding of why some research is not reproducible. Prior work has taken the first step toward assessing reproducibility of research, but has assumed certain constraints which hinder its discovery. Thus, the purpose of this study was to replicate previous work on papers within the proceedings of the *International Conference on Educational Data Mining* and develop metrics to accurately report on which papers are reproducible and why. Specifically, we examined 208 papers, attempted to reproduce them, documented reasons for reproducibility failures, and asked authors to provide additional information needed to reproduce their study. Our results showed that out of 12 papers that were potentially reproducible, only one successfully reproduced all analyses, and another two reproduced most of the analyses. The most common failure for reproducibility was failure to mention libraries needed, followed by non-seeded randomness.

All openly accessible work can be found in an Open Science Foundation project[1].

Keywords

Open Science, Peer Survey, Reproducibility, Metrics

1. INTRODUCTION

The adoption of open science and robustness of reproducibility within fields of research has incrementally gained traction over the last decade[29, 31]. This adoption trend has led to increased clarity in methodologies, easier execution of analyses, greater understanding of the underlying work, etc. However, in numerous sub-disciplines of education technology, there tends to be a lack of understanding as to why an author's research is not replicable, or even reproducible. For example, within the sub-discipline of 'Educational Data Mining', which has provided large-scale data for analyzing student learning and improve outcomes[3, 26], there are numerous analyses that, while typically falling within the reported confidence intervals, do not produce the exact results reported in the published, peer reviewed paper.

Previous works related to open science and robustness of reproducibility were conducted on the *International Conference on Learning Analytics and Knowledge* (LAK) and the *International Conference on Artificial Intelligence in Education* (AIED). Within the LAK work, 5% of papers were found to adopt some of the chosen practices needed for reproducibility; however, none were successful within a 15-minute timeframe. Within the AIED work, 7% of papers were reported to be 'potentially' reproducible through source analysis with some given assumptions; however, once again none were successful. The AIED work also collected responses from authors in association to their paper, in which 58% of authors reported that they could release a dataset or source needed for reproducibility; however, it did not improve the end result. These prior works only perform a basic overview of the potential reproducibility due to the given time limit, regardless of any extensions. In addition, certain assumptions were made which constrained whether a paper was reproducible which, while improving efficiency, provided a larger lack of understanding as to why the work was reproducible or not.

The goal of this work is to provide a deeper dive into the reproducibility of papers within the field of Educational Data Mining. Specifically, this work will replicate the results of previous work across papers published within the last two years of the proceedings of the *International Conference on Educational Data Mining* (EDM). Trained reviewers examined each paper for open science practices and reproducibility (henceforth referred to as our peer review). We further reached out to authors in an effort to obtain more information about the paper to improve reproducibility.

Each paper was given a hard limit, with minor exceptions, of 6 hours to attempt to reproduce the paper, including communication with the authors. The process needed to attempt reproduction was recorded in a document, along with a breakdown of how much time was needed to do so. If results were obtained that did not reflect those within the paper, then an additional review of the source was conducted to determine the disconnect.

[1]https://osf.io/ah5wq/?view_only=8d3a0e9b957b4f6d92fd74a5daa56d1d

Specifically, this work aimed to accomplish the following tasks:

1. Replicate previous works using the papers within the proceedings of the *International Conference on Educational Data Mining* (EDM).
    - Document and analyze which papers adopt the open science practices and associated subcategories defined by this work.
    - Communicate with the authors of the papers using a survey to measure the understanding and adoption of open science practices and receive additional information to properly reproduce or replicate the paper, if needed.
2. Attempt to reproduce the paper within a 6-hour timeframe, document any additional methodologies not reported within the paper or its resources, and determine, if necessary, why the exact results reported in the paper could not be obtained.

2. BACKGROUND

2.1 Open Science

Open science is an 'umbrella' term used to describe when the methodologies, datasets, analysis, and results of any piece of research are accessible to all[13, 31]. In addition, there are subcategories of 'open science' corresponding to individual topics created before and after the initial adoption in the early 2010s[29]. Within the first half of the decade, there were numerous issues when conducting peer reviews of other researcher's work including, but not limited to, ambiguity in methodology, incorrect usage of materials, etc. Then in the mid-2010s, large-scale studies in psychology[6] and other fields[2] were unable to be reproduced or replicated. As such, open science practices were more commonly adopted to provide greater transparency and longer-lasting robustness in a standardized format such that researchers can adapt and apply their work.

Our personal investment in documenting the adoption and robustness of research in our discipline and its subfields stemmed from our own shortcomings. Specifically, our lab ran into an issue one day where we could not reproduce our prior research. There was a lack of information on how to run the analysis code, minimal information on the provided dataset, and hard-to-diagnose issues when attempting to reproduce the results. The issues were eventually solved with communication from the original author who had since left our lab, but it motivated us to do a better job at making our work more clear and reproducible. Admitting first our own lack of adoption and ability to reproduce our work, our goals of the current work were to investigate the current adoption of open science, survey authors for their reasons for or against adoption, and attempt to reproduce their work and properly diagnose any issues that arise.

