How to Open Science: Promoting Principles and Reproducibility Practices within the Educational Data Mining Community

Aaron Haim
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts, USA
ahaim@wpi.edu

Stacy T. Shaw
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts, USA
sshaw@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

Across the past decade, open science has increased in momentum, making research more openly available and reproducible. Educational data mining, as a subfield of education technology, has been expanding in scope as well, developing and providing better understanding of large amount of data within education. However, open science and educational data mining do not often intersect, causing a bit of difficulty when trying to reuse methodologies, datasets, analyses for replication, reproduction, or an entirely separate end goal. In this tutorial, we will provide an overview of open science principles and their benefits and mitigation within research. In the second part of this tutorial, we will provide an example on using the Open Science Framework to make, collaborate, and share projects. The final part of this tutorial will go over some mitigation strategies when releasing datasets and materials such that other researchers may easily reproduce them. Participants in this tutorial will gain a better understanding of open science, how it is used, and how to apply it themselves.

Keywords

Open Science, Reproducibility, Preregistration

1. BACKGROUND

Open Science is a term used to encompass making methodologies, datasets, analyses, and results of research publicly accessible for anyone to use freely[6, 14]. This term started to frequently occur in the early 2010s when researchers began noticing that they were unable to replicate or reproduce prior work done within a discipline[13]. There also tended to be a large amount of ambiguity when trying to understand what process was followed to conduct a study or whether a specific material was used but not clearly defined. Open science, as a result, started to gain more traction to provide greater context, robustness, and reproducibility metrics with each subtopic encompassed under the term receiving their own formal definition and usage. The widespread adoption of open science began to explode exponentially when large scale studies conducted in the mid 2010s found that numerous works were difficult or impossible to reproduce and replicate in psychology[2] and other disciplines[1].

Some principles commonly referred to as part of open science and its processes: open data, open materials, open methodology, and preregistration. **Open Data** specifically targets datasets and their documentation for public use without restriction, typically under a permissive license or in the public domain[8]. Not all data can be openly released (such as with personally identifiable information); but there are specifications for protected access that allow anonymized datasets to be released or a method to obtain the raw dataset itself. **Open Materials** is similar in regard except for targeting tools, source code, and their documentation[5]. This tends to be synonymous with **Open Source** in the context of software development, but materials are used to encompass the source in addition to available, free-to-use technologies. **Open Methodology** defines the full workflow and processes used to conduct the research, including how the participants were gathered, what was told to them, how the collected data was analyzed, and what the final results were[6]. The methodologies typically expand upon the original paper, such as technicalities that would not fit in the paper format. Finally, **Preregistration** acts as an initial methodology before the start of an experiment, defining the process of research without knowledge of the outcomes[10, 11]. Preregistrations can additionally be updated or created anew to preserve the initial experiment conducted and the development as more context is generated.

2. TUTORIAL GOALS

Open science principles and reproducibility metrics are becoming more commonplace within numerous scientific disciplines. Within many subfields of educational technology, such as educational data mining, however, the adoption and review of these principles and metrics are neglected or sparsely considered[9]. There are some subfields of education technology that have taken the initiative to introduce open science principles (special education[3]; gamification[4], education research[7]); however, other subfields have seen little to no adoption. Concerns and inexperience in what can be made

A. Haim, S. Shaw, and N. Heffernan. How to open science: Promoting principles and reproducibility practices within the educational data mining community. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 582–584, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115776

publicly available to how to reproduce another's work are some of the few reasons why researchers may choose to avoid or postpone discussion on open science and reproducibility. On the other hand, lack of discussion can lead to tediousness and repetitive communication for datasets and materials or cause a reproducibility crisis[1] within the field of study. As such, there is a need for accessible resources and understanding on open science, how it can be used, and how to mitigate any potential issues that may arise within one's work at a later date.

Admitting our own initial lack of proper adoption and reproducibility first, in this tutorial, we will cover some of the basic principles of open science and some of the challenges and mitigation strategies associated with education technology specifically. Next, we will provide a step-by-step explanation on using the Open Science Framework to create a project, collaborate with other researchers, post content, and preregister a study. Using examples from the field of educational technology, we will showcase how to incorporate open science principles, in addition to practices that, when implemented, would improve reproducibility.

This tutorial will build and expand on a prior, successful tutorial at the *15th International Conference on Educational Data Mining* in 2022[1][12] and an accepted tutorial to be presented at the *13th International Conference on Learning Analytics and Knowledge* in 2023[2].

