Open Science and Educational Data Mining: Which Practices Matter Most?

Ryan S. Baker, Stephen Hutt, Christopher A. Brooks, Namrata Srivastava, Caitlin Mills
University of Pennsylvania, University of Denver, University of Michigan, University of Pennsylvania, University of Minnesota
ryanshaunbaker@gmail.com, stephen.hutt@du.edu, brooksch@umich.edu, namratas@upenn.edu, cmills@umn.edu

ABSTRACT

Open science has become an important part of contemporary science, and some open science practices (such as data sharing) have been prominent aspects of Educational Data Mining (EDM) since the start of the field. There have been recent pushes for EDM to more fully embrace the range of open science practices that are seen in other fields. In this paper, we review some of the practices that have become common in fields such as psychology, and critically examine the benefits and costs of adopting these practices within EDM. We conclude with a summary set of recommendations for the field.

Keywords

Open Science, Pre-registration, Data Sharing

1. INTRODUCTION

Though the idea can be arguably traced back to the 1600s when the first academic journals were formed, “open science” has become a popular topic in recent years [19, 66, 74]. Across fields, open science has become an increasingly important part of the way research is conducted and disseminated – with the core ideas being that research should be transparent and accessible. Such openness can take on many different forms – all of which are meant to make the processes and products of science more accessible and transparent for others – including publicly available data and/or code, and transparency in methods and materials for reproducibility. For example, authors are now routinely encouraged by conferences and journals to share their data so that findings can be replicated or extended, as well as pre-registering their hypothesis in an effort to avoid problematic practices like p-hacking that lead to high rates of Type 1 errors (i.e., claiming there is a significant result, when in fact only chance is present).

A general consensus now seems clear: scientific communities should have more open research practices in order to speed up progress and avoid situations where incorrect findings become widely believed [19, 66, 74]. Many scholars in EDM and related fields (such as learning analytics and artificial intelligence in education) have argued for wider use of open science practices [21, 29, 44, 76]. The potential benefits include more rapid improvement and decreased redundancy across scientific efforts, while also helping remove some of the barriers for early career researchers and less-funded areas of research in general. These benefits have led the EDM Society and other organizations in our field to recommend that scholars follow open science practices, for example in a one paragraph statement within the EDM2024 call for papers.

However, we have yet to reach consensus as to *which* open science practices EDM should adopt, as much of the work in our field does not fall under the traditional 5-step “scientific method” and null-hypothesis testing paradigms that have dominated open science narratives so far. The question we address here then is, *how should communities like EDM address open research practices*? In the remainder of this article, we discuss some of the key challenges and offer some recommendations for how researchers in our space may want to consider open science. We do not claim to have answered this question in full but hope to critique the notion that our community should blindly adopt what other fields are doing – or, at the other extreme, that we have already done enough. We instead argue that EDM should apply in a thoughtful and targeted manner the open research practices which are most impactful and suitable for the field of EDM.

2. BACKGROUND

Three events are commonly used to point out the existence of a reproducibility “crisis” in recent years [66]: 1) the discovery of many cases of impactful scientific fraud, 2) a series of papers highlighting questionable research practices (e.g., researcher degrees of freedom, HARKing) [38, 43], and 3) the Open Science Collaboration’s grim report that only about 36% of results of findings from top tier journals replicated [56]. Such events, though largely based in the field of psychology, have fueled interest in open science more broadly, where researchers have found similarly grim rates of reproducibility in other areas (including EDM) because of the lack of openness of data, code, or details about methods [19, 29, 30, 71].

One of the dominant narratives has been to “open up” research practices, and typically, this narrative leans strongly on recommendations based on issues from the field of psychology (see, for instance, the APA’s open science guidelines[1]). These common practices include: 1) pre-registration of and full disclosure of materials; 2) sophisticated a priori power analysis to determine sample sizes; 3) disclosure of non-significant findings in addition to significant ones; and 4) sharing data and code for reproducibility. Similar efforts have been made to establish open practices for qualitative

[1] https://www.apa.org/pubs/journals/resources/open-science

[16] and machine learning [60] as well. This has resulted in an increase of support for many open science practices, including the availability of free pre-registration sites (e.g., aspredicted.com), open data repositories (e.g., Open Science Framework), badge incentives on certain publication venues (e.g., EDM and ACM Learning At Scale), and requirements to engage in open sharing practices as a prerequisite to submission (e.g., through mandatory checklists on replication within the AAAI conference) and/or receiving funding.

However, there is some debate about whether this combination of approaches addresses the aforementioned problems and which of these practices are reasonable to apply to research processes in the field of EDM. Considering specifically the scope of work done in the EDM community, there are a number of different tensions in terms of the populations studied and research designs. For instance, whereas fields with strong laboratory research practices can set the nature and number of human participants, the fieldwork norm of education means that sample sizes are often not at the discretion of the researcher, and analyses of data are often secondary analyses and based on a combination of enrollment and parental consent. These aspects of our community's research make it difficult to engage in some open science practices, such as setting a strict a priori sample size number. Relatedly, deployment in education has an intrinsic context (measured or not), and combining contexts (e.g., classrooms, pupil populations, time periods) with the goal of establishing generality of findings has the potential to obscure context-dependent findings.

