**Workshop on Big Data in Education**

***Balancing the Benefits of Educational Research and Student Privacy***

# Advancing Educational Research and Student Privacy in the "Big Data" Era

Andrew Ho, *Harvard Graduate School of Education*

NATIONAL ACADEMY of EDUCATION

# Advancing Educational Research and Student Privacy in the "Big Data" Era

Andrew Ho, Harvard Graduate School of Education

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

## INTRODUCTION

Education is data intensive. Teachers interpret data in the form of verbal and nonverbal cues from students to adjust and improve their pedagogy. Parents receive data in the form of daily schoolwork, formal report cards, and informal stories from their children in car rides and at dinner tables. Teachers and school leaders routinely collect administrative data for compliance, monitoring, feedback, and improvement. As educational researchers, we systematize data collection and, sometimes, control the context of data creation, all to improve understanding of educational policy and practice. These systematic processes increase the usefulness of educational data but also increase the risks of exposure and harm to subjects. Collected data can be misplaced, stolen, or subjected to malicious analysis that reveals identities. Controlling the context of data creation may have a worse impact on subjects than not intervening at all.

Educational data systems are larger and more connected than ever before. As federal and state educational accountability systems have developed, administrative data have been centralized. State systems can increasingly describe the academic progress of individual students over time. Some have linkages to postsecondary and labor force outcomes. In the meantime, digital data collection and learning systems have proliferated in schools. Learning processes that were once informal, unstructured, and undocumented have become a data resource for students, teachers, parents, and researchers alike. Educational interactions are no longer bifurcated into informal classroom practices and formal administrative documentation; instead, classroom practices can be formally recorded, aggregated, and analyzed. Digital systems allow for detailed records of student struggles and successes in and out of school as well as feedback systems that can respond to student interactions in seconds rather than over semesters.

This promise has inspired high-profile efforts to collect and use educational data (e.g., inBloom, 2013) as well as high-profile backlash in the name of protecting student and family privacy (e.g., Singer, 2013). At the state level, a patchwork of legislation around family privacy and educational data records threatens to limit educational practice and educational research. At the federal level, the Family Educational Rights and Privacy Act (FERPA) provides little guidance about evolving threats to privacy from the use and analysis of digital education data. Institutional review boards (IRBs), the required evaluators and overseers of research practices in federally funded U.S. institutions, rely heavily on the same dated FERPA regulations and the 1978 Belmont Report.

In this paper, I provide a framework for evaluating the benefits of educational research using modern educational data systems and the risks to student and family privacy that this research can pose. I conclude that both have suffered from insufficient specification for this era, and I make a series of six distinctions to focus efforts on maximizing benefits while mitigating risks:

1. Between educational data and educational research: What are the privacy risks from educational research above and beyond those posed by educational environments themselves?
2. Between administrative data and "learning process" data.
3. Between primary and secondary research.
4. Among roles of policies and laws at the federal (e.g., FERPA), state, and institutional (e.g., IRBs) levels.
5. Between data ownership and data use.
6. Between potential harm and realized harm.

These distinctions help to motivate a series of recommendations for reassessing the benefits and risks of educational research for student privacy:

- Distinguish between the privacy risks posed by day-to-day educational practices and those specifically posed by educational research, above and beyond these educational practices.
- Reaffirm the benefits of research partnerships and accessible administrative data sets.
- Clarify the sufficiency of existing laws specifying proper use of these data.
- Extend both affirmation of benefits and legislation regarding proper research use of learning process data.
- Improve data security protocols and awareness of secure data practices among both researchers and practitioners; identify violations and prosecute offenders—enforce laws that already exist and expand policies as necessary when infrastructure and awareness is insufficient.
- Demystify data analytic techniques and hold them accountable to valid uses by educational practitioners.
- Following existing laws and regulations, do not release identified data to researchers without clear specification of the intended use of the data.

## DISTINCTIONS FOR DIFFERENTIATING RISKS TO PRIVACY

The benefits and the risks of educational data and research are dependent on the data type and research protocol. In this section, I draw three distinctions relevant to educational benefits and privacy risks. The first is between educational data and educational research itself. The second is between administrative data and "learning process data," with the former having more established potential and the latter showing great promise for impact on research and practice. The third distinction is between primary and secondary, with the latter using existing data and the former involving collaborations that control existing practices through treatments and interventions.

The vast majority of educational data is collected primarily for educational purposes rather than academic research, from day-to-day classroom teaching and school record keeping to the large-scale administration and monitoring of educational programs. As I review examples of risks to student and family privacy later in this paper, a useful recurring perspective is to identify the additional risks to privacy that arise from educational research, above and beyond counterfactual educational data use and collection. This is not to suggest that the additional privacy risks of educational research are always minimal, nor to suggest that, even if risks are minimal, researchers bear no ethical responsibility for advocating for privacy appropriate educational data practices. Nonetheless, privacy risks from educational research and educational practice should not be confounded.

Second, a useful broad distinction within educational data is that between administrative and process data. I define administrative data as demographic, behavioral, and educational achievement data for the purpose of administrating or monitoring educational programs and practices. Administrative data are typically collected at regular intervals on the order of months, semesters, and years. I contrast administrative data with "learning process data," or "process data" for short, which provide 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. Table 1 provides examples of variables that can comprise each data type.

Both administrative and process data can be online or offline, digital or analog. Although "online" is often a descriptor tied to modern data systems, the online nature of educational data is something of a red herring for privacy concerns. The questions that ultimately matter for privacy are how susceptible the data source is to malicious access and use, and how linkable the variables are to other variables that allow for uses beyond the initial intentions of those authorizing data use.

Finally, a consequential distinction for educational research data is that between secondary research using existing data sets and primary research that generates new data, such as those that involve interventions, treatments, and changes to existing educational interactions and contexts. Guidance for accessing educational research data is provided by FERPA, a legal framework that I review in greater detail in upcoming sections. As noted, educational research is also regulated institutionally by IRBs that evaluate potential studies in terms of their benefits and risks to participants. For secondary analyses, IRBs evaluate the risk of identifying individual students against the benefits of the research. This typically involves institutional researchers demonstrating that data are sufficiently deidentified such that the risk of exposure of personally identifiable information is minimal. For primary research that involves interventions and treatments, IRBs evaluate the proposed changes to the educational environment. They evaluate whether research benefits outweigh the risks and ensure that researchers acquire informed consent when appropriate. These long established protocols hew to principles dating back to the Belmont Report (Ryan et al., 1979). Recent efforts to update them for learning process data and exposure risks in online settings have reaffirmed rather than shifted these principles (Stevens & Silbey, 2014).

