Workshop on Big Data in Education

Balancing the Benefits of Educational Research and Student Privacy

Learning Process Data in Education Research: Panel Summary

Constance Steinkuehler, *University of Wisconsin-Madison*

NATIONAL ACADEMY of EDUCATION

Learning Process Data in Education Research
Panel Summary

Constance Steinkuehler, University of Wisconsin–Madison

Workshop on Big Data in Education:
Balancing the Benefits of Educational Research and Student Privacy

National Academy of Education
Washington, DC

NATIONAL ACADEMY OF EDUCATION 500 Fifth Street, NW Washington, DC 20001

Additional copies of this publication are available from the National Academy of Education, 500 Fifth Street, NW, Washington, DC, 20001; https://www.naeducation.org/bigdata.

Printed in the United States of America

Suggested citation: Steinkuehler, C. (2017). *Learning Process Data in Education Research: Panel Summary*. Washington, DC: National Academy of Education.

Learning Process Data in Education Research: Panel Summary

Constance Steinkuehler, University of Wisconsin–Madison

Panel:
Constance Steinkuehler, University of Wisconsin–Madison (Panel Chair)
David Pearson, University of California, Berkeley (Panel Moderator)
Ken Koedinger, Carnegie Mellon University
Zach Pardos, University of California, Berkeley
David Yaskin, Hobsons

INTRODUCTION

"Learning process data" are "a continuous or near-continuous record of usually digital interactions supporting finer-grained inferences about ongoing student progress. These data have proliferated in recent years with the evolution and expansion of digital learning systems and dramatic improvements in data storage capacity, data computation speeds, and data analytic methods."[1] Learning process data are "big" not only by virtue of being "tall"—having many participants over a given period of time—but also by virtue of being wide, fine, and deep. By "wide," we mean data that include a large number of variables of interest about any one individual. By "fine," we mean data that include multiple fine-grained observations taken across small intervals (e.g., every 10 seconds). By "deep," we mean data that are theoretically coded or labeled in some meaningful way. All four dimensions are important to the contribution of learning process data to research.

Learning process data are most commonly used to improve our theories and principles about student learning and engagement and then, based on these theories and principles, design better educational interventions. In education research, we assume that student activity during the course of a curricular intervention directly impacts what he or she learns. As noted by Koedinger and Pardos, analyses of student process data can predict, at least partially, a given student's educational outcomes. Analyses of learning process data can also determine which educational resources used within a given activity are most effective.[2] In so doing, we can assess not only what is effective and why, but also for whom. In this way, learning process data play a fundamental role in personalized learning, or the dynamic adaptation of curricular materials to the real-time needs of student users.

Learning process data give us the capacity to scale formative assessment so that it is both feasible and affordable. Such assessments can focus on the student, the program, or both. Representations of such data can serve as crucial tools for teachers, school administrators, parents, and students themselves in understanding and responding to a student's progression through course material and activities while that activity is in progress rather than after the fact. Such analyses can help identify individual students' areas of misconception or understanding of the content at hand so that struggling students can be given appropriate content and scaffolding as needed, during the instructional activities themselves. The goal is not just better assessment of student process but also better instruction.

Such capacity, however, requires two things. First, it requires thoughtful linkage between process data and administrative data including, for example, individual sociodemographic variables, performance scores, and other success metrics. To connect process to outcome, we have to tie curricular actions and materials to specific educational outcomes at the individual level. Second, it requires a knowledge base of prior analytic work built on knowledge sharing not merely in terms of results or findings but more crucially in terms of detailed data sets and specific analytic methods. For a digital learning management system, for example, to be responsive to student understanding, patterns and trends in student activity that predict varying outcomes have to be identified in the literature and those patterns must be supported empiri-

[1] Ho, A. (2017). *Advancing Educational Research and Student Privacy in the "Big Data" Era*. Washington, DC: National Academy of Education.

[2] Pardos, Z., Dailey, M., & Heffernan, N. (2011). Learning What Works in ITS from Non-traditional Randomized Controlled Trial Data. *The International Journal of Artificial Intelligence in Education* 21:47-63; Rau, M.A., & Pardos, Z.A. (2012). Interleaved Practice with Multiple Representations: Analyses with Knowledge Tracing Based Techniques. Pp. 168-171 in *Proceedings of the 5th Annual International Conference on Educational Data Mining*, Crete, Greece.

cally as to which materials when are the appropriate intervention. Building this knowledge base requires collaboration and coordination, including shared data sets and data methods.

In many respects, the privacy risks associated with process data are no different than those associated with other forms of data. In research that includes sensitive personal information, participants must give informed consent—and, in the case of minors, informed assent as well. Process data collected must remain either *anonymous* (collected without personal identifiers) or *deidentified* after collection (with personal identifiers obscured in some way). The primary privacy breach of concern, therefore, is intentional and accidental "reidentification." Reidentification risk arises from (a) malicious reconstruction or accidental release of previously obscured personal identities or (b) when linkage among multiple data sets leads to a complete enough stream of unique data points on a given participant to allow reconstruction of a given participant's unique identity.