Beyond sample size, it is often complicated to share public data in communities like EDM, where legal and policy issues (e.g., IRB, FERPA, GDPR, COPPA) are nontrivial considerations and can materially affect research design. Even the nature of which research artifacts should be shared is not a straightforward decision and often involves questions ranging from intellectual property (e.g., instruments, source code, codebooks) to protections of the privacy of the subjects taking part in the research. Finally, specific behaviors considered problematic in other fields – such as p-hacking – may be less relevant to EDM given our field's methods, but may have analogues such as using extensive hyperparameter tuning on some algorithms but not others.

In order to make recommendations specifically for the EDM community's adoption and utilization of open science practices, it is worth reflecting on the authentic research activities EDM scholars engage in, and how these activities differ from other fields. For instance, *verification of previous research*, such as null hypothesis testing and direct replication of previous findings with the exact same methods, is an important activity in fields such as psychology but is a less common goal within EDM. Instead, EDM research more often involves the *iterative refinement of research,* where the computational methods, instruments, study design, and data collection methods are modified to narrow in on the phenomena being studied. Even when conceptually replicating previous approaches, many of the contributions within EDM aim to *generalize to new contexts*, which may include new populations, tasks, or systems being used. *Exploratory analysis of novel situations* is another common EDM goal, where the system being deployed to learners is unique, or the computational or learning paradigms being used are either new or borrowed from related fields and lack baselines, targets, or expected results within the new research context. There is opportunity to support open science across each of these kinds of activities, but wholesale adoption of practices from other fields may insufficiently support these kinds of investigations.

We note that the distinction between each of these activities is fuzzy at best, and that any given EDM study may involve multiple types of research and make multiple contributions to the field. Nonetheless, thinking of EDM research in these terms allows us to consider more clearly the implications a given open science practice may have if adopted by the community. In particular, doing so provides a scaffold for reasoning about the potential negative implications of some practices (e.g. pre-registration) when adopted for certain tasks (e.g. limiting the ability of researchers to engage in exploratory analysis of new situations).

The considerations seen in other fields are not always easily transferable; and acting as if they are may do more harm than good for open science in our field. In the next section, we consider five specific open science practices – open data, open code, pre-registration, multiple comparison analysis, and measurement of context – and how these practices might best be integrated into these research activities most frequently seen in EDM.

3. RELEVANT OPEN SCIENCE PRACTICES

The open science movement has become prominent in a range of fields, but different fields have emphasized different values and aspects of open science. For instance, there has been particular emphasis on pre-registration in order to avoid cherry-picking and p-hacking within psychology and classroom studies in education [24, 53, 72], whereas reproducibility of code has been seen as particularly important within computer science [39, 45, 49]. Researchers working at the intersection of artificial intelligence and education (i.e., individuals who publish at AIED, EDM, LAK) find themselves at the nexus of several of these trends, and the question of whether we are ultimately educational researchers or computer scientists becomes particularly acute, as each of us decides which practices are relevant and irrelevant for individual projects.

In the following section, we discuss some of the open science practices and their relevance for projects in our communities, with the ultimate goal of supporting transparency, replicability, and ultimately producing a community of scholars whose work builds upon each other and where scientific errors are easily identified and corrected over time.

3.1 Open Data

One of the core types of research within our communities is the analysis of large-scale data. Such analysis might include building student models [3, 26, 37], distilling insights about learners and learning [40, 63], or a range of other purposes. Indeed, the interest in new approaches for analyzing large-scale educational data and using that data analysis for a wide number of applications is probably the key reason why the educational data mining and learning analytics and knowledge conferences (and eventually, associated journals and scientific societies) emerged.

In the years prior to the development of the educational data mining community, only a small number of researchers had access to large-scale educational data, typically through personal or institutional connections to the organizations which stewarded the data. A new scientific community was only able to form and grow when access to at least some educational data became democratized, and a broader range of researchers could obtain access to large-scale data. One of the first such sources of large-scale open educational data was the Pittsburgh Science of Learning Center Datashop [47], which offered interaction log data from a variety of interactive learning environments (but most heavily from Cognitive Tutors). Another major source was data from individual universities' online courses [18, 34, 62]. Combined, these two data sources accounted for over a quarter of the data used in papers published in the first two years of the educational data mining conference [9].

These data sources – and a few other open data sources that followed, particularly interaction log data from the ASSISTments platform [33] – have continued to be omnipresent in publications in our communities. For example, a small set of data sets are utilized in the super-majority of recent papers on knowledge tracing [1, 50]. Other types of educational data are often less broadly available, and educational data remains open to researchers rarely enough that an annual competition was created by the International Educational Data Mining Society in 2021 to give an award to the best publicly available data set. However, despite this attempt by the community to publicize excellent new open data sets, the data sets that have won the competition in its 3 years are behavioral activity data from 2 interactive learning environments and 1 online course platform. Although these data sets are themselves novel in various ways – Prihar et al. [61], for example, involves data from randomized controlled trials conducted within an interactive learning environment – they still involve the same types of behavioral activity data that have generally been available to researchers in our communities for some time.

Not all papers in our communities come from these sources. Large numbers of papers involve other forms of data, including discussion forum posts [64], multimodal data involving sensors [68, 69], interview data [70], games [23], and teacher videos [17, 46]. Some individual papers in these areas of research involve data sets which are publicly shared or are otherwise obtainable (e.g., through data enclaves). However, the vast majority of such papers involve data sets that cannot be inspected or utilized by other researchers. This distorts the research that can be conducted within our field and also reduces the reproducibility and replicability of research that is conducted. Much of the research in these areas can only be done by those who can afford to collect these types of data (or, more accurately, those who have funders willing to pay for them to collect these types of data).

