## Workshop on Big Data in Education

*Balancing the Benefits of Educational Research and Student Privacy*

# Implications of Privacy Concerns for Using Student Data for Research: Panel Summary

Marie Bienkowski, *SRI International*

NATIONAL
ACADEMY
of
EDUCATION

# Implications of Privacy Concerns for Using Student Data for Research
# Panel Summary

Marie Bienkowski, SRI International

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: Bienkowski, M. (2017). *Implications of Privacy Concerns for Using Student Data for Research: Panel Summary*. Washington, DC: National Academy of Education.

# Implications of Privacy Concerns for Using Student Data for Research: Panel Summary

**Marie Bienkowski, SRI International**

**Panel:**
Marie Bienkowski, SRI International (Panel Chair)
Susan Fuhrman, Chair, Teachers College (Panel Moderator)
Monica Bulger, Data & Society Research Institute
Adam Gamoran, William T. Grant Foundation
Kirsten Martin, George Washington University
Piotr Mitros, edX

# INTRODUCTION

An incredible amount of data is collected in the K-through-grey U.S. educational system and as part of educational research—including but not limited to administrative data, assessments, keystrokes in apps and online learning management systems, survey data, and audio/video recordings. Appropriate use of these data requires balancing multiple perspectives: the right to privacy for individual students, teachers, and parents; the needs of researchers who can make impactful discoveries from data-intensive research; the interests of commercial organizations who are building the digital tools that collect the data; and the priorities of state and local actors. Insights from analysis of large data sets come both at the macro level, for example, policy guidelines obtained from administrative data, and from the micro level, for example, tracking the causal mechanisms for learning, perhaps in the moment-by-moment enactment of pedagogy and curriculum inside classrooms and digital environments.

The concluding panel of the National Academy of Education (NAEd) workshop on student privacy in education research made an effort to summarize the complex perspectives and issues presented, and to provide overall recommendations to the NAEd for next steps. Panel members' comments reflected their backgrounds in education research, software design, working with data in large online courses, business privacy responsibilities and ethics, and public policy around data use. Audience members participated enthusiastically during the Q&A session and their comments raised new issues, promoted new resources and approaches, and reinforced the importance of issues raised in previous panels.

While many concerns and issues were raised and are discussed below, the panel emphasized that, overall, big data in education—both administrative and learning process data—have great power to improve teaching and learning through data-intensive education research, through data use by school stakeholders, and by product improvement on the part of technology companies. Parents and students, connected to education research often only through indecipherable consent and assent forms, may not reap the immediate benefits of research and development but can participate in research to support future generations' improved learning. Researchers care about the data and the findings, and see a clear need for using data that are identifiable (for example, to link data across databases); nonresearchers are interested in whether data are personally identifiable, who is the caretaker of those data (especially if data are personally identifiable), and whether, at any point in the collection or use, the process is going to negatively impact children. Researchers acknowledge this strong desire for maintaining privacy, either through deidentification or with proper safeguards on data stores. Researchers care about the promise of big data: insights that can be gleaned from big data that can then be verified experimentally. Parents and students care whether the data are permanently stored in a student's education record (in which case the student may be "labeled"), whether some populations represented in the data are disproportionately vulnerable, and what harms disclosure could bring.

Throughout the meeting, different perspectives and contexts framed remarks. In K-12 the need to preserve children's rights, conduct research of benefit to them, deal with skittish parents and school administrators, and comply with federal regulations such as the Family Educational Rights and Privacy Act (FERPA) was evident. At the same time, postsecondary representatives, collecting rich data sets on large numbers of students, ostensibly deidentified, suggested that absolute protections of privacy are difficult to guarantee in the age of hackers bent on earning prestige. As more rich audio and video data are collected on group and individual work for assessment of social-emotional and soft skills, deidentifying data is untenable.

Consumers of web-based resources, however, do not expect to release zero information; for example, in the case of free resources, the public is accustomed to models such as free email services paid for by advertising based on data *inside* the emails.[1] In the case of open online courses, for example, massive open online courses (MOOCs), use of data for research is granted in exchange for the free content. And MOOCs collect a lot of data: as Mitros pointed out about edX, “We have four years plus of [fine-grained] data on students, many of whom have completed the equivalent of undergraduate education. . . . That is across 10 million students . . . [and] a thousand courses.”

