Infrastructure for Replication in Learning Analytics

Christopher Brooks
School of Information, University of Michigan,
brooksch@umich.edu

Ryan Baker
Graduate School of Education, University of Pennsylvania,
rybaker@upenn.edu

Juan Miguel Limjap Andres
Graduate School of Education, University of Pennsylvania
miglimjapandres@gmail.com

Abstract
There has been increasing concern about the lack of replicability in psychology [2] and, while fervent discussion continues within the field and with the broader public as to the severity of the problem and the correct solutions to it [9,18], there is little doubt that there will be an impact on methodological practices [11]. In this position paper, we argue that many of the same concerns are applicable to the learning sciences as psychology, and that new methodologies and infrastructure have the potential to help mitigate these concerns, and help us to understand how broadly we can trust that published findings generalize. In particular, we focus on the opportunities afforded by the large-scale participation in massive open online courses, and argue that the data from MOOCs can enable meaningful large-scale replication and comparison if augmented with an appropriate sociotechnical infrastructure, which we describe in brief.

Replication in Education
Over the last few years, there has been increasing interest in which findings are broadly generalizable. One approach is to design very large-scale studies that incorporate diverse groups of students [8]. By selecting an appropriate group of students, it may be possible to better understand which students a finding will generalize to [13]. However, such studies are both highly expensive and may not be robust to minor variations in design -- since the same treatment, implementational approach, and measures are typically used at all sites, their findings are narrow even as they are broad, reducing the degree to which we can conclude that findings generalize.

One other risk to generalizability is when experimenter bias (whether intentional or not) influences how results are analyzed and interpreted. While the practice of removing outliers until reaching significance or running 30 tests and reporting one significant result (with no post-hoc control) are increasingly seen as problematic, the "file drawer effect" [10] where non-significant results are suppressed (often by journal editors rather than researchers) and the natural human tendency to focus on surprising and serendipitous results can lead to spurious results becoming published and widely cited. While the complete elimination of serendipity is undesirable (after all, without serendipity we would not have penicillin or use salicylic acid cosmetically), bias can be reduced in several ways. One approach for this is to encourage pre-registration of scientific hypotheses before a study is carried out. Infrastructure such as the Open Science Framework [19] aims to support this activity. This can be done in either a descriptive manner, where the researcher outlines study, methods, and hypotheses then conducts the study, or in a prescriptive manner, where the research must first submit the registration information to a body of peers before the study can be conducted. This latter form intends to have two effects: an increase in the number of experts involved in the replication in the design phase, and (important for the investigator) reviewed pre-registration often comes with a guarantee of publication regardless of whether the outcome discovers a positive, negative, or null result.

Another path which has encouraged replicability is the greater sharing of data. In education, the Pittsburgh Science of Learning Center DataShop [6], an online repository of student interaction data with educational software, has led to literally dozens of published papers by secondary researchers. This infrastructure, now being iterated upon with the LearnSphere project [14], provides a layer of access and control services to educational datasets, allowing researchers to share de-identified data and use common tools in their analyses.

A Proposal for Discussion

We argue that existing infrastructure is necessary but not sufficient to support replication in the learning sciences. In particular, some analyses, such as discourse analysis, may require access to sensitive information which may be difficult or impossible to fully de-identify. While limited data of this type has been shared broadly, it remains infeasible to share large amounts of this type of data widely. In addition, there remains considerable data which institutions consider proprietary and which therefore cannot be simply downloaded by outside researchers. While exemplary organizations such as Carnegie Learning and ASSISTments [5] share data widely through engines such as the PSLC DataShop, they cannot share all data (such as student identifiers), and many other organizations are unable to share data even to the level which these organizations do.

One alternate approach is to create a **cooperative dataset infrastructure**, where data is not available for download, but can be used in new analyses. Large-scale proprietary data, such as MOOC data or educational software data, could be made available on this platform. Unlike data from traditional randomized controlled trials, this data could be drawn from the regular use of online platforms such as Coursera and edX, as well as data from publishers that is not currently available on DataShop. Given the massive quantities of data held currently by institutions such as the University of Pennsylvania, and the University of Michigan, along with other institutions such as the Universities of Edinburgh, Harvard, and Stanford, there are great potentials for conducting broad and replicable research, where we determine what findings hold in what courses and for which students. But we must unlock this data in a fashion that allows researchers to analyze it while protecting the goals of the data owners as well.