TABLE 1. Examples of Variables That Can Be Included in Administrative and Learning Process Data Sets

| Administrative Data | Learning Process Data |
|---|---|
| Name | Number of clicks |
| Birthdate | Time spent on site |
| Parent/guardian name | Average time per visit |
| Address | Last activity |
| Social Security Number | Number of videos viewed |
| Gender | Time spent on videos |
| Race/ethnicity | Interim test scores |
| National school lunch eligibility | Achievement score |
| English learner status | Engagement score |
| Disability status | Learning trajectory |
| Transcript (courses and grades) | Network score |
| Grade point average | Predicted grade |
| Standardized test scores | Number of forum posts |
| Absences | Words per forum post |
| Suspensions | Topic model score |

## DESCRIBING THE VALUE OF EDUCATIONAL RESEARCH

As researchers and IRBs assess the benefits of educational research compared to its privacy risks, a number of common appeals and types of benefits can be identified. Under the broad heading of effectiveness research, or "what works," examples include the Coleman Report (Coleman et al., 1966), which began a generations-spanning discussion about the relative contribution of school and nonschool factors to achievement outcomes and disparities (e.g., Hanushek, 1989; Lareau, 2000). There is also broad descriptive work—what Reardon (2004) describes as educational epidemiology. He and his colleagues recently released a data set of district-level outcomes derived from administrative data (Reardon et al., 2016) and allowed the public to visualize it online (Rich et al., 2016). These efforts rely on administrative student data to reveal inequities and enable identification of possible sources and remedies.

A third appeal of the value of research is to partnerships: collaborations between researchers and practitioners that themselves spur improvements. In a recent column, Susan Dynarski (2015) made a strong case for educational research in the face of privacy concerns. She argued for the sufficiency of existing federal laws and IRBs, which together limit data access to those with legitimate educational interests and research that has maximized benefits over risks. Regarding partnerships, she cited findings impossible without research collaborations between local researchers and schools, including high college dropout rates among Chicago high school graduates (Healy et al., 2014) and positive effects of a prekindergarten program in Boston Public Schools (Weiland & Yoshikawa, 2013). She identified three benefits to such research partnerships: outside expertise that benefits the school system, improved knowledge of the state of education, and improved identification of effective initiatives.

A similar type of appeal is to data and findings that are of more immediate use to local actors, including students, teachers, and schools. These appeals are so widespread that there are reports with headline-ready titles like, "19 Times Data Analysis Empowered Students and Schools" (Zeide, 2016). As an example of these types of projects, all of which necessitate some sharing of educational data, the Institute of Education Sciences lists 33 funded projects for its grant competition titled, "Researcher-Practitioner Partnerships in Education Research," a rate of more than 8 projects per year since 2013 (Institute of Education Sciences, 2016). I excerpt a representative list of titles and abstracts in Table 2 to illustrate this kind of value proposition, enabled by educational data sharing. Some of these are projects involving only administrative data, and some involve learning process data, as well.

There are particular types of benefits to research using process data that contrast with those of research using administrative data. These are due primarily to the more continuous temporal nature of process data and thus the in-the-moment learning that this research can describe and elucidate. Process data in education and psychology have revealed canonical learning trajectories in science and mathematics (National Research Council, 1999, 2001).

At a larger scale, carefully designed interactive systems can administer assessments that reveal learning progressions (Koedinger et al., 2015). Some research also takes advantage of these fast feedback cycles to iterate through multiple experiments, enabling identification and deployment of effective interventions (Heffernan & Heffernan, 2014).

Administrative data in the form of end-of-year test scores are still default educational outcomes in large-scale policy evaluations. In contrast, process data are helping to advance theories of learning in many domains as well as changing conceptions of how outcomes should be reimagined to support continuous learning (Bennett, 2015). A longstanding literature has

TABLE 2. Five Examples of IES-Funded Researcher-Practitioner Partnerships Using Student Data, from 33 Funded Projects Over 4 Years

| Year | Principal Investigator | Institution | Partner(s) | Title | Student Data |
|---|---|---|---|---|---|
| 2013 | Julian Betts | University of California, San Diego (UCSD) | San Diego Unified School District, San Diego Education Research Alliance at UCSD | Academic Trajectories and Policies to Narrow Achievement Gaps in San Diego | On-track indicators for all students with available prior-year administrative data: approximately 85,000 students attending all 170 schools in the district. |
| 2013 | Roger Weissberg | Collaborative for Academic, Social, and Emotional Learning | Washoe County School District (Nevada) | Creating a Monitoring System for School Districts to Promote Academic, Social, and Emotional Learning | Self-report and administrative data from all students in grades 5-12. The district serves more than 60,000 students in more than 90 schools. |
| 2014 | Paul Strand | Washington State University | Washington State Education Services District | A Partnership to Improve the Use of a Developmental Assessment Framework in Kindergarten | Assessment data from approximately 3,140 kindergarten students and 128 kindergarten teachers. |
| 2015 | June Ahn | University of Maryland | District of Columbia Public Schools | Blended Learning at Scale—Implementation and Analysis of Student Achievement in District of Columbia Public Schools | Administrative data for students in grades 2-8 in Washington, DC, public schools. Learning process data from blended learning technology platforms. |
| 2015 | Therese Dozier | Virginia Commonwealth University | Chesterfield County, Hanover County, Henrico County, and Richmond City Public Schools (Virginia) | META Researchers and Practitioners in Partnership (RPP) to Enhance Data Use Practice that Improves Student Learning | Eighty teachers from four school districts, using their student data from interim assessments, classroom assessments, and state assessments. |

presented games and simulations as models of this approach (Dede, 2015; Gee, 2003; Klopfer, 2008; Steinkuehler & Williams, 2006).