Another perspective that emerged is that of the “end user” who uses a digital learning system and thereby produces data (even if that production is done unwittingly or without being obvious). Many panelists felt that the users, especially K-12 students, are forgotten in discussions of student data. In an age where data and information literacy are increasingly necessary skills, big data collected by researchers could be shown to be of value and benefit to the *subjects* of the research; instead of thinking of students as data points, perhaps when engaging in primary collection, researchers could use this as an opportunity, where appropriate, to talk to students and teachers about data, their data, their lives converted to data points, and use these conversations as teachable moments to inspire an interest in research and data literacy.

Yet another perspective, perhaps less well represented, is that of the technology companies and vendors who are operating under a specific business model (including paying employees) that make some use of the data collected. How does a company use data if a school is getting software free of charge? How does a company balance using data to improve their product versus building knowledge to improve the field?

## BACKGROUND AND DEFINITIONS

Several panelists and audience members reminded the participants of important definitions; it takes work to remember the correct terms and accompanying connotations. For example, regarding privacy versus confidentiality, the *Institutional Review Board Guidebook*, published by the Office for Human Research Protections[2] states:

> Privacy can be defined in terms of having control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others.
>
> Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others, in ways that are inconsistent with the understanding of the original disclosure without permission.

As Martin pointed out, Helen Nissenbaum at New York University views privacy as the appropriate flow of “sharing” within the context.[3] Martin defines privacy as a “social contract” with “negotiated norms of what information should be used and how within a community.” She explained that under this theory, privacy looks different in different contexts and,

[1] As Google’s Terms of Service points out: “Our automated systems analyze your content (including emails) to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection. This analysis occurs as the content is sent, received, and when it is stored.”

[2] Office for Human Research Protections (OHRP). (1993). *Institutional Review Board Guidebook*. Available at https://archive.hhs.gov/ohrp/irb/irb_preface.htm.

[3] Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press.

importantly, privacy is not seen in tension with research but instead part of the social contract between the stakeholders.

These frames can be used to describe expectations for research practices that include stakeholders in the "contract" especially because their behaviors help shape the norms. In research with human subjects, researchers provide subjects with descriptions of their confidentiality procedures that help subjects decide what control (i.e., privacy) they are willing to give up. When subjects agree to participate in research, they expect that the confidentiality agreements that researchers propose will be upheld. When confidential information is divulged, privacy is violated and information that a subject believed would be kept private is not. In the era of big data kept in networked devices, confidentiality procedures can be more difficult to uphold due to more elaborate requirements for computer information security.

Information privacy is the appropriate protection, use, and dissemination of information about individuals. Information security, essential to information privacy, means preventing unauthorized access to data and includes standards that can be followed to maintain proper access to data. Security breach is a legally defined term that requires notice to the individuals and remediation. In contrast, a data incident is not necessarily a breach; however, it may be more harmful and more impactful to research work because of the perception. A leakage of data, an accidental sharing of data, or the misuse of data can have the same impact on an organization as an actual breach in terms of public perception and the press coverage.

Another framework for thinking about more than the legal limits on data use involves three levels. There is the risk of actually breaking the law, whether that harms a person or not. The second level is actually violating privacy through security incidents but not breaking the law. The third level is perception: incidents that are not against the law, that most would not consider a privacy or security violation, but that show up in the press and cause concern. Understanding the social contract around privacy according to these levels can help researchers interpret possible or actual incidents, and the consequences.

## COMMUNICATING THE VALUE OF RESEARCH

There were many calls for better presentation of the value of data-intensive education research. Evidence of how big data are already being used to improve learning should be highlighted, as well as examples of how other industries use such data. The consensus was that researchers have done a poor job explaining the value of collecting and using student data for educational research. Research has informed landmark policies such as early education, school breakfasts, and teachers paired with veteran mentors. As Bulger pointed out, none of these important policies, which were informed by research, is promoted as "this is brought to you by research from XYZ [organization] using student data from over 100,000 kids in 45 school districts over a five-year period." Researchers do need to make such research contributions more visible, and, as a community, researchers also need to understand that a lack of transparency breeds mistrust and misinformation.

As an example of communication, a representative of the Consortium for School Networking (CoSN) described how it has created user-friendly tools on data and privacy intended for school districts. Their observation is that school districts are not effective at communicating with parents or guardians why they collect data, generally relying on acceptable use policies (essentially the same types of agreements required when someone downloads software). A long set of pages in small text that a parent has to sign does not instill trust. By partnering with schools, CoSN created a simpler, multilingual, and more visual infographic that is intended to be used by the principal to communicate data procedures to the parents, perhaps at an informational event such as back-to-school night. Other privacy tips are described at cosn.org/privacy.