Changing this practice requires attention in advance as it can be difficult to share data once collected if the proper groundwork has not been laid. Two key challenges need to be addressed: authorization to share and deidentification of the data. First of all, the default preference for most institutional review boards (IRBs, a legal oversight requirement of much research in the United States, with analogies such as ethics boards in other countries) generally is to prevent open data access. Making data available has potential for disclosure, and thus, many IRBs will prefer that data not be shared, even in deidentified form (especially if data is from a protected class, such as children), and will not approve protocols that propose to share data outside of the original research group, either in advance or after the fact. One key step to getting IRB agreement to data sharing is to include provisions for future sharing within the original consent form used in a study (if there is such a consent form). In cases where no consent form is available for the original data collection (such as in learning platforms already being used at scale), providing end users (or their guardians) with a clear statement of how data is used for research prior to the student using the learning system can also facilitate future data sharing. Second, an increasing number of school districts now refuse to agree to student data being shared for research, based in some cases on local legislation and in other cases on direct lobbying of school districts [6]. While projects have attempted to create standardized IRB agreements and school agreements (e.g., the ASSISTments platform [33]), these efforts often fail to scale, as individual IRBs and schools often have "house rules" which require negotiations to be one-by-one and custom [14]. Still, the effort to achieve these agreements can have dividends for our field, expanding access to data to a broader range of researchers and contributors, who often have new and useful ideas.

Clarifying procedures for data deidentification can also make it more feasible to share data. Arguably, no data can ever be truly and conclusively considered deidentified (see, for instance, [75], which demonstrates the reidentification of a class's clickstream data using a newspaper article on a class field trip), but steps can be taken to reduce risks of reidentification. These steps vary in difficulty depending on the type of data represented. For example, video data can display student and teacher faces [4, 11]; keystroke data can be subject to reidentification through characteristic patterns of movement [41]; and within discussion forum and interview data, participants occasionally share identifiable information such as their employers or cities or names [13]. These limitations can be addressed through various human and automated processes. For instance, discussion forum and interview transcript data can be exhaustively checked by multiple humans for personally identifying information (PII) [15], and contemporary large language models can also be employed to check for PII [13]. AI technologies may also be able to obfuscate characteristic movements in sensor or keystroke data (perhaps at some loss of information, as per [48]) and to resolve faces into facial action units in ways that discard recognizable features [10, 51].

3.2 Open Analyses

A second core goal and principle is reproducibility and replicability of the analyses done on collected data. Within the EDM community, analyses are most commonly done through code (e.g. python, R), and sharing one's code is a positive practice which can enable scholarly review, promulgate novel techniques throughout the community, and decrease repetitive work.

However, simply sharing code isn't always enough to guarantee reproducibility – notebooks, for example, though they represent everything that was done, also allow researchers to run code out of order and in non-systematic ways [59, 73]. In addition, issues like code rot/decay, where code no longer runs due to changes in packages it depends on – [20], and dependency hell, where a package that code depends on other packages that have changed [12], can make existing code no longer function the same way. As papers such as [12] have discussed, these challenges can be hard to surmount if not planned for in advance, leading to many communities having very low proportions of code that are actually runnable after the fact. In a recent survey of papers published in educational data mining, [27] found that only 2.4% of papers studied involved code that could be run within a 6-hour time limit. This finding may seem overly pessimistic, as many code bases can be coaxed to run with further effort and cooperation from the original researchers; however, these researchers are not always available, able, or willing to cooperate with future researchers [22].

One partial solution is to use containerized approaches such as Docker [12, 35] or Kubernetes [65]. These containerized approaches contain not just the code itself, but the libraries needed for the code to run. Developed carefully, a combination of source code and a container can make it so that code can be run in many computational contexts and continue to be used for longer. However, these approaches can create significant barriers to entry, as many researchers (even with programming backgrounds) find Docker and other containerization approaches challenging to use. For example, the MORF project [35] has found that the use of Docker represents one of the largest barriers to external users, even after providing a range of code examples.

3.3 Pre-registration

One methodological approach that is heavily used in other fields is pre-registration, where the full details of an upcoming study and analysis methods are published in full, before any data is collected [53]. Pre-registration as a practice is now widespread (if still not dominant) in a range of fields, from psychology to medicine to political science [54, 2, 55]. Pre-registration is cited as a way to prevent p-hacking and, when combined with journals that accept articles based on pre-registration (generally called registered reports [5, 52, 57]), it creates incentives to commit to an analysis approach before the fact, rather than re-running analyses in several ways after collecting data. There have, therefore, been recent calls to adopt pre-registration in our communities as well [28, 32, 31, 29].

However, pre-registration may not be suitable for all kinds of investigations being undertaken by the EDM community. On the positive side, some research is intended to be confirmatory in nature, closely tracking the classic "scientific method" – an initial hypothesis is generated, an experiment is developed to test whether that hypothesis is true or false (with some set of findings clearly not supporting the hypothesis), the experiment is conducted, data is analyzed, and a conclusion on a small set of pre-defined questions is obtained. This is the case which pre-registration was originally designed for [72] and it is a case where pre-registration makes a great deal of sense. Considerable research in educational technology – more in AIED/IJAIED than in the EDM or LAK communities – is of this nature.