The practical benefits of process data for education arise from its "big data" features, which Laney (2001) described as volume, velocity, and variety. In education, these refer to numbers of student observations, the frequency of observations, and the number of types of observations (e.g., Table 1), respectively. Cope and Kalantzis (2016) describe big data in education as purposeful or incidental recording of continuous interactions incorporating varied data types that are accessible, durable, and subject to analysis. If data-collection contexts are well designed, these features can enable precise estimation of student learning trajectories. If feedback is well designed, these trajectories can facilitate student and teacher interactions that promote learning.

Instead of adopting the "big data" moniker to describe nonadministrative data, I use the "administrative" and "process" distinction. The labels more clearly describe the source of the data (in the case of administrative data from classroom, school, district, state, and federal records) and the use of the data (in the case of process data, describing and supporting con-

tinuous learning processes). Both administrative data sets and process data sets can be "big" in size. Both can be linked through unique identifiers to other variables and data sets. Both can be online or offline, digital or analog. Both can be recorded by and analyzed by software. Both have demonstrated usefulness in existing research and considerable promise for future research. And both raise concerns about family and student privacy.

## PRIVACY THREATS AND LEGAL PROTECTIONS IN EDUCATIONAL RESEARCH: FOUR CASES

Privacy issues in educational contexts are notable for their ability to distract from educational processes, which themselves require interpersonal interactions among students and educators. The concept of privacy itself is famously nebulous. In his book *Understanding Privacy,* Solove (2008) describes privacy as "a concept in disarray" (p. 1). He includes quotes from Hyman Gross, "the concept of privacy is infected with pernicious ambiguities" (1967), and Jonathan Franzen, "privacy proves to be the Cheshire cat of values: not much substance, but a very winning smile" (2003). These quotations reflect a concern, in the case of education, that unfocused debates about educational privacy will distract from legitimate educational processes and progress. Solove concludes that we should focus not on privacy in the abstract but on "specific activities that pose privacy problems" (p. 10). I discuss recent events that have posed privacy problems and attempt to disentangle the issues they raise.

### The Gayden Case: Individual Harm Versus Systematic Disclosure Risks in Research

In a case described by Bathon (2013), a Minnesota Public School seventh grader, named in the legal opinion as Kevin Gayden, was being called "dumb" and "stupid" at school. His taunters had found his school records near a dumpster in the school parking lot. According to the case, these records included information about the student's school and family history, as well as descriptions of his intellectual and functional abilities (Minnesota State Court of Appeals, 2006). The jury decided that the school district had violated the Minnesota Government Data Practices Act, which requires the school district to "establish appropriate security safeguards for all records containing data on individuals" (Minnesota Statute 13.05, n.d.).

The Gayden case raises the most tangible consequences of a disclosure of entrusted records: others using these records to cause direct harm to an individual. This case is relevant because I suspect it encapsulates the anxiety that students and families feel about unauthorized use of their educational data. At the same time, it is quite distinct from the cases I raise next. I consider six features of the Gayden case important to distinguishing among confounded privacy issues. Each feature can take the form of a question that clarifies whether the use of educational data for research is relevant, and whether it raises a new concern or one covered by existing statutes.

1. In the Gayden case, the data comprised the student's educational record, which was being used for educational purposes. It was neither released by a researcher nor intended for use in a research study. This distinction helps to clarify whether a privacy concern relates to educational data records generally or educational research data specifically. A legislative or policy solution that improves privacy in education should not necessarily be extended to educational research unless research itself poses additional threats. If

additional threats from research exist, these should be targeted specifically by additional solutions in turn. To be clear, it is essential to secure student records, and a variety of state policies have established supports and incentives for data security (Vance, 2016). The more specific focus of this paper is on additional privacy issues raised by educational research.

2. The release involved administrative paper records, not digital records. This distinction helps to clarify whether a privacy concern is educational data records generally or digital data records specifically. Digital data repositories tend to have larger numbers of records (due to the same features of accessibility, scalability, and searchability that make them promising for research), although the Gayden case makes clear that paper records are no guarantee of nondisclosure.
3. The Gayden release caused direct and demonstrable harm. In my search of recent incidents, I found no other example of direct and demonstrable harm from release of student records on the order of the Gayden case. Herold (2014a) draws a similar conclusion. He described a case where dentists may have used student directory information to target low-income students eligible for Medicaid with unnecessary dental procedures. A case was filed and settled, and a U.S. Senate report recommended ousting one dental management company from Medicaid (U.S. Committee on Finance, 2013). As horrific as unnecessary dental procedures on children are, these are both prosecutable under existing law and only speculatively connected to disclosure of student directory information. This distinction helps to clarify whether the privacy concern is one with any precedent for realized harms. However, as Herold acknowledges, it is difficult to directly connect a large-scale data breach to, for example, individual identity theft years later. This should not be confused with the absence of any consequences.
4. The Gayden release was unintentional—the result of sensitive documents carelessly tossed in the trash. The result was a civil case against a district rather than a criminal case against a hacker or discloser. The Privacy Rights Clearinghouse (n.d.) maintains a database of data breaches, which must be reported under laws in 47 states (Mintz-Levin, 2016; National Conference of State Legislatures, 2016). The Privacy Rights Clearinghouse database lists 777 educational breaches comprising 14.8 million records, including misplaced laptops, lost flash drives, hacking, malware, and unintended disclosures. Among unintentional releases, I have not been able to connect any to educational researchers or learning process data. Among intentional, malicious releases, similar to Dynarski (2015), I found no large-scale intrusion whose specific intent seems to be acquiring student achievement data, let alone learning process data. Instead, intentional breaches involve administrative data with the conspicuous inclusion of data with obvious financial value in legal or illegal markets, such as social security numbers.
5. The Gayden release was a violation of state law, not FERPA. As Schultze (2009) clarifies, FERPA was enacted under congressional spending power, and its force arises from its ability to terminate federal funds. It does so in cases of systematic, not individual, violations of student privacy, leaving consequences for individual releases like those in the Gayden case to state laws.
6. Finally, the harm incurred by the release arose directly from the data in the records, and the identification of sensitive student information required no additional analysis or linkage to other data. This contrasts with the case in the next section.

The Gayden case involved the unintended release of sensitive administrative data about a student that led to realized harm, a case already prosecutable under existing state laws. It does not involve research at all. I contrast the Gayden case with one involving learning process data explicitly for research, where I unpack a key protection in the use of student data for research: deidentification.