2.2 Data Mining

Data Mining is a term used to describe the extraction of previously unknown or potentially useful information from some piece of data[5, 24]. Originally known as 'Knowledge Discovery in Databases' (KDD), it has since expanded to apply the collected information in numerous fields and contexts. Within education, 'Educational Data Mining' has helped collect data on how students learn and teachers provide information at numerous levels (e.g. classroom, school, district) to better improve a student's understanding and outcomes[3, 26]. There were a few workshops in educational data mining since 2005, but in 2008, the *International Conference on Educational Data Mining* (EDM) was created[1] and took the role of hosting research which collected and analyzed large-scale data in educational settings. The collection and analysis associated with data mining practices tend to correspond with those related to open science and is typically a common topic due to developing proper and secure policies[32]. As such, papers submitted to the EDM conference will be used as the dataset for this work.

3. METHODOLOGY

3.1 Open Science Peer Review

To answer RQ1, we adopted the methodology from the previous works. We evaluated every full paper, short paper, and poster paper from the previous two EDM proceedings: the the *15th International Conference on Educational Data Mining*[2] and the *14th International Conference on Educational Data Mining*[3]. Reproducibility of older years was likely to be more difficult as papers become older as software might no longer exist or is outdated or the dataset or source required had been taken down for some reason. Thus, only the last two years were considered. Both proceedings are divided into subsections 'Full Papers' (synonymous with research articles in previous works), 'Short Papers', or 'Poster Papers' (synonymous with posters in previous works). The papers within the proceedings of the *15th International Conference on Educational Data Mining* were identified by their digital object identifier (DOI)[4]. The papers within the proceedings of the *14th International Conference on Educational Data Mining* were identified by their page number within the proceedings[5]. As the identifiers for each proceeding were different but functionally equivalent, they were referred to as unique identifiers (UID). Each review captured a UID, the proceedings the paper was a part of, and the subsection the paper was listed under. Each review for a paper was given a maximum time limit of 15 minutes because of logistical constraints (e.g. non-specified or degraded links, nested resources within citations, etc.). In addition, an explanations document was created which justified why a specific choice was made in the review. If a choice was self-explanatory, the justification was omitted (e.g., no preregistration was linked in the paper, no README was located in the source). Any links within the paper that no longer reference the original resource were marked as degraded and reported in the explanations document.

Open Methodology is a term that says the details of the col-

[2]https://zenodo.org/communities/edm-2022/
[3]https://educationaldatamining.org/EDM2021/EDM2021Proceedings.pd
[4]There was no DOI associated with the proceedings itself, so the citation is a footnote with a link to the community group on Zenodo.
[5]The proceedings of the *14th International Conference on Educational Data Mining* had no DOI. As such, the page number in the proceedings were used. A separate link was provided to the virtual page for each paper as well.

lection, methods, and evaluation of a research project are accessible and usable by all[13]. Compared to a paper, the methodologies typically represent every possible step and resource needed for another researcher to reproduce or replicate the research themselves. All papers submitted to the *15th International Conference on Educational Data Mining* are licensed under the Creative Commons Attribution 4.0 International License[6], or CC-BY-4.0 for short, and are considered 'Open Access'. The papers within the proceedings of *14th International Conference on Educational Data Mining* are unlicensed; however, EDM treats them as 'Open Access' regardless, so they are considered as such for this work.

Open Data is a term that says the dataset(s) associated with the research project is accessible and can be used by all[15, 16]. These datasets are typically specified with a license or are part of the public domain. A dataset is marked as being open if the paper contains a link, or a link to another paper with a link, to the dataset. If the paper mentions explicitly that the dataset can be requested from the authors, then it will be marked as 'on request'. If the paper does not use a dataset, such as for theoretical or development topics, then the field is marked as non-applicable. The licensing on the dataset was not considered as researchers are unlikely to be as familiar with them and are normally ambiguous or too complex to properly understand[11, 25]. A separate field is provided for the documentation of the data which is marked if there exists a location where the dataset's fields are mapped to its associated description. A partial marking for the documentation can be met if there is at least one field documented at some location.

Open Materials is a term that captures whether technologies – including open source software[22, 9], freeware, or non-restrictive services – can be used by all. A paper has open materials if the paper contains a link, or a link to another paper with a link, to all the materials and source the authors used. A partial marking was assigned if there is at least one material mentioned. If there are no materials used, such as for argumentative or theoretical papers, then the field was marked as non-applicable. The documentation for the materials and source, which provides understanding on how to use them[7], also had a field, along with a partial equivalent if the materials or source was not fully documented. If the source was available, then two more fields were considered: the README which contained information on the source and potentially some setup instructions[12] and a license field which said that the source can be used openly[22, 28, 8].

A **preregistration** describes the processes conducted for the paper before the research takes place to prevent hypothesizing after results are known and p-hacking observations[19, 18, 30]. A preregistration can be altered by creating a new preregistration to preserve the initial methodologies. A paper has a preregistration if there is a link within the paper to some location hosting the preregistration (e.g., Open Science Framework[7], AsPredicted[8]). If a preregistration is unnecessary, then the field is marked as non-applicable.