Other research in our community is concerned with generalization, and involves studies where a method or intervention is reproduced, but with one or more aspects of the sample or the instrument intentionally manipulated [37]. For instance, an intervention that worked in one population of learners may be tested with a different group of learners, or it may be applied later in the course than was the case in the initial study. In the case of a machine learned model, a model that performed better than other alternatives in one learning system or population may be tested in different learning systems or populations. For these machine learning cases, preregistration may again make sense and be appropriate, but there are some limitations. First, it can be hard to verify that pre-registration is taking place prior to research rather than after, since no data collection, IRB, or classroom procedures need to take place after preregistration – it comes down to the researcher attesting that the analysis (which generally could be done at any time) occurred after pre-registration rather than before. Second, generalization studies often raise additional questions as to the mechanisms of findings that may not have been clear during the planning phase.

A third, very common (see [8]) category of research in our community involves prediction modeling. In some cases, it may be appropriate to pre-register such an analysis with some of the same caveats as generalization research (e.g., the difficulty of verifying whether an analysis was done before or after pre-registration). Doing so may, in fact, save quite a bit of time for a researcher. Take, for example, a researcher planning a knowledge tracing analysis where their plans for training/test validation or comparison algorithms do not match guidelines or standards in the field (see next subsection for a further discussion of this). If they submit their analysis design for a "registered reports" review at pre-registration, these issues may be caught before they do the work. This would avoid the situation where the author's paper is rejected by EDM/LAK/AIED, requiring the author to redo their analyses and resubmit (as well as the situation where they choose not to redo their analyses and simply resubmit as-is to a lower-tier venue). However, in other cases, pre-registration of prediction modeling may have negative impacts. Take, for example, the case where prediction modeling is focused on engineering an automated detector of a complex construct. Within this type of prediction modeling, feature engineering is an iterative process (see, for instance, [67]) where model developers use their initial results to refine their feature engineering plan. In this case, preregistration would limit a researcher to a set of variables chosen in advance, hampering their ability to improve their model to the best degree possible. In a situation of this nature, it may be better to do model development thoroughly and without limitations in a first data set and then collect an entirely new data set and test the model with that data set (perhaps pre-registering that final step, but it is not clear that preregistering the final step is as important as collecting the new data set).

A fourth, again very common (see [8]) category of research in our community involves exploratory research. Many of the methods of EDM and LAK – clustering or association rule mining for instance – are entirely exploratory, and researchers do not know what they will find when they apply these methods. These methods are not designed to be used to make claims about generality, do not have statistical tests or anything analogous, and are generally considered propositive (i.e., generating new ideas or lines of investigation) rather than dispositive (i.e., selecting between alternate hypotheses). As such, pre-registration seems inappropriate for this type of research and may discourage researchers from following up interesting preliminary findings with more in-depth exploratory analysis.

Overall, pre-registration is designed to prevent a specific type of problematic research – p-hacking and reporting secondary analyses as if they were primary. For cases where there is essentially no *p* to hack, and analyses are acknowledged to be exploratory or engineering-focused (e.g. prediction modeling for detector development), pre-registration seems unnecessary and even limiting.

3.4 Avoiding Comparison Hacking

A practice that is similar to p-hacking, but perhaps of more immediate concern to the types of secondary data analysis predominant within educational data mining and learning analytics, is a practice we refer to as *comparison hacking*. In comparison hacking, a researcher is working on a new algorithm variant, and finds ways to make their new algorithm variant perform better than previous algorithm variants. Comparison hacking can take on two forms, which are not mutually exclusive.

First, a researcher can conduct comparisons within a single data set and try a range of different variants until they find one that performs substantially better than past approaches. This form of comparison hacking does not distort the past approaches' performance, but essentially over-fits at the researcher level rather than the level of an individual algorithm, by trying approach after approach until one performs better. This approach, which is somewhat analogous to p-hacking, has also been referred to as "graduate student descent" [25].

In its second form, a researcher compares a single new algorithm variant to other existing algorithm variants, but adjusts the flexibility of fit of their algorithm or for the testing procedure. For instance, a researcher may use existing published hyperparameters for other algorithms but tune their own algorithm's hyperparameters to the current data. This practice can even be justified as representing prior algorithms fairly by not modifying them in any fashion, but in practice doing so gives the new algorithm the scope to better fit to the current data set than previous algorithms. Alternatively, a researcher may search for different ways to conduct testing, such as different training/test splits or different random seeds, in order to find the method that makes their algorithm variant appear to perform better than other algorithm variants.

This second form of comparison hacking appears, at a glance, to be common within research on knowledge tracing, particularly recent research on deep knowledge tracing (DKT) variants that appear outside of the premier publication venues in our community. There are a surprising number of papers that introduce new DKT variants that are slightly or subtly different than existing variants yet obtain much better reported performance, only to drop substantially in performance in the next paper which itself reports much better performance for its own new approach, all with subtly different evaluation methods. While individual papers bearing these characteristics may be innocent of any ill-intended attempts at comparison hacking, the overall pattern of published papers in this area suggests that this practice is widespread.

What can be done? One option is to bring together a group of researchers in the area to agree on fair rules and guidelines for comparing algorithms in this area, as well as guidelines for when these rules should be applied (for example, comparing multiple new variants of an algorithm to each other in order to answer scientific questions about a new mechanism could involve different rules than comparing a single new algorithm variant to existing best practice). These guidelines could then be communicated to reviewers as an expectation for the practices that future papers should follow. This approach would not prevent non-compliant papers from appearing at other venues, where reviewers may be unaware of the guidelines, but could provide a quick way for researchers to identify which papers are more likely to be free from comparison hacking.