Another example is what the education community (e.g., school systems) has done at studentdataprinciples.org with the Data Quality Campaign. Forty national associations were involved in voting on belief statements around data and privacy, and the participants learned that there is consensus at the high level of these statements. The resulting ten parent- or legislator-friendly statements about what these organizations believe about why school systems collect data and how we will protect data are good examples for the education research community.

The challenge to the research community is to become as clear and precise as the education community is working to be about why data-intensive research is important to our society. Why is it important to you as a parent? What does it mean for your child? Beliefs and experiences shared by the audience revealed that superintendents and principals want to be part of a bigger purpose as long as it is well articulated and of value to teaching and learning.

As Bulger explained, challenges to the use of student data have conflated national testing with *all* educational research, any student data collection with third-party sharing for advertising purposes, and big data with existing practices of student tracking and infinite remediation loops. Trust is important to establish, and transparency helps build trust through answers to basic questions such as, What are you going to do with my data? What measures will you take to protect my data?

## FORMING MEANINGFUL AND EFFECTIVE PARTNERSHIPS, INCLUDING WITH TECHNOLOGY COMPANIES

Partnerships with school districts, with states, and with business and industry all enhance big data research. But partnerships are not easy to accomplish and may require that researchers set aside their own research agenda in favor of the stakeholder's goals. These partnerships, between technology experts, researchers, policy makers, and advocates including parents and educators, are essential because they establish a framework for trust. They set out standards and norms of practice that allow school districts to trust that researchers will address questions whose answers will be consequential to decision making and will not use the data for purposes other than those for which they have been granted.

If schools and districts routinely use research results in their own decision making, establishing partnerships can be easier. The Institute of Education Sciences (IES) has funded two research and development centers that are looking at knowledge utilization—including research-backed knowledge—in schools by superintendents, principals, and teachers. Early findings suggest that superintendents and principals use research on a daily basis, although use at the teacher level lags.[4]

Although there was limited representation at the workshop from technology companies who make digital learning tools (including online learning environments) and from companies who provide data warehousing services for schools (K-12 or postsecondary), they did feature prominently in discussions of partnering and data sharing. First, researchers must be informed that companies handling student data must abide by FERPA. The U.S. Department of Educa-

[4] Means, B., Padilla, C., & Gallagher, L. (2010). *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*. Washington, DC: U.S. Department of Education. Available at http://eric.ed.gov/?id=ED511656.

tion's Privacy Technical Assistance Center[5] makes clear that providers of online educational services who work with districts and with personally identifiable information are covered under the "school official exception" educational use portion of FERPA.[6] Second, researchers, state educational agencies, and school districts must remember that companies have a "business model" that governs a price they put on the value they can obtain from use of data. On the other hand, learning technology companies themselves have a rationale for using data to improve their products. The question becomes whether the present value a school obtains is worth the "cost" of the data that is given to the vendor. It is also the case that efforts at schools co-developing technology or investing in startups can backfire as journalists and others accuse them of experimenting on our children ("used as whetstones to hone a badly flawed product being pitched as cutting-edge technology…"[7]).

Panelists discussed the need for researchers to work with technology developers to collect data useful to research. At present, researchers must adjust their research questions to the data that technology developers of learning platforms (MOOCs or personalized learning systems) deem relevant to collect. Increasing partnerships with such companies will increase the value of the data collected. There are examples where education researchers are working with technology vendors to use data for improvement,[8] and companies in the learning space are interested in collecting the right data for improvement.

## BETTER TRAINING FOR EDUCATION RESEARCHERS

There was also a call for better preparation for researchers in two topic areas. The first is more widespread adoption, among social, behavioral, and education researchers, of the standard, compliance-oriented human-subjects training. Existing human-subjects regulation training is accessible online through organizations such as the Collaborative Institutional Training Initiative, is used by federally mandated institutional review boards (IRBs) for compliance purposes, and covers the history of human-subjects research protection, researcher's responsibilities, and ethics. However, it does not deal with nonregulatory issues (e.g., state laws for data use) nor does it cover technical issues needed to keep data secure. It also does not reflect the current and all-too-real concerns about student privacy held by parents, students, and advocates. Therefore, in addition to recommending broad adoption of standard human-subjects training in the education research community, the panel recommended training or workshops based on collected best practices around data security and privacy, and the preparation or curation of materials that reflect the concerns about student privacy held by parents, students, and advocates.