3. TUTORIAL ORGANIZATION

The tutorial will occur over half a day and focuses on introducing some common open science principles and their usage within education technology, providing an example on using the Open Science Framework to create a project, post content, and preregister studies, and using previous papers to apply the learned principles and any additional reproduction mitigation strategies. An outline of this tutorial can be found below:

- First, we will provide a presentation on an overview of a few problems when conducting research. Using this as a baseline, we will introduce open science and its principles and how they can be used to nullify some of these issues and mitigate others. In addition, we will attempt to dispel some of the misconceptions of these principles.

- Second, we will provide a live example of using the Open Science Framework (OSF) website to make an account, create a project, add contributors, add content and licensing, and publicize the project for all to see. Afterwards, we will provide a guide to creating a preregistration, explaining best practices, and identifying how to create an embargo. Additional features and concerns, such as anonymizing projects for review and steps required to properly do so, will be shown.

- Third, we will discuss reproducibility metrics within work when providing datasets and materials. This will review commonly used software and languages (e.g. Python, RStudio) and how, without any steps taken, most work tends to be extremely tedious to reproduce or are not reproducible in general. Afterwards, we will provide some mitigation strategies needed to remove these concerns.

- Finally, we will take some existing papers either from the author's own research or from prior education technology conferences that do not meet some open science principles or cannot easily be reproduced and apply what has been learned across the entire tutorial. We will use a few papers, each containing different issues, and apply the necessary steps needed to reproduce the results within the paper.

3.1 Dissemination of Information

The dissemination of information for this tutorial will be provided before and after the conference. Before the conference, information about the tutorial itself will be stored on an OSF project, containing references to the papers used within the final part of the tutorial, any slides to be used within the conference, and additional resources that could provide better understanding of the issues and nuances of avoiding open science and reproducibility metrics. A website separate to the OSF project will also be set up containing the following information for ease of consumption; however, this will only be used as an alternative to the project in case the website disappears at some point in the future.

After the conference, any resources created or recordings taken will be uploaded to the project for preservation. Alternative links will be provided to separate sites for more formal hosting (e.g. videos on YouTube). As this tutorial wants to repeat and expand upon open science and reproducibility at prior workshops across conferences, an additional project will be created on the OSF website containing components pointing to all previous conferences and resources discussed.

3.2 Organizers

Aaron Haim[3] is a Ph.D. student in Computer Science at Worcester Polytechnic Institute. His initial research focuses on developing software and running experiments on crowdsourced, on-demand assistance in the form of hints and explanations. His secondary research includes reviewing, surveying, and compiling information related to open science and reproducibility across papers published at education technology and learning science conferences.

Stacy T. Shaw[4] is an Assistant Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute. She is an ambassador for the Center for Open Science, a catalyst for the Berkeley Initiative in Transparency in Social Sciences, and serves on the EdArXiv Preprint steering committee. Her research focuses on mathematics education, student experiences, creativity, and rest.

Neil T. Heffernan[5] is the William Smith Dean's Professor of Computer Science and Director of the Learning Sciences & Technology Program at Worcester Polytechnic Institute.

[1] https://osf.io/m7cnr/

[2] https://doi.org/10.17605/osf.io/kyxba

[3] https://ahaim.ashwork.net/

[4] http://stacytshaw.com/

[5] https://www.neilheffernan.net/

He is the founder of ASSISTments, an online learning platform which provides immediate feedback for students along with actionable data for teachers. Heffernan has been pushing open science with his graduate students in recent years. He has also started to push the Educational Data Mining committee to broaden their promotion and support of open science.

4. REFERENCES

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

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

[3] B. G. Cook, L. W. Collins, S. C. Cook, and L. Cook. A replication by any other name: A systematic review of replicative intervention studies. *Remedial and Special Education*, 37(4):223–234, 2016.

[4] A. García-Holgado, F. J. García-Peñalvo, C. de la Higuera, A. Teixeira, U.-D. Ehlers, J. Bruton, F. Nascimbeni, N. Padilla Zea, and D. Burgos. Promoting open education through gamification in higher education: The opengame project. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'20, page 399–404, New York, NY, USA, 2021. Association for Computing Machinery.

[5] 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.

[6] 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.

[7] M. C. Makel, K. N. Smith, M. T. McBee, S. J. Peters, and E. M. Miller. A path to greater credibility: Large-scale collaborative education research. *AERA Open*, 5(4):2332858419891963, 2019.

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

[9] B. Nosek. Making the most of the unconference, 2022.

[10] 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.

[11] 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.

[12] S. Shaw and A. Sales. Using the open science framework to promote open science in education research. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 853–853. International Educational Data Mining Society, Jul 2022.

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

[14] 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.