[6]https://creativecommons.org/licenses/by/4.0/
[7]https://osf.io/registries
[8]https://aspredicted.org/

3.1.1 Undergraduate Interpretation

In contrast to previous works, the peer review was handled by two trained undergraduate research assistants, referred to as 'Reviewers' in the explanations document. Undergraduates are typically pressed upon to conduct and publish research prior to graduation for better advancement within their career[14, 27, 23]. As such, it stands to reason that papers should be geared towards the understanding of undergraduates assuming the requisite knowledge. Due to undergraduate interpretation, it was expected to see a higher level of adoption as previous works tended to be highly specific and nuanced when evaluating whether a given subcategory was adopted.

To mitigate any misconceptions or inaccuracies between the reviewers, each reviewer was randomly assigned ten papers that another reviewer reviewed and provided their own review. Both reviews are provided within the explanations document in an arbitrary order.

As a final precaution, the lead on the research project, referred to as the 'Meta-Reviewer', was responsible for resolving any disputes or disagreements within the provided reviews. If two reviewers disagreed on a particular section, the meta-reviewer had the final say as to what was reported. Additionally, if either reviewer asked for verification on a particular review, the meta-reviewer provided the requested feedback and correct markings. Finally, the meta-reviewer lightly reviewed the results of the reviewers for any major inaccuracies in understanding or logic and corrected them as necessary.

3.2 Author Survey

Authors were allowed to provide input to the peer review performed using a survey. For each paper submitted to the two EDM conferences, an email was sent out to the first author[9]. To avoid issues involving the email server (e.g. email marked as spam, denied due to too many receipts), authors with multiple papers published in the proceedings were sent a single email containing the papers they should complete the survey for[10]. As an added measure to improve the number of survey responses, a separate, mass email was sent prior to the survey to notify authors about the survey and what email it would be sent from. The survey responses were publicly released and linked by their UI as stated in our International Review Board (IRB) study. Additionally, the author information provided was removed from the released dataset. The survey itself was sent on November 29th, 2022 and currently continues to collect responses. This work reports on responses collected up to January 3rd, 2023.

The survey asked for the name and email of the author and the UI of the associated paper. The content of the survey was separated into six subsections: data, materials, preregistration, preprint, reproducibility and replicability, and resource degradation.

[9]The first author was assumed to be the corresponding author as EDM does not provide any formal way of marking so.
[10]This email survey was conducted in parallel with two separate research projects for other conferences to mitigate the issues mentioned above. The other research projects will be reported at a later time.

3.2.1 Data

The data section was used to collect information on the dataset and documentation used within the paper. The author first reported whether the dataset is publicly available, is private but can be shared on request, or if the dataset cannot be shared at all. In the case where a dataset was not used or does not correspond with one of the above categories, an additional 'other' option was available with an appropriate text box. If the dataset was not publicly accessible, the author was asked to provide their reasoning as to why. If the dataset could be shared either publicly or on request, the author was asked to provide the location of the dataset along with its associated license. If a link was provided but the dataset could not be released publicly, the link would be scrubbed from the publicly released dataset. This would provide a relatively secure way to share data that may contain sensitive information. All questions were shown for full transparency.

3.2.2 Materials

The materials section was responsible for collecting information on the materials, source, and documentation used within the paper. The questions in this section are the same as those within the data section except replaced with material-related keywords.

3.2.3 Preregistration

The preregistration section was responsible for collecting information on an available preregistration, if applicable, for the paper. The author was asked to report on whether there is a public, private, or no preregistration made for the paper. If a preregistration was not applicable (e.g. theoretical paper, argumentative paper) or did not fit into one of the available categories, an additional 'other' option was available with an appropriate text box. For available preregistrations, whether public or private, the author was requested to provide the associated link. If no preregistration was made, the author was asked to provide their reasoning as to why.

3.2.4 Preprint

The preprint section documented information on an available **preprint**, a paper that usually proceeds formal peer review and publication in a conference or journal[4, 10], for the paper. The author was asked to report on whether a preprint was available for the paper. If a preprint was not applicable or did not fit into one of the available categories, an additional 'other' option was available with an appropriate text box. If a preprint was present, the author was requested to provide the associated link. If no preprint was created, the author was asked to provide their reasoning as to why.

3.2.5 Reproducibility and Replicability

The reproducibility and replicability section documented information needed to properly reproduce or potentially replicate the associated paper. Towards replication, the author was asked to provide any additional methodologies that were not reported in the original paper. Towards reproduction, the author was asked to provide any necessary setup instructions needed to properly connect the dataset to the source and run the associated analysis. This included, but was not limited to, file locations, software versions, setup scripts, etc. If any of the above information was not provided within the paper or its citations, the author was asked to provide their reasoning as to why.

3.2.6 Resource Degradation

The resource degradation section documented information on resources reported within the papers that no longer exist at the specified location. The authors were asked to review their resources for any that no longer exist or point to an incorrect location and provide alternatives if possible. If the resources were degraded, the author was asked to explain what happened to the original resource.

3.3 Reproducibility

An experiment or study is **reproducible** when the exact results reported in the paper can be produced from a static input (e.g. dataset, configuration file) and deterministic methodology (e.g. source code, software)[17, 21, 20]. While reproducibility is the simplest form of reviewing the results of a paper, in practice, there are differing levels of what defines a complete reproduction. For this work to answer RQ2, we assume that a paper is reproducible when the dataset and analysis used in the original paper returns the exact same results and figures as those reported. If either the dataset or analysis method is not present, found within the 15-minute timeframe in the paper or its resources, or provided within the author's survey response, then the paper will be marked as non-reproducible. If the paper does not use a dataset or analysis method or does not run an experiment or study in general, then reproducibility will be marked as non-applicable.