CONNECTING LEARNING PROCESS DATA TO ADMINISTRATIVE DATA

Learning process data are not commonly used themselves as the basis for grading or advancement in classrooms, but they are often connected to grades and other forms of administrative data (e.g., test scores and sociodemographic data) in order to differentiate which actions lead to what outcomes and to determine whether and how aspects of the intervention works for some subgroups of students and not others. Administrative data are crucial for tying learning process to its outputs and inputs, both the assessed gains in knowledge, practice, and disposition that result from a given piece of instruction as well as the demographic variables (most notably, prior knowledge) that shape and inform what is learned.

There are obvious risks involved, however, when grades or demographics are added to learning data process sets: the risk of reidentification increases, as do the stakes. Much of the research to date involving learning process data has been conducted among adult learners in voluntary contexts like massive open online courses (MOOCs). In nonvoluntary contexts such as K-12 public schools, and when minors are involved, additional precaution is needed to protect student privacy. Two things are needed. First, those handling data within schools—school administrators and not just researchers—need training on how to handle such data. Even though current staff in public K-12 schools already handle sensitive data, there is a pervasive lack of training on all levels. Second, an infrastructure and agreed set of procedures for linking process and administrative data in K-12 settings could track and mitigate such issues, yet no such system currently exists. At present, there is no shared data architecture or agreed-upon protocol to centralize or standardize such data and therefore reduce and detect risks. Until such sociotechnical systems are developed, we have no way to track threats to privacy let alone attenuate them as we should. The development of such systems would also allow educational researchers to get out of their lab and work more directly and meaningfully in schools.

SHARING DATA AND ANALYTIC METHODS TO ACCELERATE RESEARCH

Using learning process data to improve theory and practice in education requires a knowledge base that goes beyond shared results; it requires sharing the basis of those results, data corpora and methods for data curation, analysis, visualization, and reporting. Process data analysis is an emerging domain that borrows techniques from adjacent fields such as Bayesian

knowledge tracing,[3] a mainstay of intelligent tutoring systems for estimating cognitive mastery, and neural networking approaches, yet much of our analytic tools and procedures are currently under development. Researchers need to be able to share data across their academic and institutional silos in order to conduct basic research at scale and, in a word, get beyond mere intuition.

To date, however, there is little formal capacity for such work at scale. The largest current effort toward these ends is the DataShop,[4] created by the LearnLab at the Pittsburgh Science of Learning Center. DataShop is the world's largest repository of learning interaction data,[5] containing more than 850 data sets. For example, through such learning process data sets, researchers can look at trends in performance over time (e.g. error rates) and then, using this trend curve, improve the intervention to reduce overall time to mastery. LearnSphere[6] is another noteworthy effort, which attempts to integrate methods with data to facilitate collaboration and knowledge sharing across institutions.

Here, there is a tradeoff between sharing and privacy protection. As the risk of deidentification increases, access must necessarily decrease. For example, data sets that contain no administrative data and no links to other data sets pose little to no risk of reidentification; thus, they can often be shared publicly under Family Educational Rights and Privacy Act (FERPA) standards. But as such features of the data increase, access must be increasingly regulated through institutional review board (IRB) oversight at respective research institutions and through differing levels of access. Such governing bodies create a chain of responsibility should data be mishandled in some way. Access rights shift with reidentification risk, creating a sliding scale of differentiated access depending on risk, need, and reward. Although we currently have no standardized data-sharing infrastructure or protocols to guide this process, DataShop and similar efforts provide clear proof of concept. The end points—public access on no-risk data, on the one hand, and strictly controlled access to risky data, on the other—are easily determined; the middle ground requires better IRB processes and shared policy. Such mechanisms for data sharing would enable much more rapid progress in the domain.

CONCLUSION

Large corpora of learning process data allow us to develop empirically based principles for learning that, in turn, guide the design and development of educational materials and activities as well as teaching practice. Although the collection of such data do raise issues of privacy, appropriate data architecture and protocols can be developed to centralize or standardize the collection of such data, their linkage to administrative data, and their collaborative sharing across institutions that can detect and mitigate privacy threats while improving our capacity to conduct empirically driven and classroom-focused research.

There are broader concerns with assessment regimes in our schools, however, particularly that these regimes are either not equitable or not increasing equity with regard to race/ethnicity and socioeconomic status. Learning process data may raise unique privacy concerns or amplify

[3] Corbett, A.T., & Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4(4):253-278.

[4] See http://learnlab.org/datashop.

[5] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM Community: The PSLC DataShop. In *Handbook of Educational Data Mining*, edited by C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker. Boca Raton, FL: CRC Press.

[6] See http://learnsphere.org.

existing concerns. When there are concerns that systems are not equitable or increasing equity, collecting data that are summative already raises privacy concerns, but when looking at process data, which are detailed data collected on what students are doing and their behaviors, heightened equity and possibly privacy concerns may exist.

Nevertheless, there is also a significant opportunity cost to deterring the collection and use of process data that is borne not only by educational researchers but also and perhaps more consequentially by schools and students themselves. To not use process data is to hamstring our capacity to understand student process, to improve our instruction based on what demonstrably works, and to tailor materials and processes to individual students in an empirically grounded way. Thus, it is worth the investments required to develop sociotechnical infrastructure to manage and mitigate risks that may be involved.