### The HarvardX-MITx Deidentified Data Set: Addressing the Threat of Reidentification

In 2014, I and colleagues from Harvard and the Massachusetts Institute of Technology released a report (Ho et al., 2014) that described characteristics of an emerging online learning context: the massive open online course (MOOC). The report included the demographics of participants and defined new terms and new variables relevant to the heterogeneous registrants and their asynchronous interactions. Shortly after releasing the report, we also released a public data set that contained some of the variables that generated the descriptive statistics (MITx & HarvardX, 2014). To my knowledge, it remains the only publicly accessible participant-level MOOC data set; it has been downloaded more than 5,500 times.[1] In follow-up papers (Angiuli et al., 2015; Daries et al., 2014), members of the research team demonstrated that the deidentification process necessary to release the data to the public degraded statistical inferences and precluded direct replication and extension of results in the original report. We described the tension between making data available for widespread replication—a key feature of good science—and protecting the identities of those in the data set.

We released the HarvardX-MITx deidentified data set after ensuring that it contained no personally identifiable information (PII). I excerpt the full FERPA definition of PII below and add emphasis to subpart (f), which acknowledges that seemingly nonidentifying data like zip codes, birthdates, and gender can, in combination, identify many individuals uniquely (Sweeney, 2000). Daries et al. (2014) describe how learning process data from open online interactions can be particularly susceptible to triangulation by combining variables in the data set and/or linking them to variables acquired elsewhere. For example, a supposedly deidentified data set may not include a student's name but includes her course grade and her number of online forum posts. Online, an analyst can count the number of forum posts made by each username. Linking the two data sets identifies the student's username with her course grade.

**Personally Identifiable Information (Authority: 20 U.S.C. 1232g):**

The term includes, but is not limited to— (a) The student's name; (b) The name of the student's parent or other family members; (c) The address of the student or student's family; (d) A personal identifier, such as the student's social security number, student number, or biometric record; (e) Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name; (f) **Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty**; or (g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates. "Record" means any information recorded in any way, including, but not limited to, hand writing, print, computer media, video or audio tape, film, microfilm, and microfiche.

[1] Tsinghua University's open online course initiative, XuetangX, sponsored a data mining competition with publicly available MOOC learning process data (KDD Cup, 2015). The data and a successful prediction method are described briefly by a member of the winning team (Ozaki, 2015). The data are no longer available at the original website.

Data preprocessing can reduce the likelihood of reidentification by rounding data into coarser and coarser categories, until no unique combinations of values exist. Daries et al. (2014) interpret guidance from the U.S. Department of Education's Privacy Technical Assistance Center (2012) to decide on a conservative threshold for nonuniqueness, where all combinations of variables have at least five observations of the same combination. This criterion is known as *k*-anonymity (Sweeney, 2002), where Daries et al. (2014) choose $k = 5$. Once data are coarsened to this extent, we determined that the data may be released under a FERPA exception (emphasis added):

> § 99.31 (b)(1) De-identified records and information. An educational agency or institution, or a party that has received education records or information from education records under this part, may release the records or information without the consent required by § 99.30 after the removal of all personally identifiable information provided that the educational agency or institution or other party has made a **reasonable determination that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information.**

Angiuli et al. (2015) illustrate the privacy information tradeoff by showing degraded interpretations of research data for increasing values of *k*. Other privacy solutions that balance research needs include differential privacy (Dwork, 2006), where statistical queries of raw data are limited in number and resolution to prevent identification of individuals. This also suffers from distortion in practical research scenarios (Bambauer et al., 2014).

These technical privacy solutions have greatest value in situations where we desire to make the data available to the public, where adherence to click-through agreements is neither incentivized nor practically enforceable. These deidentification strategies are motivated by an extreme interpretation of the FERPA stipulation that a "reasonable person in the school community" cannot reidentify the data, where the reasonable person is someone proficient with scraping large amounts of data online and who makes a concerted effort to reidentify data through linkages to other available data sources. In situations where recipients of the data are researchers who are bound to agreements through the strength of affiliated institutions, with civil and criminal penalties associated with any efforts they make to breach, release, or reidentify data, the added value of these technical privacy solutions does not seem to warrant the harm done to the data.

Instead, research is already enabled by more straightforward data preprocessing, where sensitive identifiers like names and social security numbers are replaced by other unique identifiers before transfer to research teams, and detailed memoranda of understanding govern the storage and use of the transferred data. A variety of solutions serve other organizations like the Internal Revenue Service (IRS) and the Social Security Administration. Mervis (2014) describes how the IRS made tax records available to researchers by releasing a perturbed data set for researchers to test their code and, in some cases, hiring researchers as federal employees.

In 2008, the National Research Council (NRC) convened a relevant workshop titled *Protecting Student Records and Facilitating Educational Research* (National Research Council, 2009). There, representatives from Michigan (Barbara Schneider) and North Carolina (Helen Ladd and Jeff Sellers) each described partnerships between researchers and the state that enabled privacy appropriate research. Dynarski (2014) has described specific proposals for enabling educational research, particularly longitudinal research, using administrative data. Card et al. (2010) have made a similar case to the National Science Foundation.

For educational data, protocols would either be FERPA compliant by ensuring the data released are deidentified, as defined above, or by considering the data as PII and eligible for release to researchers under a separate exception, where "the disclosure is to organizations conducting studies for, or on behalf of, educational agencies or institutions to (A) Develop, validate, or administer predictive tests; (B) Administer student aid programs; or (C) Improve instruction." The full quote and context is here, with emphasis added:

> § 99.31 Under what conditions is prior consent not required to disclose information?
>
> (a) An educational agency or institution may disclose personally identifiable information from an education record of a student without the consent required by § 99.30 if the disclosure meets one or more of the following conditions:
>
> … (**6)(i) The disclosure is to organizations conducting studies for, or on behalf of, educational agencies or institutions to: (A) Develop, validate, or administer predictive tests; (B) Administer student aid programs; or (C) Improve instruction.** (ii) An educational agency or institution may disclose information under paragraph (a)(6)(i) of this section only if— (A) The study is conducted in a manner that does not permit personal identification of parents and students by individuals other than representatives of the organization that have legitimate interests in the information; (B) The information is destroyed when no longer needed for the purposes for which the study was conducted; and (C) The educational agency or institution enters into a written agreement with the organization that— (1) Specifies the purpose, scope, and duration of the study or studies and the information to be disclosed; (2) Requires the organization to use personally identifiable information from education records only to meet the purpose or purposes of the study as stated in the written agreement; (3) Requires the organization to conduct the study in a manner that does not permit personal identification of parents and students, as defined in this part, by anyone other than representatives of the organization with legitimate interests; and (4) Requires the organization to destroy or return to the educational agency or institution all personally identifiable information when the information is no longer needed for the purposes for which the study was conducted and specifies the time period in which the information must be returned or destroyed.