Although we allocated 15 minutes for each paper to find its dataset or analysis, if we were able to track these down, each paper was given a hard limit of 6 hours to reproduce the results reported in the paper. If any action exceeded the 6 hour limit, then the action was stopped and only the exported results were considered with any reasonable educated guesses on the rest of the runtime. The 6 hour limit was only extended if the reproduction could be assumed to be completed within an additional hour. To provide a better and more accurate understanding of the amount of time taken, the collected metric was broken down into three time periods: setup, execution, and debugging. A timing site[11] was used to manually track how long each section took along with the total time. If any breaks were taken by the reviewer, the timers and all actions were stopped and recorded in the explanations document until the reviewer resumed working.

The **setup time** tracked the time taken for all tasks prior to the first execution of the analysis. This includes downloading the dataset and source, setting up the necessary environment, and following information provided within the README, if available. Information that can be assumed from the source was not provided during the setup phase to better simulate cases where a researcher would run the source assuming they had all the necessary libraries installed from previous runs. This time was likely to vary between reviewers depending on factors such as connection speed and should be taken with a grain of salt. The **execution time**

[11] https://stopwatch.online-timers.com/multiple-stopwatches

tracked the time taken during the execution of the program. This began when the program was ran (e.g. command, button) and stopped when the program finished executing or crashed. This time was the total time on execution any might included multiple runs. Any specific information was recorded in the explanations document. The **debugging time** tracked the time taken between executions when the analysis crashed. Any diagnoses made which corrected the issue was reported in the explanations document. A perfectly reproducible analysis should have minimal to no debugging time.

All reproducibility tests were run on a single big data machine used within the author's lab. The machine was chosen for two reasons. First, as a big data machine, it can run numerous calculations relatively quickly depending on the efficiency of the analysis. Second, it runs a Unix-based operating system with a Bash shell which most scripts provided by researchers are typically for. For benchmarking purposes, the specifications of the machine are listed in Appendix B.

3.3.1 Python

If the environment needed to reproduce the source used Python[12], then the following steps were taken:

1. If a specific version of Python was specified, download and select the version of Python.
2. Create a empty virtual environment using 'venv'[13] and activate it.
3. Follow any setup steps specified by the analysis.
4. If the analysis is in a Python (.py) file:
   (a) Run the file using the 'python' command.
5. If the analysis is in a Python Notebook (.ipynb):
   (a) Install 'ipykernel' and 'notebook' using the 'pip' command.[14]
   (b) Open the notebook and specify the kernel used as the one within the virtual environment.
   (c) Run the notebook.

3.3.2 R

If the environment needed to reproduce the source used R[15], then the following steps were taken:

1. If a specific version of R was specified, download and select the version of R.
2. Create a new project using RStudio[16] or another IDE that can use 'packrat'[17][18].

[12]https://www.python.org/
[13]This is the recommended way for Python 3; however, there are other methods to do so.
[14]If the path is improperly configured, the command may need to be prefixed with 'python -m'.
[15]https://cran.r-project.org/
[16]https://posit.co/products/open-source/rstudio/
[17]https://cran.r-project.org/package=packrat
[18]'packrat' is the most commonly used option for managing R dependencies. It is not the only method.

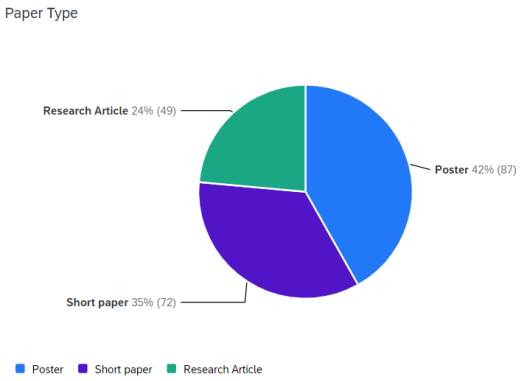

Figure 1: A representation of the review on the full papers, short papers, and poster papers published within the proceedings of the 15th and 14th EDM conferences.

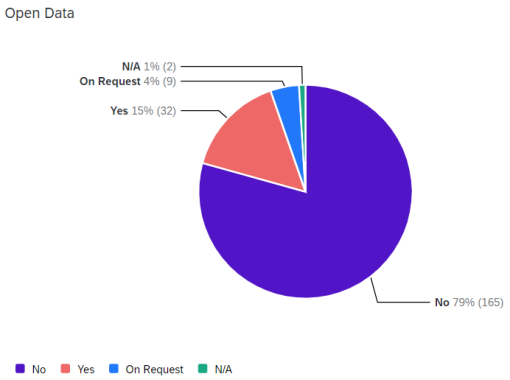

Figure 2: A representation of the review on the adoption of open data within papers published in the proceedings of the 15th and 14th EDM conferences.

3. Follow any setup steps specified by the analysis.
4. Run the R script.

4. RESULTS

4.1 Peer Review

As shown in Figure 1, across the 99 papers published in the 15th proceedings and the 109 published in the 14th proceedings, there were 49 full papers (research articles), 72 short papers, and 87 poster papers.