Another approach, already in progress, is for periodic “neutral” comparisons to be conducted by a group with no specific ties to any algorithm variant, on a range of different data sets. A particularly high-quality example of such a comparison is seen in Gervet et al. [26].

3.5 Representing Context

A fifth important type of open science practice is fully representing the context of a study, not just so that it can be replicated, but also so that the findings can be better understood. Doing so enables researchers conducting later research syntheses to be able to study the factors that lead to different studies obtaining contradictory results (see for instance, findings in [42], that showed that differences in affect dynamics patterns reported depended on learners’ nationality, age, and research setting). Unfortunately, papers in our communities often do a poor job at reporting on the learners who are being studied (see [58]) and the learning systems that contribute the data being studied. Given the lack of generalization of many findings, high quality and comprehensive discussion (or availability of information on) of specific studies will help us to understand why findings manifest in some cases but not others (again, as in [42]).

Three forms of context may be relevant to provide information on. A first relevant area is the learner experience and instrument of learning which was studied. Many papers provide a screenshot and a one paragraph representation of the task, but this information is often insufficient. Ideally, a study will report extensive details of the learning activity, task, or platform; on the pedagogy it utilizes; on the content being taught; how it was communicated to students and integrated into their regular learning curriculum (where appropriate). In cases (such as conference proceedings) where space is inadequate, this sort of information can be provided in supplemental materials sections or in a repository within GitHub or the PSLC DataShop. No standard exists for all of these types of reporting, but by providing additional information, the authors make it easier for later researchers to compare studies and make hypotheses for why findings do not generalize.

Second, it is relevant to report on the learners being studied. Baker and Hawn [7] and the Penn Center for Learning Analytics wiki[2] report on the range of factors for which algorithmic bias has been reported in educational technology, including race, ethnicity, gender, socioeconomic status, national origin, native language and dialect, urbanicity, migrant and military-connected status, type of school (public or private), and parental educational background. Algorithmic bias serves as an indicator of differences in how students interact with the learning experience, or other differences in the meaning of variables, and as such, these known factors (as well as other factors such as religion or the experience of being a minority in one’s educational setting) are likely to influence the results of studies in this space and as such may be appropriate to report to understand the differences between studies.

Finally, it is relevant to report on contextual factors that represent an interaction between the learner themselves and the content they experience. For instance, past aspects of a learner’s educational background and socioeconomic status, as well as the region where they grew up, may lead to differences in prior knowledge of the content being experienced in the learning system, differences in student interest in the topic being studied, or different levels of initial familiarity and comfort with the pedagogy being used. We do not as a field have a full representation of the types of variables of this nature which are relevant for our field’s work, but reporting on as many of these as is practically feasible is likely to enhance later attempts to synthesize across findings.

4. OPEN SCIENCE AND THE FUTURE OF EDM

As can be seen, there are a range of open science practices that could be adopted in EDM. Within this paper, we have discussed some of the most common and/or relevant practices. It is important to note that the activities outlined in the prior sections of this paper are not intended to be restrictive or exhaustive. Instead, this work aims to be illustrative of both the benefits (and risks) of adopting open science practices in order to encourage reflection among all of us about how we can implement open science in a responsible way that makes sense for our field.

One key theme in our discussion is that EDM research and the work of our community diverges from the traditional research norms seen in Psychology and similar disciplines. While Psychology’s replication crisis has been a catalyst to the rise of open science practices (and subsequent proposed solutions), EDM should not simply mimic these approaches. There is a potential danger to trying to fit

[2] https://www.pcla.wiki/index.php/Algorithmic_Bias_in_Education

our work into the mold of another discipline; doing so may hinder much of our research in undesirable ways. As discussed above, pre-registration is a clear example of a practice that makes great sense in many contexts, but may be inappropriate for many types of EDM research. Learning from other disciplines is valuable, but it is crucial for EDM to tailor its open science goals and practices to our field's particular needs and challenges, such as coping with strict data agreements and addressing algorithmic biases that negatively impact protected groups. We recommend that the community (perhaps with support from a collective of related societies) work to articulate our open science goals and best practices formally – or at least more formally than what exists currently. Forging a clear, field-specific understanding of what open science means and should mean to us will help us avoid the pitfall of measuring our success by the standards of other fields.

Once we have achieved this understanding, we then must embed the appropriate practices, expectations, and norms in the community. The conference and journal proceedings are artifacts of the community, and thus, these are the most impactful places we can push the field towards appropriate practices. Pulling from our own field and our scientific knowledge of self-regulated learning, if individuals must reflect on their practice throughout the process, they will more critically examine the steps they will engage in. One method of doing this is to go beyond current statements in calls for papers and badges to require authors to complete an open science checklist – similar to other fields' reproducibility checklists [60]- at the time of paper submission. Even if reviewers were to pay minimal attention to those checklists, in the theme of SRL, this checklist could still have a positive effect on author practices, as reflection alone is powerful for conceptual change and skill development.