Educational research has thus been protected on two fronts. First, researchers must ensure that their data are reasonably deidentified by setting up firewalls between PII and the data they receive, to enable "a reasonable determination that a student's identity is not personally identifiable." This intended use is made even more explicit in the next paragraph, where the legislation spells out a mechanism for use: "for the purpose of education research by attaching a code to each record that may allow the recipient to match information received from the same source" (FERPA, n.d.). Second, if educational researchers cannot claim that data are deidentified, they must ensure that the use is "for, or on behalf of, educational agencies or institutions to … improve instruction."

The breadth of the exception to "improve instruction" is notable, leaving latitude for researchers and organizations collaborating on such studies to make arguments for both direct and indirect mechanisms of improvement. As Schultze (2009) has argued in his interpretation of the opinion by Breyer (2002), "(1) FERPA is indeed vague and not easily understood; (2) educational experts, not courts, should be the primary arbiters of FERPA's language; and (3) Congress intended FERPA to leave plenty of room for common sense and effective teaching methods, not to curtail smart teaching choices that happen to disclose innocuous information implicitly and indirectly" (Schultze, 2009, p. 232).

This is consistent with flexibility in interpreting regulations that would enable educational research. In any secondary analysis of data without direct identifiers, an inadvertent

disclosure of an individual's identity may be unlikely enough, and an educational researcher unlikely enough to be incentivized toward such a disclosure, that restricting distribution of educational research data on the basis of advanced reidentification methods may be unnecessary. We should be able to reconcile the fact that deidentified data are, with concerted effort, reidentifiable, with the fact that the researcher undertaking such effort would be violating an agreement for uncertain ends. Existing protocols that require memoranda of understanding from institutionally affiliated researchers have been sufficient to ensure compliance. Restricting the possible benefits of educational research by imposing additional hurdles or penalties is only warranted if there is legitimate reason to think these will decrease risks above and beyond existing legislation.

It is less clear whether institutions should distribute putatively deidentified student-level data to online data repositories that allow access after one-click acceptances of online agreements. In the summary of the NRC workshop (2009), participants emphasized the importance of establishing trust between researchers and practitioners by developing research partnerships. A "hit-and-run" approach to data access, where unconnected researchers collect theoretically reidentifiable secondary data and disappear, may not be supported by FERPA for three reasons: (1) it is not "for, or on behalf of, educational agencies or institutions," (2) it is less likely to "improve instruction" in any direct way, and (3) it decreases accountability to memoranda of understanding, including agreements not to reidentify data.

## The Facebook Experiment: Corporate Versus Academic Responsibility for A/B Testing

In 2012, Facebook conducted an "A/B test"—industry parlance for a randomized experiment—that tested whether users who viewed posts with less positivity were subsequently more likely to post with positive words than those who viewed posts with less negativity. The researchers described this as an investigation of whether "emotional states can be transferred to others via emotional contagion" (Kramer et al., 2014, pp. 8788-8790). The effect was small, decreasing from around 5.3 percent positive words to 5.2 percent, but the difference was statistically significant, in part due to the substantial statistical power derived from having around 155,000 accounts per condition. Researchers from Cornell University obtained IRB approval to analyze the secondary data, and they published their findings in the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) in 2014.

The publication of the Facebook experiment caused many to raise questions about the ethics of, to use the authors' metaphor, infecting people with a contagion without informed and affirmative consent required under standard human subjects research protocols. Solove (2014) observed that Facebook carried out the research under the Data Use Agreement (DUA) that Facebook users opt into, which notes in part that Facebook "may use the information we receive about you ... for internal operations, including troubleshooting, data analysis, testing, research and service improvement." Solove observes that clicking through a data use agreement is in no way affirmative consent under the IRB guidelines to which federally funded researchers must adhere. Instead, the Cornell IRB approved the study on the basis that they were using secondary data that Facebook had already collected.

The PNAS editor-in-chief described the decision in a follow-up editorial:

> Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the

> authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out. (Verma, 2014, p. 10779)

The Facebook experiment raises difficult questions about whether and if so how to ensure that for-profit companies adhere to the same or similar ethical guidelines as researchers who directly or indirectly receive federal funding. Many pledges and general documents exist that reflect the same principles as those in the Belmont Report (Stevens & Silbey, 2014; Student Privacy Pledge, 2014), but there is no independent oversight of any internal screening for A/B experimental conditions, nor any comparable incentive to the loss of federal funding. At the very least, corporations should be aware of the foundational research principles. The Facebook experiment would have been improved had the treatment been more positive content and the control had been business as usual. The so-called contagion of negative content, with a result that can be construed as harm, was the most problematic component and could have easily been avoided.

One approach is for researchers to not participate in analysis of secondary data that have not been collected under standard IRB guidelines. Meyer (2014) argues persuasively that such shaming and shunning would simply drive A/B testing underground (if it has not already), with corporations conducting research without reaching out to academics or publication outlets. Solove (2014) and Hill (2014) argue that user expectations are a reasonable measure of whether or not something is ethical, and that the Facebook experiment was outside what most would expect. This implies that one solution to the problem is to ensure that this kind of experimentation is expected by users; however, public awareness campaigns and click-through DUAs are no guarantee.

### The Case of inBloom: Without Specified Uses, Imagined Uses

In 2011, the Shared Learning Collaborative, an initiative funded largely by the Gates Foundation and the Carnegie Corporation, launched with the goal of creating a common data infrastructure across schools, districts, and states. In 2013, it relaunched as the nonprofit organization inBloom, with pilot partnerships in nine states, continuing its emphasis on personalized learning, individualized instruction enabled by data dashboards, and partnerships with third-party education technology companies. Just more than 1 year later, on April 21, 2014, inBloom announced it would wind down its operations. In CEO Iwan Streichenberger's announcement of the closing, he cited "generalized public concerns about data misuse" (Herold, 2014b). Well-organized opposition to inBloom led to state legislation and public sentiment that the organization could not overcome.