[5] See ptac.ed.gov.

[6] See https://tech.ed.gov/wp-content/uploads/2014/09/Student-Privacy-and-Online-Educational-Services-February-2014.pdf.

[7] See http://www.aclumich.org/article/guyette-how-eaas-buzz-program-exploited-detroits-most-vulnerable-kids.

[8] See https://ww2.kqed.org/mindshift/2014/05/06/how-are-teachers-and-students-using-khan-academy.

## WAYS TO SAFEGUARD BIG DATA AND MAKE IT AVAILABLE TO INFORM POLICY

In some big data environments, there are thousands of researchers handling educational data under thousands of different security policies. For example, Mitros noted that edX data is handled by hundreds of researchers, which makes for a very large "security perimeter." While some felt that the standards for education research security should not have to be at the heightened level of security for the financial industry, others were adamant about the serious privacy concerns with student data and thus the need for significant security protections. While some posited the different nature of financial data and Individual Education Plans (IEPs)—one leading to possible financial gains if garnered and the other without such gains—others argued that IEPs are highly confidential and that hackers choose to access these data not for financial gain but for the sport of it. Additionally, Gamoran pointed out that much of the data used by education researchers is already being collected as education data; thus, the researchers are not creating new incentives for hacking because the data are already collected.

It was agreed that there should be different controls and access for different types of data, including data centers and controlled access. When the controls are stricter and more trusted, it is possible to avoid the data loss that comes with deidentification. Any rule that requires aggregation to a certain block level for any type of research means that important populations could get overlooked. One of the problems with aggregation is that minority groups and small, special, vulnerable populations get lost. Eric Hanushek, a workshop participant, shared an example of a state-sponsored research laboratory where all of the microdata that the state collects on any subject, teachers, students, etc. are stored and a standard process was created for applying for use of parts of these data by individual researchers.

Gamoran suggested that the volumes of data that the government already collects should be made better available for and used for research purposes. He stated his hope that the bipartisan congressionally mandated Commission on Evidence-based Policy[9] interprets its mandate broadly to not only link federal data sets but also to prepare the data for use by researchers, create the structures to support such research from inside and outside the government, and encourage the use of evidence for smarter policy decisions. Participants agreed that the goal is not to say that evidence should drive the policy, but rather that evidence should inform the policy. Evidence should be at the table when policies are discussed.

Panelists sounded alarms about potential data breaches especially as the types of data collected become more difficult to deidentify (e.g., audio and video data) and as individual opt-out becomes more difficult (e.g., due to an individual's presence in teams and collaborative work). Educational researchers have been engaging in the collection, storage, safeguarding, and use of student data since before big data, but the data are more available to hackers, the climate is more sensitive, and deidentification is more difficult.

Perceptions versus harms and risks were discussed with the issue of awareness that needs to be raised everywhere. Education researchers are not the only voice at the table or partners emphasizing student privacy. Moreover, parents and students do not necessarily differentiate education data from research data.

[9] See https://www.obamawhitehouse.archives.gov/omb/management/commission_evidence. "The 15-member Commission is charged with examining all aspects of how to increase the availability and use of government data to build evidence and inform program design, while protecting privacy and confidentiality of those data."

## NEXT STEPS AND RECOMMENDATIONS

Participants in the workshop felt that there were several steps forward.

First, researchers should *improve the ways they communicate* the value of research, find ways to clearly communicate the confidentiality procedures put in place for data, and help shape community norms around the privacy contract between researchers and research participants:

- Clearly explain why we collect data and how we protect those data. Researchers should describe, for those involved in research, what data they are collecting, what their research is, who funds it, what its likely benefit is, and what other research it resembles or builds on. The consequences of "opting out" in terms of evidence and impact should be made clear to participants, as should the need to participate in research that may not offer immediate benefits to participants. Communicating requires nonacademic, simple language. We need to explain that we collect data to determine how every child can succeed. In addition to explaining what we will do with the data and why we will do it, we need to explain what we will never do with the data, such as sell it. We also must explain how we will ensure the protection of the data.
- Build training for researchers to communicate the value of research, and also to communicate the nuanced differences between educational research data, data collected as part of normal education practice, and state- and federal-held data sets. Using terms familiar to parents can help differentiate who does what kind of testing: national, state, and district standardized tests; school- and teacher-administered interventions and tests; and those administered as part of a research study. And within all data to be collected, making clear the differences between administrative versus process data (for parents and subjects) will help them understand who collects and who controls the data. For example, a researcher may be given eighth grade standardized mathematics test scores (controlled by the district or school), backend process data from an online math learning system (provided by a commercial entity), and responses from a survey (administered by the research organization). These distinctions need to be explicit.
- Help improve IRBs with a clearinghouse on best practices, examples, and a robust professional community of those making decisions in the education sector. Areas for improvement include improving communications with subjects through better assent and consent forms, better understanding of how big data impact education, and when big data research status is exempt or expedited. Additionally, researchers who are using student data should have a clear privacy policy on their site and discuss the guidelines that researchers and their organizations follow when using data, including who can get access to data and for what purposes. These visible signs can help everyone understand how researchers are using data.
- Leverage existing resources and approaches for helping educate school-based partners on data and privacy. For example, CoSN created resources through knowledge-building activities for the education sector, building fundamental knowledge with a privacy toolkit covering FERPA, the Children's Online Privacy Protection Act (COPPA), the Protec-

tion of Pupil Rights Amendment (PPRA), and basic information security.[10] Resources should address the myths and the misinformation with respect for the perspectives of any vocal minority. In the higher education sector, educate partner organizations about efforts to use data appropriately for research.[11]
- At both the local and national levels, promote the value of research by framing the conversation in terms of how using student data can improve education, including for the most vulnerable. This conversation should describe the long and productive history of use of student data in educational research, describe what typically is and is not done with student data (for example, how research use differs from commercial use), and gather examples of positive data use.[12] Simultaneously, researchers should become familiar with efforts to limit school and company use of data[13] to ensure that distinctions are made between research and commercial uses of data. The field as a whole should create a strong evidence base to inform policy and public media discussions around student privacy versus education research, should have ready explanations about what education research is and why it is important, and should have concrete examples of why various actors should care about it.

Second, researchers should *take information security seriously*, address the security concerns of parents, learning institutions, and advocates, understand the risks in disclosure, and enlist the help of their organization's information technology department to rethink security frameworks.

- Establish best practices and guidelines for engaging schools and set evaluation criteria for what success looks like in terms of safeguarding student privacy. Even though data collected for research purposes may not be more attractive or valuable to hackers than that stored in banks or e-commerce sites, there are harms that could arise from disclosure; the public does not evaluate data breaches differently if they come from schools, companies, or research institutions.
- Hold education researchers to the same standards as companies in terms of safeguarding student data, and expect that schools and districts will want to verify security procedures. Producers of data shared with researchers should be able to trust and verify that their data are held safely if the risks of disclosure of those data are significant. Although human-subjects training tells us that assessing risk involves understanding the probability of disclosure and the severity of the harm that could occur from disclosure, research involving minors and/or public perceptions of disclosures should increase the "severity rating" of the harm.

[10] Another example, https://ferpasherpa.org/wp-content/uploads/2016/05/EduGuide_DataPrivacy_516.pdf, comes from a privacy partnership, FERPA|Sherpa, and is directed at teachers who may adopt online tools and apps that may collect student data.

[11] For example, the 2014 work on Learning Research in Higher Education, captured at http://asilomar-highered.info.

[12] See the Data Quality Campaign, http://dataqualitycampaign.org.

[13] For example, does a "Student Privacy Bill of Rights" such as advocated by the Electronic Privacy Information Center (https://epic.org/privacy/student/bill-of-rights.html) inhibit research use of data?

## CLOSING THOUGHTS

Participants agreed that we are in a new era where we can collect fine-grained data on every student action longitudinally: we can see when a student changes a page in a textbook and exactly what text they read. These data lead to new responsibilities, new uses, and new concerns. New ideas, such as a federal clearinghouse that holds process data for the public good and controls access by researchers, journalists, and individuals, are being offered, while states create a patchwork of regulations on student data.

Education researchers found great value in hearing the diverse views and perspectives of those present and felt both the need and the challenge to work on the myriad of issues around student privacy in educational research. Perhaps the most motivating challenge came near the end from Keith Krueger: "Frankly, you are late to the table. Legislators, school boards are already out there making these decisions [about limiting data use]. . . . If you want to build trust, you need to clearly articulate the why. Why . . . should we care that this research happens? . . . You need to talk very clearly about what we believe: that data should not harm kids. . . . It ought to be used to lift up [kids]. . . ."