As shown in Figure 2, 32, or 15%, of papers used a dataset that was already or made openly available. 5% mentioned that the dataset could be requested. Out of those 15% with openly available data, 69% had full documentation on the dataset while the other 31% had partial documentation.

As shown in Figure 3, 31, or 15%, of papers used materials and made the source openly available. 20% used at least on openly available materials. Out of those 15% with openly available materials, 45% had full documentation while 55%

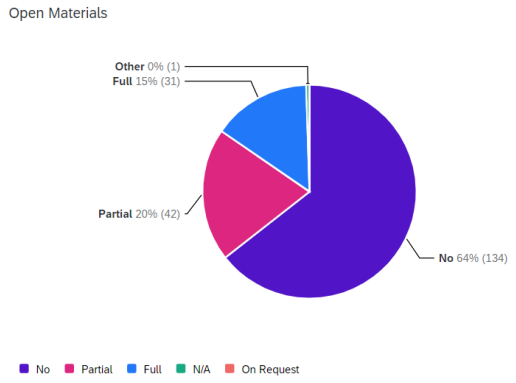

Figure 3: A representation of the review on the adoption of open materials within papers published in the proceedings of the 15th and 14th EDM conferences. The other encompasses the 1 non-applicable paper and the 0 papers in which you could request the materials.

had partial documentation. Additionally, 94% of the open materials had a README while 44% had a permissive license provided with the source.

As shown in Figure 4, only three, or 1%, of the papers had a preregistration linked to it. One of the papers was a short paper while the remaining two were poster papers. One paper was determined to be non-applicable for having a preregistration as it was a concept discussion.

Finally, as shown in Figure 5, nine, or 4%, of the papers provided dataset links that were no longer located in its original location. Two were full papers, five were short papers, and the remaining two were poster papers. Six, or 3%, provided material links that were no longer available. One was a full paper, four were short papers, and the remaining one was a poster paper.

4.2 Author Survey

Out of the 208 surveys sent, only 13, or 6%, complete responses were received. In addition, 7% of the surveys did not reach their destination in a timely fashion: 2 received auto response emails about a delay in reading the email, 2 were denied by the mail server, and 10 emails were no longer available or locatable on the mail server.

Out of 13 responses, 3 papers reported that their datasets were publicly available, 5 papers reported that their dataset could be requested, and 5 papers reported that they cannot share their datasets. Out of the 8 public and on request responses, 5 did not mention in the paper that they could share or request the dataset. Out of the 10 on request and cannot share responses, 6 mentioned they do not have the rights or necessary license to release the dataset, 3 mentioned that the dataset contains sensitive information due to an IRB or some other committee, and 1 mentioned they simply did not have enough time to go through the process of reviewing and potentially publicly releasing a dataset.

For materials, 9 reported that they could make their ma-

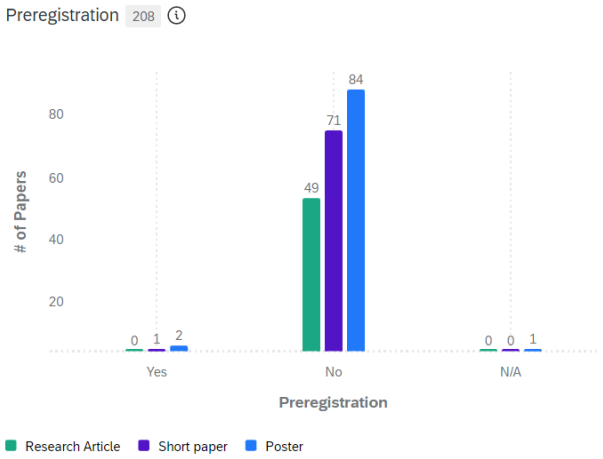

Figure 4: A representation of the review on the adoption of preregistrations within papers published in the proceedings of the 15th and 14th EDM conferences split by paper type.

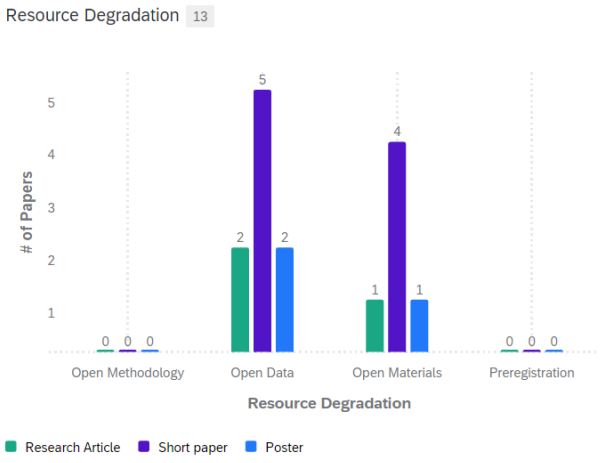

Figure 5: A representation of the review on the degradation of resources within papers published in the proceedings of the 15th and 14th EDM conferences split by paper type.

terials and source public, 3 reported that they could share their materials and source on request, and 1 mentioned that they cannot release their materials and source. Out of the 12 public and on request responses, 8 did not mention in the paper that they could share or request the materials or source. Out of the 4 on request and cannot share responses, 3 mentioned that the source contains references to sensitive information from the associated dataset while 1 mentioned they simply did not have the time nor motivation to go through the process of reviewing and potentially publicly releasing their materials and source.