In New York, for example, education law was changed to add that "an educational agency may opt out of providing personally identifiable information to a SLISP [shared learning infrastructure service provider] or data dashboard operator for the purpose of creating data dashboards" (McKinney's Education Law, 2014). The law also prohibited state-level agreements: "the commissioner and the department are hereby prohibited from providing any student information to a SLISP." In California, similarly, an operator of such a service must "delete a student's covered information if the school or district requests deletion of data under the control of the school or district." And operators cannot "use information, including persistent unique identifiers, created or gathered by the operator's site, service, or application, to amass a profile about a K-12 student except in furtherance of K-12 school purposes."

The demise of inBloom offers lessons for addressing privacy concerns in educational research. Key features cited by critics of inBloom included the for-profit motivations of potential inBloom partners, the unsecure nature of online data, and unease about the use of data for student and teacher classification (Herold, 2014a; Parent Coalition for Student Privacy, 2014). Both inBloom and the backlash to inBloom placed little emphasis on learning process data, focusing instead on the benefits and risks of online integration and interoperability of existing, mostly administrative, data. In particular, critics articulated fears that partners would use the data to advertise services, and that student data could become a commodity sold to other parties and uncontrollable in bankruptcy proceedings. At around the same time, companies began to address these concerns by drafting and signing the Student Privacy Pledge (2014), where signers commit themselves against disclosing, sharing, or selling data beyond legitimate educational purposes. More than 200 companies have signed the pledge, but it has no legal standing.

Like FERPA, state and proposed federal laws have explicit acknowledgments of research. A recent federal proposal, the Student Digital Rights and Privacy Act, includes the clarification that "nothing in this Act prohibits an operator from ... disclosing de-identified and aggregated covered information for research and development, including (i) research, development, and improvement of educational sites, services, and applications; and (ii) advancements in the science of learning" (H.R. 2092). Indeed, the greater threat of privacy concerns for research may not be that legislation will prohibit research but that districts will opt out of participation and collaboration, potentially restricting the generalizability of research findings to those who participate in data sharing.

The essential substantive limitation of inBloom was its underspecified theory of action for improving educational practice. Data in and of itself do not improve practice unless they can answer relevant questions that participants in the process are asking, with answers that inform subsequent actions. The inBloom strategy seemed to start with the promise of compiling data and letting third-party operators identify the questions that data could answer. A convergent approach would have invested more proportionately in identifying the questions that students, parents, and teachers were asking from the beginning, as well as the actionability of answers. The dashboard metaphor is discouragingly apt because of how rarely drivers actually use the varied elements in traditional dashboards, and how rarely these elements inform actions beyond refueling and slowing when speeds exceed limits. Focusing schools, service providers, and researchers together on specific data uses that hold the promise of benefiting students, teachers, parents, and administrators, over the near or long term, is both consistent with FERPA and a reminder of the benefits of legitimate educational data use.

## The Problem with Prediction

I have argued elsewhere that current data-driven movements have focused disproportionately on predictive and diagnostic modeling (Ho, 2015). Future gains in this area are more likely to advance statistical and algorithmic theory than make significant gains in practice, whereas improvement of more holistic feedback loops, facilitating the cycle from question generation to data analysis to informed decisions, are both poised for and demanding of greater progress. Predictive and diagnostic modeling continue to advance in the areas of intelligent tutoring systems and games, where the users are individual students in contexts outside of or indifferent to formal schooling contexts. These include controlled online environments (Koedinger et al.,

2013) as well as massive open online courses (Whitehill et al., 2014). In school environments where FERPA and privacy issues are backdrops, research that promises useful instructional information without reciprocation or partnership will predictably and should rightly be treated with skepticism.

Data mining in the absence of theory or design will always be incomplete as an educational intervention. We may achieve accurate predictions, but, in an educational context, we want our predictions to be wrong. Precisely, they should be biased negatively, where every prediction is communicated to teachers and learners in an actionable way and ultimately overcome. Inserting data analytics between students and teachers in classroom contexts in a way that improves teaching and learning continues to be challenging and an effort too rarely undertaken (for exceptions, see Klopfer & Perry, 2014; O'Rourke et al., 2016). The failures of initiatives like inBloom are less technical than political and substantive; data gathered without engaging teachers and parents are reframed as an attack on, or an end around, formal schooling. Importantly, existing legislation allows for research data use in the context of legitimate educational interests, a testament that the trust that ameliorates privacy concerns is born not only of stated intentions but investment in outreach and relationships.

## RECOMMENDATIONS: PARALLEL PATHS FORWARD

This paper has reviewed key contrasts and cases in educational research in the era of big data, process data, corporate research, and mergeable, reidentifiable data sets. These distinctions and cases provide possible strategies for advocating for and advancing privacy appropriate educational research. I list these recommendations as part of a framework for consideration.

First, reassert that educational data are an inherent necessity to everyday educational practice. Education is fundamentally data intensive—maximizing privacy in the abstract requires withholding data that inform teaching and learning. This is part of the reason why student privacy laws that emphasize "the right to be forgotten" are not only infeasible but poorly framed. They suggest that education is a service rendered to an individual as opposed to a participation in an exchange. A right to be forgotten interferes with the right of others—fellow students, teachers, counselors, and principals—to remember and to learn. If educators are already collecting data for legitimate educational purposes, we can focus on the additional privacy risks, like those associated with sharing deidentified data with researchers bound by agreements, and evaluate their added risks with clarity and specificity. This also rightly raises standards whenever researchers seek to partner to collect data beyond those that educators are inclined to generate or seek as part of their natural educational practices.

Second, reaffirm the research benefits of research partnerships and accessible administrative data sets. Clarify that established secure protocols and partnerships with reputable researchers have existed for years (National Research Council, 2009), with no realized harms to student privacy, and with considerable benefits in the form of providing schools with expertise, identifying effective programs, and revealing achievement disparities (Dynarski, 2015). Third, reaffirm the legal protections for such research under existing laws. These include FERPA allowances for research for educational organizations to improve instruction, as well as FERPA exceptions for deidentified data, particularly those distributed to researchers with established incentives to comply with signed agreements that forbid reidentification.