As we publish (and review) in these academic spaces, we have a responsibility to keep open science and replicability front and center. It is imperative that our work provide comprehensive retrospective reports accurately detailing our full processes – both what was successful and what was unsuccessful. Sharing a full account of work performed will aid interpretation and building on past work. It is critical for these descriptions to be thorough; publications or supplementary materials must include sufficient detail to enable the exact replication of our work, as well as reproducible code. While excessive technical details can sometimes disrupt the flow and coherence of a manuscript, this challenge can be effectively addressed by including supplementary documents/appendices or providing digital repositories and code for precise replication instructions. Even in terms of the process of open science itself, we have a responsibility to be more thorough and share the challenges we have faced in implementing these practices; it is only through understanding the challenges of the field that we can create a robust path forward.

In terms of supporting reproducibility and reuse of code, Docker (or Kubernetes) containers present considerable utility for promoting long-term runability. We propose that the field should move towards sharing Docker containers as a standard rather than sharing code or repositories in isolation. Docker containers can complement these existing code sharing methodologies. However, creating Docker containers does have a "barrier for entry". As mentioned above, there is some technical skill required to create (and use) such containers. As a community, we must ensure that if we do decide that Docker containers present a best practice for the field, we also provide adequate support so that this practice does not become exclusionary to some researchers.

When it comes to data, Open Data is a desirable goal for our field, but it is necessary to acknowledge the numerous constraints upon how data is shared. Data sharing needs to be considered before data is even collected when signing data agreements in order to achieve open data. Privacy-preserving data enclaves (such as MORF [35]) may aid us in making data available for reproduction and reuse, without directly sharing. It is also important to note that changes in legislation and policies surrounding data are likely to impact data sharing and data sharing practices (see discussion in [36]); as regulations change, the challenges for researchers and practitioners in our field are likely to change as well.

5. CONCLUSION

In conclusion, open science is a worthwhile endeavor for almost all data-driven scientific fields. EDM/AIED/LAK are no different. In this paper, we argue that we need to reflect on our open science practices to ensure that our community decides on standard and expected practices thoughtfully and appropriately, so that the practices that are recommended, encouraged, and (hopefully) systematically adopted reflect the nuances of our field. Such an approach should be bottom-up, with community members sharing their challenges and contributing to the definition of our community goals to improve open science. But as we move towards consensus, building these practices into conference and journal reviewing processes will help us to build a field where research is open, transparent, robust, and valid.

6. ACKNOWLEDGMENTS

We would like to thank Chelsea Porter for assistance in document preparation and editing.

7. REFERENCES

[1] Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge Tracing: A Survey. *ACM Computing Surveys, 55*(11), 1–37. https://doi.org/10.1145/3569576

[2] Al-Durra, M., Nolan, R. P., Seto, E., & Cafazzo, J. A. (2020). Prospective registration and reporting of trial number in randomised clinical trials: Global cross sectional study of the adoption of ICMJE and Declaration of Helsinki recommendations. *BMJ*, m982. https://doi.org/10.1136/bmj.m982

[3] Almoubayyed, H., Fancsali, S., & Ritter, S. (2023). Generalizing Predictive Models of Reading Ability in Adaptive Mathematics Software. *Proceedings of the 16th International Conference on Educational Data Mining*, 207–216. https://doi.org/10.5281/ZENODO.8115782

[4] Andrejevic, M., & Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology, 45*(2), 115–128. https://doi.org/10.1080/17439884.2020.1686014

[5] *APS Registered Reports*. (n.d.). Association for Psychological Science -– APS. Retrieved January 5, 2024, from https://www.psychologicalscience.org/publications/replication

[6] Baker, R.S. (2023) The Current Trade-off Between Privacy and Equity in Educational Technology. In G. Brown III, C. Makridis (Eds.) The Economics of Equity in K-12 Education: Necessary Programming, Policy, and Systemic Changes to Improve the Economic Life Chances of American Students, pp. 123-138. Lanham, MD: Rowman & Littlefield.

[7] Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*, 1-41.

[8] Baker, R.S.J.d., Inventado, P.S. (2014) Educational Data Mining and Learning Analytics. In J.A. Larusson, B. White (Eds.) *Learning Analytics: From Research to Practice*. Berlin, Germany: Springer

[9] 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.

[10] Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019

[11] Banzon, A. M., Beever, J., & Taub, M. (2023). Facial Expression Recognition in Classrooms: Ethical Considerations and Proposed Guidelines for Affect Detection in Educational Settings. *IEEE Transactions on Affective Computing*, 1–13. https://doi.org/10.1109/TAFFC.2023.3275624

[12] Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review, 49*(1), 71–79. https://doi.org/10.1145/2723872.2723882

[13] Bosch, N., Crues, R. W., & Shaik, N. (2020). "Hello, [REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 39–49.

[14] Briggs, R. (2022, March). The Abject Failure of IRBs. *The Chronicle of Higher Education.* Retrieved February 5, 2024, from https://www.chronicle.com/article/the-abject-failure-of-irbs

[15] Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., & Goodman-Williams, R. (2023). Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data. *Advances in Methods and Practices in Psychological Science, 6*(4), 25152459231205832. https://doi.org/10.1177/25152459231205832

[16] Class, B., de Bruyne, M., Wuillemin, C., Donzé, D., & Claivaz, J.-B. (2021). Towards Open Science for the Qualitative Researcher: From a Positivist to an Open Interpretation. *International Journal of Qualitative Methods, 20*, 16094069211034641. https://doi.org/10.1177/16094069211034641

[17] Colliot, T., & Jamet, É. (2018). Understanding the effects of a teacher video on learning from a multimedia document: An eye-tracking study. *Educational Technology Research and Development, 66*(6), 1415–1433. https://doi.org/10.1007/s11423-018-9594-x

[18] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain* (pp. 41–50).