Towards reproducibility, only 1 mentioned additional information was necessary to reproduce their work while 2 mentioned that the information on the source should be enough to do so. The provided resources did not have an effect on the reproducibility of the papers within Section 4.3.

1 survey response mentioned that they did create a preregistration and provided a link to it while 12 did not. Out of the 12 who did not create a preregistration, 4 believed that one was not necessary during the beginning of the research project, 1 did not remember the option existed, and 6 did not know what a preregistration was. 1 provided no response.

5 survey responses reported that they did create a preprint while 8 did not. Out of the 5 that created a preprint, only 4 links were provided. Out of the 8 that did not create a preprint, 2 believed that one was not necessary, 2 did not remember the option existed, 2 did not know what a preprint was, and 1 did not believe it was fair to the review process to create a preprint. 1 provided no response.

No survey responses reported anything about their resources no longer existing at the specified location.

4.3 Reproducibility Metrics

Only 12, or 6% of papers, were able to attempt reproduction. 2 papers were unable to be timed due to logistical reasons during setup. One paper requested a Python dependency which was no longer obtainable in an official capacity. The other paper required arguments to run the Python script which were not defaulted. There was no indication as to what the value of those arguments might be, so there was an infinite number of potential combinations. As such, the paper was deemed to be non-reproducible.

5 papers passed the 6-hour hard limit. 1 paper was excused because of the additional overflow, but it did not allow all the results to be completed. 2 papers were still running during the execution time when the 6-hour limit passed; however, only one produced intermediate results that could be compared. 1 paper crashed 30 minutes before the limit and provided intermediate results. The remaining paper was being debugged as there were a number of version incompatibilities between the Python libraries preventing execution which was specific configured.

In addition to the paper that crashed within the 6 hour limit, the version incompatibilities paper contained numerous crashes and changes to the source code before it the version incompatibility was found.

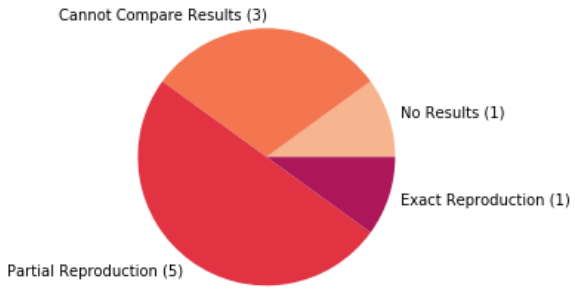

Figure 6: A representation of the test results obtained while reproducing papers published in the proceedings of the 15th and 14th EDM conferences.

Only 9 of the 10 tested papers required some amount of debugging. The remaining paper, while needing no debugging, produced numerous results that did not line up with those reported. Out of the 9 papers which required debugging, all 9 were missing some unreported dependency that needed to be downloaded. 2 papers failed as it did not create the necessary directories to read or write files to.

As shown in Figure 6, out of the 10 tested papers, only 6 produced results that could be potentially linked to the paper or its resources. 3 papers provided results but not in a comparable form to the paper. The remaining paper passed the 6-hour time limit due to version incompatibilities. Only one paper, a poster paper, exactly provided the results expressed within the paper; however, some of the results had to be pulled from an intermediate variable that was not printed. The remaining five provided some of the results reported in the paper; however, only two papers could safely mitigate the inaccuracies due to the confidence interval.

Further source analysis revealed the five papers which did not exactly provide the results mentioned in the paper was due to non-seeded randomness: the seed, or initial value, which in most cases makes the numbers generated by the algorithm fixed instead of random is not set to a deterministic value. Some papers do partially set the seed for some generators but not all.

5. LIMITATIONS

A number of limitations within previous works replicated for RQ1 are still applicable to this paper due to human intervention and limited resources. For the peer review, this includes the subjectiveness of the author's review on the proceedings papers and mitigation through an explanations document. For the author survey, this includes the nonexistent fallback strategy, confusion of email and survey instructions, and the limited responses.

As the peer reviews were conducted by undergraduate research assistants, there are likely some misconceptions between the instructions given, the understanding of the papers, and the explanations for their choices. To better stan-

dardize and mitigate these concerns, the undergraduates were each given a standardized set of explanations which could be used during review. In addition, examples were given to better understand the relationship between the review topic and its corresponding phrases within the papers and associated resources. As an added precaution, the undergraduates could ask a graduate student to perform a meta-review or review other undergraduates' reviews in either agreement or disagreement.

When testing for reproducibility, the total time spent had a hard limit, with one exception, of 6 hours. 2 of the available papers were halted due to this limit; however, only one did not produce intermediate results that could be compared to a paper. It would be useful to properly test the execution for the entire time provided a large number of machines were available.

Additionally, the timing was performed manually instead of through timers associated with the application. There could be slight overestimations in the amount of time taken to reproduce. On the other hand, software timers are ill-suited for such a task as they are typically not multilingual and may not be available for all software.