Fourth, extend the affirmation of these benefits and protocols to educational research using learning process data. As I have posited, privacy concerns are not likely to be heightened by

research uses of learning process data above and beyond online use of administrative data. In fact, a focus on learning process data could ameliorate privacy concerns because of the specificity of the scope of its intended use, assuming that this use is to identify canonical learning progressions, guide instruction, and accelerate student learning.

Fifth, improve data security protocols and awareness of secure data practices among both researchers and practitioners. Identify violations and prosecute offenders—enforce laws that already exist and expand policies as necessary when infrastructure and awareness are insufficient. Sixth, demystify data analytic techniques for dashboards and other metrics intended for practitioner use. Assess their validity not merely by predictive accuracy but by the demonstrated likelihood that interpretations and uses of resulting metrics will be appropriate.

Finally, for identified data, establish partnerships to ensure a bridge between exploratory predictive modeling, on the one hand, and, on the other, the use of predictions to inform teaching and learning practices. The specified use of identified data must extend beyond "prediction" to informing teacher and student actions that ultimately render predictions of, for example, low student achievement, incorrect. Identified educational data should not be granted with an assumption that predictive modeling efforts alone will lead educators and students to more positive outcomes when they might just as easily lead to stigmatization and entrenchment. For identified data, close partnerships between analysts and educators should improve both statistical analyses and substantive responses to their results.

## REFERENCES


Angiuli, O., Blitzstein, J., & Waldo, J. (2015). How to De-Identify Your Data. *Communications of the ACM 58*(12):48-55. Available at http://cacm.acm.org/magazines/2015/12/194640-how-to-de-identify-your-data/abstract.

Bambauer, J., Muralidhar, K., & Sarathy, R. (2014). Fool's Gold: An Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment and Technology Law* 16:701-755.

Bathon, J. (2013, October). How Little Data Breaches Cause Big Problems for Schools. *THE Journal* 40(10):26-29.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education* 39:370-407.

Breyer, S. (2002). *Gonzaga Univ. v. Doe*, 536 U.S. 273, 278.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding Access to Administrative Data for Research in the United States. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1888586.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of Educational Opportunity. Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Cope, B., & Kalantzis, M. (2016). Big Data Comes to School: Implications for Learning, Assessment, and Research. *AERA Open* 2(2):1-19.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Seaton, D. T., Ho, A. D., & Chuang, I. (2014). Privacy, Anonymity, and Big Data in the Social Sciences. *Communications of the ACM* 57(9):56-63. Available at http://cacm.acm.org/magazines/2014/9/177926-privacy-anonymity-and-big-data-in-the-social-sciences/fulltext.

Dede, C. (Ed.). (2015). Data-Intensive Research in Education: Current Work and Next Steps: Report on Two National Science Foundation-Sponsored Computing Research Education Workshops. Washington, DC: Computing Research Association. Available at http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf.

Dwork, C. (2006). Differential Privacy. In *International Colloquium on Automata, Languages, and Programming*, edited by M. Bugliesi et al..Berlin, Germany: Springer-Verlag.

Dynarski, S. (2014). Building Better Longitudinal Surveys (on the Cheap) Through Links to Administrative Data. Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Educational Research Community. Washington, DC: National Academy of Education. Available at http://www.naeducation.org/cs/groups/naedsite/documents/webpage/naed_160695.pdf.

Dynarski, S. (2015, June 14). When Guarding Student Data Endangers Valuable Research. *The New York Times.* Available at http://www.nytimes.com/2015/06/14/upshot/when-guarding-student-data-endangers-valuable-research.html.

Family Educational Rights and Privacy Act. (n.d.). 20 U.S.C. § 1232g; 34 CFR Part 99. Available at https://ed.gov/policy/gen/guid/fpco/pdf/2012-final-regs.pdf.

Franzen, J. (2003). How to Be Alone. New York: Picador.

Gee, J. P. (2003). What Video Games Have to Teach Us About Learning and Literacy. New York: Palgrave/Macmillan.

Gross, H. (1967). The concept of privacy. *New York University Law Review, 42*, 34-53.

Hanushek, E. A. (1989). The Impact of Differential Expenditures on School Performance. *Educational Researcher* 18(4):45-51.

Healy, K., Nagaoka, J., & Michelman, V. (2014). The Educational Attainment of Chicago Public Schools Students: A Focus on Four-Year College Degrees. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Heffernan, N., & Heffernan, C. (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: 470-497.

Herold, B. (2014a). Danger Posed by Student-Data Breaches Prompts Action. *Education Week.* Available at http://www.edweek.org/ew/articles/2014/01/22/18dataharm_ep.h33.html.

Herold, B. (2014b). inBloom to Shut Down Amid Growing Data-Privacy Concerns. *Education Week*. Available at http://blogs.edweek.org/edweek/DigitalEducation/2014/04/inbloom_to_shut_down_amid_growing_data_privacy_concerns.html.

Hill, K. (2014, June 29). Facebook doesn't understand the fuss about its emotion manipulation study. *Forbes*. Available at https://www.forbes.com/sites/kashmirhill/2014/06/29/facebook-doesnt-understand-the-fuss-about-its-emotion-manipulation-study/#1550db2166db.

Ho, A. D. (2015). Before "Data Collection" Comes "Data Creation." In *Data-Intensive Research in Education: Current Work and Next Steps*, edited by C. Dede. Pp. 34-36. Washington, DC: Computing Research Association.

Ho, A. D., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The First Year of Open Online Courses. HarvardX and MITx Working Paper No. 1. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263.

inBloom. (2013). Some Facts About inBloom, Inc. Available at http://www.classsizematters.org/wp-content/uploads/2013/05/inBloom-reply_Assemblyman-ODonnell.pdf.

Institute of Education Sciences (2016). Search Funded Research Grants and Contracts. Available at https://ies.ed.gov/funding/grantsearch/index.asp?mode=1&sort=1&order=1&searchvals=&SearchType=or&slctProgram=81&slctGoal=0&slctCenter=0&FundType=1&FundType=2.

KDD Cup. (2015). KDD Cup 2015: Predicting dropouts in MOOC. Available at http://kddcup2015.com/information.html.

Klopfer, E. J. (2008). Augmented Learning: Research and Design of Mobile Educational Games. Cambridge, MA: MIT Press.