[19] Echtler, F., & Häußler, M. (2018). Open Source, Open Science, and the Replication Crisis in HCI. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3170427.3188395

[20] Eick, S. G., Graves, T. L., Karr, A. F., Marron, J. S., & Mockus, A. (2001). Does code decay? Assessing the evidence from change management data. *IEEE Transactions on Software Engineering, 27*(1), 1–12. https://doi.org/10.1109/32.895984

[21] Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., … & Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science, 4*(3), 25152459211027575.

[22] Gardner, J., Brooks, C., Andres, J. M., & Baker, R. (2018). Replicating MOOC predictive models at scale. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10. https://doi.org/10.1145/3231644.3231656

[23] Geden, M., Emerson, A., Rowe, J., Azevedo, R., & Lester, J. (2020). Predictive Student Modeling in Educational Games with Multi-Task Learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 654–661. https://doi.org/10.1609/aaai.v34i01.5406

[24] Gehlbach, H., & Robinson, C. D. (2018). Mitigating Illusory Results through Preregistration in Education. *Journal of Research on Educational Effectiveness, 11*(2), 296–315. https://doi.org/10.1080/19345747.2017.1387950

[25] Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., Leinonen, J., & Huttunen, H. (2019). *HARK Side of Deep Learning—From Grad Student Descent to Automated Machine Learning.* https://doi.org/10.48550/ARXIV.1904.07633

[26] 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. https://doi.org/10.5281/ZENODO.4143614

[27] Haim, A., Gyurcsan, R., Baxter, C., Shaw, S. T., & Heffernan, N. T. (2023a). How to Open Science: Debugging Reproducibility within the Educational Data Mining Conference. *Proceedings of the 16th International Conference on Educational Data Mining (EDM '23)*, 114–124. https://doi.org/10.5281/ZENODO.8115651

[28] Haim, A., Heffernan, N., & Shaw, S. T. (2022). *How to Open Science: Promoting Principles and Reproducibility Practices within the Learning Analytics Community*. https://doi.org/10.17605/OSF.IO/KYXBA

[29] Haim, A., Shaw, S., & Heffernan, N. (2023a). *How to Open Science: Promoting Principles and Reproducibility Practices within the Educational Data Mining Community.* https://doi.org/10.5281/ZENODO.8115776

[30] Haim, A., Shaw, S. T., & Heffernan, N. (2023b). *How to Open Science: A Reproducibility Author Survey of the Artificial Intelligence in Education Conference.* https://edarxiv.org/xkmfw/download?format=pdf

[31] Haim, A., Shaw, S. T., & Heffernan, N. T. (2023c). How to Open Science: Promoting Principles and Reproducibility Practices Within the Artificial Intelligence in Education Community. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking*

Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky (Vol. 1831, pp. 74–78). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_11

[32] Haim, A., Shaw, S. T., & Heffernan, N. T. (2023d). How to Open Science: Promoting Principles and Reproducibility Practices within the Learning @ Scale Community. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 248–250. https://doi.org/10.1145/3573051.3593398

[33] Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

[34] Hershkovitz, A., & Nachmias, R. (2009). Consistency of Students' Pace in Online Learning. In *Proceedings of the 2nd International Conference on Educational Data Mining* (pp. 71–80).

[35] Hutt, S., Baker, R. S., Ashenafi, M. M., Andres-Bray, J. M., & Brooks, C. (2022). Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology, 53*(4), 756–775. https://doi.org/10.1111/bjet.13231

[36] Hutt, S., Das, S., & Baker, R. S. (2023). The Right to Be Forgotten and Educational Data Mining: Challenges and Paths Forward. *Proceedings of the 16th International Conference on Educational Data Mining, EDM 2023.* https://eric.ed.gov/?id=ED630886

[37] Hutt, S., Grafsgaard, J. F., & D'Mello, S. K. (2019). Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300726

[38] Ioannidis, J. P. (2005). Why most published research findings are false. *PloS Medicine, 2*(8), e124.

[39] Ivie, P., & Thain, D. (2019). Reproducibility in Scientific Computing. *ACM Computing Surveys, 51*(3), 1–36. https://doi.org/10.1145/3186266

[40] Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C., & Brooks, C. (2018). How Do We Model Learning at Scale? A Systematic Review of Research on MOOCs. *Review of Educational Research, 88*(1), 43–86. https://doi.org/10.3102/0034654317740335

[41] Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing, 11*(2), 1565–1573. https://doi.org/10.1016/j.asoc.2010.08.003

[42] Karumbaiah, S., Baker, R.S., Ocumpaugh, J., Andres, J.M.A.L. (2023) A Re-Analysis and Synthesis of Data on Affect Dynamics in Learning. *IEEE Transactions on Affective Computing, 14*(2), 1696-1710.

[43] Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

[44] Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023, March). Towards more replicable content analysis for learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 303-314).

[45] Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. University of California Press.

[46] Kizilcec, R. F., Bailenson, J. N., & Gomez, C. J. (2015). The instructor's face in video instruction: Evidence from two large-scale field studies. *Journal of Educational Psychology, 107*(3), 724–739. https://doi.org/10.1037/edu0000013

[47] Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining, 43*, 43-56.