Finally, the timer categories could be more specified and less generalized. Each timer only represents the length of each section rather than individual sections for how long a specific task took. For the setup and debug categories, these specific sections would not be as useful since different reviewers might take different lengths of time to setup or determine an issue. For the execution category, while it would be useful to know how much time was needed to reproduce the results, it would be better suited as a benchmark from the original author who had already ran the methodology successfully and without issue.

6. FUTURE WORK

Future work should include another round of reproducibility tests on different machines. Each test would provide a valid benchmark on the execution length of the code and serve as a robust measure to validate the reproducibility in numerous circumstances. In addition, results that were inaccurate due to randomness could be averaged to provide a more accurate estimate of the results compared to those reported. Authors could be recruited to run reproducibility tests either voluntarily or through giveaways; however, it would require the authors to have a greater understanding of computer science rather than those needed to provide their analyses.

Another direction for future work could view the impact of conferences which promote open science and reproducibility measures to compare them to those without them. In addition to previous work on author responses and this work on reproducibility metrics, a comparison could be made between the promoting and non-promoting conferences to see whether the adoption of such practices have improved the robustness of research within the discipline.

7. CONCLUSION

Approximately 35% of papers met a partial definition of the chosen open science practices with 5% able to attempt reproducibility with the combined peer review and author survey responses. With the additional time compared to previous works, one paper provided the exact results reported in the paper, while two papers mostly provided the reported results. In addition, while all of the papers needed to download unreported libraries to properly execute the source, the non-exact results collected could all be attribute to non-seeded randomness.

In-depth reproducibility tests and source analysis greatly increases the robustness of an author's paper. The two main issues within the paper might not seem relevant to most authors, but they are likely to have some lasting impact in the future. Library compatibility may not seem useful in a year or two, but after half a decade or so, trying to run the same analysis might prove to be impossible as it did with two of the papers in this work. As for non-seeded randomness, most researcher would agree that as long as the obtained value is within the confidence interval, then it should considered replicable. However, a lack of stability across papers might lead to one reproduction compared to another reproduction, which is not guaranteed to be within each other's confidence interval. As such, deterministic results provide greater robustness and stability such that it can stand the test of time.

Most of the issues can be simplified down to a few additional actions necessary to provide deterministic results. Taking Python analyses as an example, the libraries could be exported with the source by running the 'pip freeze' command. Any source of randomness within Python or popular libraries can also be seeded such as 'random.seed' or, for numpy[19], 'numpy.random.seed'. Other languages or sources are not much different. In the cases where libraries are no longer present, the container itself can be wrapped and provided using services like containerd[20]. By providing these simple, quick actions, the robustness of research, and open science in general, could be greatly improved.

8. ACKNOWLEDGMENTS

BLINDED FOR REVIEW

9. REFERENCES

[1] M. Al-Razgan, A. S. Al-Khalifa, and H. S. Al-Khalifa. Educational data mining: A systematic review of the published literature 2006-2013. In T. Herawan, M. M. Deris, and J. Abawajy, editors, *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, pages 711–719, Singapore, 2014. Springer Singapore.

[2] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, May 2016.

[3] R. Baker et al. Data mining for education. *International encyclopedia of education*, 7(3):112–118, 2010.

[4] P. E. Bourne, J. K. Polka, R. D. Vale, and R. Kiley. Ten simple rules to consider regarding preprint submission. *PLOS Computational Biology*, 13(5):1–6, 05 2017.

[5] M.-S. Chen, J. Han, and P. Yu. Data mining: an overview from a database perspective. *IEEE*

[19] https://numpy.org/

[20] https://containerd.io/

Transactions on Knowledge and Data Engineering, 8(6):866–883, 1996.

[6] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[7] B. Dagenais and M. P. Robillard. Creating and evolving developer documentation: Understanding the decisions of open source contributors. In *Proceedings of the Eighteenth ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE '10, page 127–136, New York, NY, USA, 2010. Association for Computing Machinery.

[8] A. Engelfriet. Choosing an open source license. *IEEE software*, 27(1):48–49, 2009.

[9] J. Johnson-Eilola. Open source basics: Definitions, models, and questions. In *Proceedings of the 20th Annual International Conference on Computer Documentation*, SIGDOC '02, page 79–83, New York, NY, USA, 2002. Association for Computing Machinery.

[10] J. Kaiser. The preprint dilemma. *Science*, 357(6358):1344–1349, 2017.

[11] M. Khayyat and F. Bannister. Open data licensing: more than meets the eye. *Information Polity*, 20(4):231–252, 2015.

[12] M. Koskela, I. Simola, and K. Stefanidis. Open source software recommendations using github. In E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. C. Lopes, editors, *Digital Libraries for Open Knowledge*, pages 279–285, Cham, 2018. Springer International Publishing.

[13] P. Kraker, D. Leony, W. Reinhardt, and G. Beham. The case for an open science in technology enhanced learning. *International Journal of Technology Enhanced Learning*, 3(6):643–654, 2011.

[14] M. C. Linn, E. Palmer, A. Baranger, E. Gerard, and E. Stone. Undergraduate research experiences: Impacts and opportunities. *Science*, 347(6222):1261757, 2015.

[15] J. C. Molloy. The open knowledge foundation: Open data means better science. *PLOS Biology*, 9(12):1–4, 12 2011.

[16] P. Murray-Rust. Open data in science. *Nature Precedings*, 1(1):1, Jan 2008.