As such, in an infrastructure like this one, the data would be retained solely by the institution (or perhaps by a trusted broker). Rather than downloading data, any researcher conducting an analysis would offer code or analyses (in a syntax such as R, or a production-system language). This code would then be applied to the data and the results would be made available both to the researcher and more broadly. This code would be required to be made available to the wider scientific community (perhaps under a delay or publication embargo, so that researchers were able to publish their findings first), so that any result obtained could be re-checked, verified, criticized, modified as appropriate, and re-used where applicable. Again, this platform would not be in competition with LearnSphere (and in fact is likely to be a natural outgrowth of a platform like LearnSphere), but would supplement it for organizations that cannot share their data in the same fashion as the contributors to the DataShop do.

A more controversial approach might be the addition of a **sociotechnical infrastructure** (Edwards 2003) which augments the cooperative infrastructure with a social/legal contract binding those who participate. Participation might entitle researchers access to underlying datasets, metadata, methods, code, and results, but requires that they participate equally by providing more datasets, metadata, methods, code, and results. Unlike the cooperative dataset infrastructure, this sociotechnical infrastructure would be a cooperative, where datasets, metadata, methods, and results are actively iterated upon, augmented, and changed. Like many open source software projects, access to this cooperative imparts a duty on the accessor.

Such a sociotechnical infrastructure would be most successful if it accommodates the different interests and techniques of different intellectual traditions in the broad and intellectually diverse learning sciences community. For instance, we might expect that learning theorists might be interested in the relationship between discussion forum content and educational frameworks, such as [7,16], while linguists might be interested in the relationship

between discussion forum content and student success [4]. Thus for the same set of data, one research project could provide human annotations through open or closed coding, while another could provide feature extraction based on natural language semantics. Not only can the two investigations then look at specific questions within the same dataset potentially providing insight into cross disciplinary findings, but future investigations by others can use the same constructs (human annotations or features) either for replication or comparison.

Existing Examples

One well-known candidate for evolving into such an infrastructure is the PSLC DataShop/LearnSphere. However, it may also be useful to advance other platforms which are more specifically focused on large-scale replication of research findings.

Some elements of the **cooperative dataset infrastructure** may be found in MORF, the MOoc Replication Framework at the University of Pennsylvania, which is designed to replicate research results across MOOCs. In MORF, MOOC data can be ingested into a common research format. This format is similar to MoocDB [15], but tailored to the specific research goals of the platform. Researchers can then input a finding to check for, realized as an "if-then" production rule. The production rule is then tested against the MOOC data sets currently ingested within the framework. The wide number of published findings from MOOC data (for instance, [3,4,12,17] provide considerable grist for such a project. Despite being broadly interested in the same topic, each of these experiments is impossible to compare to others in that they differ in several ways:

(a) Feature engineering approaches used,
(b) Outcome of prediction,
(c) Dataset characteristics,
(d) Machine learning techniques used

By contrast, in an engine such as MORF, each variant on the findings can be readily tested and compared within the same data sets. In a first "proof of concept" pilot study, 21 previously published findings from the MOOC literature were tested within the context of a single MOOC's data set [1]. While this falls far short of the eventual goal of testing hundreds of findings in hundreds of MOOCs, it demonstrates that research of considerably broader scope than most MOOC research can be feasibly carried out in such a platform.

Existing elements of a **sociotechnical infrastructure** are harder to find. The closest known by the authors in the MOOC space is the EdX Research Data Exchange (RDX)[1]. The RDX is a voluntary exchange of MOOC data which is open to charter members of the EdX organization, and as of writing 19 members have elected to join the RDX. The RDX data is de-identified, but contains artifacts which are highly personalized (e.g. discussion fora messages), and membership to the RDX gives access to all of the combined data of institutions involved in the RDX. This infrastructure is governed by a separate legal agreement among members.

The RDX demonstrates an existing issue when sharing learning systems data -- the advent of cloud computing has brought with it increase ambiguity as to ownership and stewardship of student data. The RDX is in part possible because a single vendor is involved, enabling the activity. But it's less clear how feasible cross vendor data-sharing solutions are in the highly competitive and deeply integrated MOOC space.