[48] Leinonen, J., Ihantola, P., & Hellas, A. (2017). Preventing Keystroke Based Identification in Open Data Sets. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, 101–109. https://doi.org/10.1145/3051457.3051458

[49] LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering, 14*(4), 13–17. https://doi.org/10.1109/MCSE.2012.38

[50] Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). *A Survey of Knowledge Tracing*. https://doi.org/10.48550/ARXIV.2105.15106

[51] Mase, J. M., Leesakul, N., Figueredo, G. P., & Torres, M. T. (2023). Facial identity protection using deep learning technologies: An application in affective computing. *AI and Ethics, 3*(3), 937–946. https://doi.org/10.1007/s43681-022-00215-y

[52] *Nature Human Behaviour Registered Reports*. (n.d.). Retrieved January 5, 2024, from https://www.nature.com/nathumbehav/submission-guidelines/registeredreports

[53] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

[54] Nosek, B. A., & Lindsay, D. S. (2018). Preregistration Becoming the Norm in Psychological Science. *APS Observer*. https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

[55] Nyhan, B. (2015). Increasing the Credibility of Political Science Research: A Proposal for Journal Reforms. *PS: Political Science & Politics, 48*(S1), 78–83. https://doi.org/10.1017/S1049096515000463

[56] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

[57] *Open Science Registered Reports*. (n.d.). Retrieved January 5, 2024, from https://www.cos.io/initiatives/registered-reports

[58] Paquette, L., Ocumpaugh, J., Li, Z., Andres, J.M.A.L., Baker, R.S. (2020) Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining, 12*(3), 1-30.

[59] Pimentel, J. F., Murta, L., Braganholo, V., & Freire, J. (2019). A Large-Scale Study About Quality and

Reproducibility of Jupyter Notebooks. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 507–517. https://doi.org/10.1109/MSR.2019.00077

[60] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving Reproducibility in Machine Learning Research (a Report from the NeurIPS 2019 Reproducibility Program). *The Journal of Machine Learning Research, 22*(1).

[61] Prihar, E., Syed, M., Ostrow, K., Shaw, S., Sales, A., & Heffernan, N. (2022). Exploring common trends in online educational experiments. In *Proceedings of the 15th International Conference on Educational Data Mining*.

[62] Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data Mining Algorithms to Classify Students. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational Data Mining 2008, The 1st International Conference on Educational Data Mining, Montreal, Québec, Canada, June 20-21, 2008. Proceedings* (pp. 8–17). www.educationaldatamining.org.

[63] Sabourin, J. L., & Lester, J. C. (2014). Affect and Engagement in Game-Based Learning Environments. *IEEE Transactions on Affective Computing, 5*(1), 45–56. https://doi.org/10.1109/T-AFFC.2013.27

[64] Sha, L., Raković, M., Lin, J., Guan, Q., Whitelock-Wainwright, A., Gašević, D., & Chen, G. (2023). Is the Latest the Greatest? A Comparative Study of Automatic Approaches for Classifying Educational Forum Posts. *IEEE Transactions on Learning Technologies, 16*(3), 339–352. https://doi.org/10.1109/TLT.2022.3227013

[65] Shah, J., & Dubaria, D. (2019). Building Modern Clouds: Using Docker, Kubernetes & Google Cloud Platform. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 0184–0189. https://doi.org/10.1109/CCWC.2019.8666479

[66] Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology, 69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

[67] Slater, S., Baker, R.S., Wang, Y. (2020) Iterative Feature Engineering Through Text Replays of Model Errors. *Proceedings of the 13th International Conference on Educational Data Mining*, 503-508.

[68] Srivastava, N., Nawaz, S., Lodge, J. M., Velloso, E., Erfani, S., & Bailey, J. (2020). Exploring the usage of thermal imaging for understanding video lecture designs and students' experiences. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 250–259. https://doi.org/10.1145/3375462.3375514

[69] Srivastava, N., Nawaz, S., Newn, J., Lodge, J., Velloso, E., M. Erfani, S., Gasevic, D., & Bailey, J. (2021). Are you with me? Measurement of Learners' Video-Watching Attention with Eye Tracking. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 88–98. https://doi.org/10.1145/3448139.3448148

[70] Tsai, Y.-S., Singh, S., Rakovic, M., Lim, L.-A., Roychoudhury, A., & Gasevic, D. (2022). Charting Design Needs and Strategic Approaches for Academic Analytics Systems through Co-Design. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 381–391. https://doi.org/10.1145/3506860.3506939

[71] Wacharamanotham, C., Eisenring, L., Haroz, S., & Echtler, F. (2020). Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376448

[72] Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Van Der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

[73] Wang, J., Li, L., & Zeller, A. (2020). Better code, better sharing: On the need of analyzing Jupyter notebooks. *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*, 53–56. https://doi.org/10.1145/3377816.3381724

[74] Wilson, M. L. L., Resnick, P., Coyle, D., & Chi, E. H. (2013). *RepliCHI: The workshop. CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 3159–3162. https://doi.org/10.1145/2468356.2479636

[75] Yacobson, E., Fuhrman, O., Hershkovitz, S., & Alexandron, G. (2021). De-identification is Insufficient to Protect Student Privacy, or – What Can a Field Trip Reveal? *Journal of Learning Analytics, 8*(2), 83–92. https://doi.org/10.18608/jla.2021.7353

[76] Kitto, K., Manly, C.A., Ferguson, R., & Poquet, O. (2023) Towards more replicable content analysis for learning analytics. *Proceedings of the 13th Annual Learning Analytics and Knowledge Conference,* 303-314.