[17] E. National Academies of Sciences, P. Affairs, E. Committee on Science, B. Information, D. Sciences, C. Statistics, B. Analytics, D. Studies, N. Board, D. Education, et al. *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C., USA, 2019.

[18] B. A. Nosek, E. D. Beck, L. Campbell, J. K. Flake, T. E. Hardwicke, D. T. Mellor, A. E. van 't Veer, and S. Vazire. Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10):815–818, Oct 2019.

[19] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[20] B. A. Nosek, T. E. Hardwicke, H. Moshontz, A. Allard, K. S. Corker, A. Dreber, F. Fidler, J. Hilgard, M. Kline Struhl, M. B. Nuijten, J. M. Rohrer, F. Romero, A. M. Scheel, L. D. Scherer, F. D. Schönbrodt, and S. Vazire. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1):719–748, 2022. PMID: 34665669.

[21] P. Patil, R. D. Peng, and J. T. Leek. A statistical definition for reproducibility and replicability. *bioRxiv*, 1(1):1–1, 2016.

[22] B. Perens et al. The open source definition. *Open sources: voices from the open source revolution*, 1:171–188, 1999.

[23] J. K. Petrella and A. P. Jung. Undergraduate research: Importance, benefits, and challenges. *International journal of exercise science*, 1(3):91, 2008.

[24] G. Piateski and W. Frawley. *Knowledge Discovery in Databases*. MIT Press, Cambridge, MA, USA, 1991.

[25] G. K. Rajbahadur, E. Tuck, L. Zi, Z. Wei, D. Lin, B. Chen, Z. M. Jiang, and D. M. German. Can I use this publicly available dataset to build commercial AI software? most likely not. *CoRR*, abs/2111.02374:1–1, 2021.

[26] C. Romero and S. Ventura. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[27] S. H. Russell, M. P. Hancock, and J. McCullough. Benefits of undergraduate research experiences. *Science*, 316(5824):548–549, 2007.

[28] H. Schoettle. Open source license compliance-why and how? *Computer*, 52(08):63–67, aug 2019.

[29] B. A. Spellman. A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, 10(6):886–899, 2015. PMID: 26581743.

[30] A. E. van 't Veer and R. Giner-Sorolla. Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology*, 67:2–12, 2016. Special Issue: Confirmatory.

[31] R. Vicente-Saez and C. Martinez-Fuentes. Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88:428–436, 2018.

[32] A. Zuiderwijk and M. Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014.

APPENDIX

A. STANDARD PHRASES

This was a list of standard phrases used within the explanations document which was used to provide information or justifications on a given paper. The text might have been changed or further elaborated when used:

- The raw dataset and materials do not seem to be provided anywhere.
  - This was used when there is no information or links provided on the dataset or materials within the paper or its sub-resources. This might have also been used if it took longer than 15 minutes to located the associated resource(s).

- The raw dataset does not seem to be provided anywhere.
  - This was used when there is no information or links provided on the dataset within the paper or its subresources. This might have also been used if it took longer than 15 minutes to located the associated resource(s).
- The data documentation is likewise nonexistent.
  - This was used when there was no information within the paper on any documentation of the columns of the dataset. This was typically used in conjunction with papers that did not provide the dataset.
- Some data documentation is represent through <location>, and as such it will be marked as partial.
  - This was used when a column within the dataset was found to be marked in a paper or its subresources. The 'location' was replaced with the section or link the description was located.
- Open Materials include <materials>.
  - This was used whenever a paper contained materials that were not mentioned in the source or that the source was not provided for in the paper. The 'materials' was replaced with a list of the materials and links to their locations, if possible.
- The full analysis is not provided, so the materials fields will be marked as partial.
  - This was used when the source was unavailable when materials were present, or when the source did not seem to provide the ability to replicate all results provided within the paper.
- The paper seems to be argumentative in nature to create a new theoretical idea to use in the field. As such, all of the fields will be marked as non-applicable.
  - This was used when a paper talked about or elaborated on a concept rather than conduct an experiment or study. It marked all the available open science topics as non-applicable.

B. COMPUTER SPECIFICATIONS

B.1 Hardware Components

- AMD Ryzen Threadripper 2950X[21]
- NVIDIA GeForce RTX 3090[22]
- Corsair VENGEANCE LPX 128GB (4 x 32GB) DDR4 DRAM 2133MHz C18 Memory Kit
- WD Blue SN550 NVMe SSD (WDC WDS200T2B0C-00PXH0)[23]

B.2 Software Components

Some of the software components are considered the default if no specific version was specified in Section 3.3.

- Ubuntu 20.04.5 LTS[24]
- Linux Kernel 5.15.0-53-generic
- GNU bash 5.0.17(1)-release (x86_64-pc-linux-gnu)
- Python 3.8.10[25]
- R version 4.2.2 Patched (2022-11-10 r83330)[26]

[21] https://www.amd.com/en/product/7926
[22] https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3090-3090ti/
[23] https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/product/internal-drives/wd-blue-nvme-ssd/product-brief-wd-blue-sn550-nvme-ssd.pdf
[24] https://releases.ubuntu.com/focal/
[25] https://www.python.org/downloads/release/python-3810/
[26] https://cran.r-project.org/bin/linux/ubuntu/