Considerable work remains to create a socio-technical infrastructure that will help education to get past the problem of replicability that has plagued social psychology. By leveraging all of the data that is already available, and creating an infrastructure that supports research, we may be able to advance research, understanding where findings apply, and for whom.

[1] See http://edx.readthedocs.io/projects/devdata/en/latest/rdx/index.html

Conclusion

The intent of this position paper is to surface the issue of replication in learning analytics and the learning sciences more broadly, and to propose for discussion the need for cooperative dataset and sociotechnical infrastructures. We have focused here on the MOOC domain, which we we perceive to be of high interest to a broad community, and where the vendor presence (and in some cases, activity) minimizes the need for significant integration efforts. Our interest in bringing this forward to the LAK 2017 Workshop on Methodology in Learning Analytics (MLA) is to gain insight from workshop participants, as well as potentially form new partnerships, as we bring these ideas forward to granting agencies.

References

1. Juan Miguel L. Andres, Ryan S. Baker, George Siemens, Dragan GAŠEVIĆ, and Catherine A. Spann. Replicating 21 Findings on Student Success in Online Learning. Retrieved from http://www.columbia.edu/~rsb2162/TICLReplicationManuscript.pdf
2. Monya Baker. First results from psychology's largest reproducibility test. *Nature News*. https://doi.org/10.1038/nature.2015.17433
3. Christopher Brooks, Craig Thompson, and Stephanie Teasley. 2015. A Time Series Interaction Analysis Method for Building Predictive Models of Learners Using Log Data. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (LAK '15), 126–135.
4. Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. Combining Click-stream Data with NLP Tools to Better Understand MOOC Completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (LAK '16), 6–14.
5. Heffernan, Neil T and Heffernan, Cristina Lindquist. 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.
6. Kenneth R. Koedinger, Ryan S. J. D. Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. A Data Repository for the EDM Community. In *Handbook of Educational Data Mining*. 43–55.
7. Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: a cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 15–24.
8. Frederick Mosteller and Robert F. Boruch. 2002. *Evidence Matters: Randomized Trials in Education Research*. Brookings Institution Press.
9. Katie M. Palmer and Katie M. Palmer. 2016. Psychology Is in Crisis Over Whether It's in Crisis. *Wired*. Retrieved December 11, 2016 from https://www.wired.com/2016/03/psychology-crisis-whether-crisis/
10. Robert Rosenthal. 1979. The file drawer problem and tolerance for null results. *Psychological bulletin* 86, 3: 638.
11. Barbara A. Spellman. 2015. A Short (Personal) Future History of Revolution 2.0. *Perspectives on psychological science: a journal of the Association for Psychological Science* 10, 6: 886–899.
12. Colin Taylor, Kalyan Veeramachaneni, and Una-May O'Reilly. 2014. Likely to stop? Predicting Stopout in Massive Open Online Courses. *arXiv [cs.CY]*. Retrieved from http://arxiv.org/abs/1408.3382
13. Elizabeth Tipton. 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of educational and behavioral statistics: a quarterly publication sponsored by the American Educational Research Association and the American Statistical Association* 39, 6: 478–501.
14. Carnegie Mellon University. Carnegie Mellon Leads New NSF Project Mining Educational Data To Improve Learning-CMU News - Carnegie Mellon University. Retrieved December 11, 2016 from https://www.cmu.edu/news/stories/archives/2014/october/october2_learnsphere.html
15. Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary Pardos, and Una-May O'Reilly. 2013. Moocdb: Developing data standards for mooc data science. In *AIED 2013 Workshops Proceedings Volume*, 17.
16. Alyssa Friend Wise, Yi Cui, and Jovita Vytasek. 2016. Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 188–197.

17. Wanli Xing, Xin Chen, Jared Stein, and Michael Marcinkowski. 2016. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in human behavior* 58: 119–129.
18. Ed Yong. 2016. Psychology's Replication Crisis Can't Be Wished Away. *The Atlantic*. Retrieved December 11, 2016 from http://www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/
19. Home. Retrieved December 11, 2016 from https://osf.io/