Klopfer, E. J., & Perry, J. (2014). UbiqBio: Adoptions and Outcomes of Mobile Biology Games in the Ecology of School. *Computers in the Schools* 31:43-64.

Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., & Stamper, J. (2013). New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization. *AI Magazine* 34(3):27-41.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z., & Rose, C. P. (2015). Data Mining and Education. *Wiley Interdisciplinary Reviews: Cognitive Science* 6:333-353. doi: 10.1002/wcs.1350.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* 111:8788-8790.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Stamford, CT: META Group. Available at https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lareau, A. (2000). Home Advantage: Social Class and Parental Intervention in Elementary Education. Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Mervis, J. (2014, May 22). How Two Economists Got Direct Access to IRS Tax Records. *Science Insider*. Available at http://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records.

Meyer, M. N. (2014). Misjudgements will drive social trials underground. *Nature* 511:265.
Minnesota State Court of Appeals. (2006). Unpublished opinion. Available at https://mn.gov/law-library-stat/archive/ctapun/0604/opa050649-0418.htm.
Minnesota Statute 13.05 (n.d.). Available at https://www.revisor.mn.gov/statutes/?id=13.05.
Mintz-Levin (2016). State Data Security Breach Notification Laws. Available at https://www.mintz.com/newsletter/2007/PrivSec-DataBreachLaws-02-07/state_data_breach_matrix.pdf.
MITx & HarvardX. (2014). HarvardX-MITx Person-Course Academic Year 2013 De-Identified Dataset, Version 2.0. Harvard Dataverse. Available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147.
National Conference of State Legislatures. (2016). Security Breach Notification Laws. Available at http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.
National Research Council. (1999). How People Learn: Brain, Mind Experience, and School. Edited by M. Suzanne Donovan, John D. Bransford, and James W. Pellegrino. Washington, DC: National Academy Press.
National Research Council. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Edited by J. W. Pellegrino, N. Chudowsky, and R. Glaser. Washington, DC: National Academy Press.
National Research Council. (2009). Protecting Student Records and Facilitating Education Research: A Workshop Summary. Washington, DC: The National Academies Press.
O'Rourke, E., Chen, Y., Ballweber, C., & Popovic, Z. (2016). Personalized Learning and Its Behavioral Impact on the Classroom Ecosystem. *Educational Psychologist* 36(2):89-101.
Ozaki, K. (2015). Techniques (tricks) for data mining competitions. Available at https://speakerdeck.com/smly/techniques-tricks-for-data-mining-competitions.
Parent Coalition for Student Privacy. (2014). Letter to Congress. Available at http://www.studentprivacymatters.org/wp-content/uploads/2014/07/Parent-Coalition-for-Student-Privacy-Letter-to-Congress-7-22-14.pdf.
Privacy Rights Clearinghouse. (n.d.). Chronology of Data Breaches. Available at http://www.privacyrights.org/data-breach.
Privacy Technical Assistance Center. (2012). *Frequently Asked Questions—Disclosure Avoidance*. Washington, DC: U.S. Department of Education. Available at http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf.
Reardon, S. F. (2004). Examining patterns of development in early elementary school using ECLS-K data. *Education Statistics Quarterly* 6(3):16-18.
Reardon, S. F., Kalogrides, D., Ho, A. D., Shear, B., Shores, K., Fahle, E. (2016). *Stanford Educational Data Archive.* http://purl.stanford.edu/db586ns4974.
Rich, M., Cox, A., & Bloch, M. (2016, April 29). Money, Race, and Success: How Your School District Compares. *The New York Times.* Available at http://www.nytimes.com/interactive/2016/04/29/upshot/money-race-and-success-how-your-school-district-compares.html.
Ryan, K. J., Brady, J., Cooke, R., Height, D., Jonsen, A., King, P., Lebacqz, K., Louisell, D. W., Seldin, D. W., Stellar, E., & Turtle, R. H. (1979). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Washington, DC: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.
Schulze, L. N. (2009). Balancing Law Student Privacy Interests and Progressive Pedagogy: Dispelling the Myth That FERPA Prohibits Cutting-Edge Academic Support Methodologies. *Widener Law Journal* 19:215-276.
Singer, N. (2013, October 5). Deciding Who Sees Students' Data. *The New York Times*. Available at http://www.nytimes.com/2013/10/06/business/deciding-who-sees-students-data.html.
Solove, D. J. (2008). Understanding Privacy. Cambridge, MA: Harvard University Press.
Solove, D. (2014). Facebook's psych experiment: Consent, privacy, and manipulation. Available at https://www.linkedin.com/pulse/20140630055215-2259773-the-facebook-psych-experiment-consent-privacy-and-manipulation.
Steinkuehler, C. A., & Williams, D. (2006). Where Everybody Knows Your (screen) Name: Online Games as "Third Places." *Journal of Computer-Mediated Communication* 11:885-909.
Stevens, M. L., & Silbey, S. S. (2014). The Asilomar Convention for Learning Research in Higher Education. Asilomar, CA. Available at http://asilomar-highered.info/asilomar-convention-20140612.pdf.
Student Privacy Pledge. (2014). Available at https://studentprivacypledge.org.
Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper #3. Pittsburgh, PA: Carnegie Mellon University.

Sweeney, L. (2002). k-anonymity: A model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems* 10:557-570.

U.S. Committee on Finance. (2013). Joint Staff Report on the Corporate Practice of Dentistry in the Medicaid Program. Available at http://www.adea.org/uploadedFiles/ADEA/Content_Conversion_Final/policy_advocacy/Documents/emailDist/Report_on_Corporate_Dentistry.pdf.

Vance, A. (2016). Trends in Student Data Privacy Bills. *Policy Update* 23(13):1-2. Alexandria, VA: National Association of State Boards of Education. Available at http://www.nasbe.org/wp-content/uploads/Vance_2016-State-Final.pdf.

Verma, I. M. (2014). Editorial expression of concern and correction. *Proceedings of the National Academy of Sciences of the United States of America* 1111:10779.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills. *Child Development* 84:2112-2130.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2014). Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout. *Proceedings of the 8th International Conference on Educational Data Mining.* Pp. 171-178.

Zeide, E. (2016). 19 Times Data Analysis Empowered Students and Schools: Which Students Succeed, and Why? Washington, DC: Future of Privacy Forum.