Big Data in Education

Balancing the Benefits of Educational Research and Student Privacy

A Workshop Summary

NATIONAL
ACADEMY
of
EDUCATION

INTRODUCTION

This is a critical time to understand the benefits and risks of educational research using large data sets. Massive quantities of educational data can now be stored, analyzed, and shared. State longitudinal data systems can track individual students from pre-K through college and work. Districts and schools keep detailed data on individual academic performance, behavior, and educational needs. Schools provide portals for parents to check student assignments and grades. Software developers and researchers collect data from applications used for instruction that include keystroke-level information about student decision making.

Meantime, "big data" are becoming a "big concern" to parents and privacy advocates who worry that student information will be placed "in unreliable hands or put to nefarious uses."[1] Commonly voiced concerns focus on privacy breaches, hacking, the use of data by commercial software developers for marketing purposes, and the possibility that sensitive information (e.g., regarding learning disabilities, behavior problems, or test scores) might limit future opportunities for students.

These privacy concerns have inspired some major policy shifts. Emerging state and federal legislation aim to set limits on how much data can be collected, how long they are stored, and how they are used for research, commercial, and other purposes. At the federal level, these include calls for changes to the Family Educational Rights and Privacy Act (FERPA), such as prohibiting research that is not aimed at improving the instruction or testing of the specific students involved in the research, requiring parents to "opt in" to all research, and requiring the deletion of student records when a student leaves an institution.[2] States have passed student privacy laws that have variously sought to restrict the types of information that can be collected, limit access to data by third parties, and beef up state agency privacy safeguards.[3]

As legislators and the public discuss state and federal policy changes, a full accounting of educational research benefits and risks is necessary. Longitudinal data systems that track the progress of individual students have enormous potential for illuminating why some students thrive and others do not. There are many questions to which parents, practitioners, and policy makers would like answers that can be explored by linking administrative data with achievement data and tracking students over time. The use of learning process data is opening a window into student thinking and learning in ways that can substantially improve educational outcomes. Interactive technologies are able to record correct and incorrect answers to questions, response times, students' requests for hints, and the aspects of the software with which students interact. When such data are linked to administrative data, there is the promise of understanding, with greater depth than ever before, how student background and prior experiences interact with learning. Now personalized instruction can improve motivation and learning for these students.

Combining administrative and learning process data can provide evidence about cognitive processes and instructional effects that were unimagined just a decade ago. Researchers can use administrative data and learning process data to inform policies that improve student learning and teaching for all students.[4] In order to understand and improve teaching and learning, education data must be available to researchers, and the privacy of children and families must be protected.

A recent workshop held by the National Academy of Education (NAEd), upon which this report is based, reviewed the benefits of educational research using modern data systems, the risks to the privacy of families and children, and technical and political solutions for maximizing benefits and minimizing risks. For more information on the workshop, including a background paper, panel videos and summaries and further resources, visit naeducation.org/bigdata.

COMMON LANGUAGE

Terms such as big data and privacy are seen with increasing frequency. There are weekly, if not daily, news reports about the use of large-scale data in research or as a means to solve real-world problems. Privacy concerns and security breaches, including the release of personally identifiable information, are also widely reported. Large government and private-sector databases have been hacked, exposing billions of users to potential identity theft; trusted employees have stolen classified governmental information and widely disseminated it; and certain organizations pride themselves on hacking into systems and publicly releasing the information.

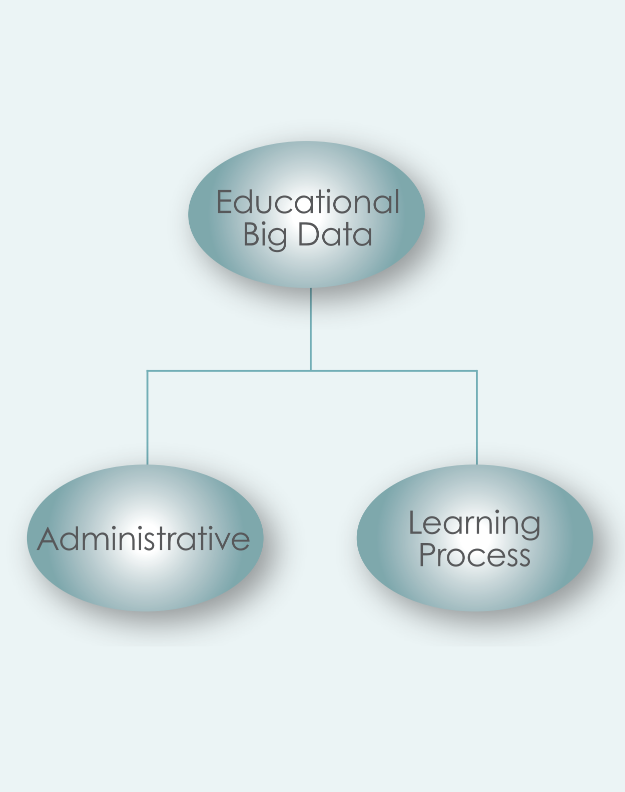
In debates concerning educational data and privacy, many common terms such as privacy, security, and confidentiality are used interchangeably. Additionally, the scope and breadth of educational big data are not always fully understood. Following are definitions of these terms in the context of the discussion, as well as a helpful categorization for using these terms in the education arena.

What Are Big Data?

Big data are extremely large data sets that defy traditional data-processing applications. But what are “extremely large” or “traditional data-processing applications”? In a much-cited article, Laney (2001) describes big data in terms of volume, velocity, and variety.[5] In the education context, this refers to the “numbers of student observations, the frequency of observations, and the number of types of observations, respectively.”[6]

Moreover, in the educational context, big data typically take the form of administrative data and learning process data, with each offering their own promise for educational research as well as raising their own privacy concerns.

The benefits of big data for educational research often arise when data sets are combined and merged. For example, learning process data, combined with administrative data such as demographics and test scores, can provide insights into how to address educational inequities in faster feedback cycles.[7]



Administrative Data

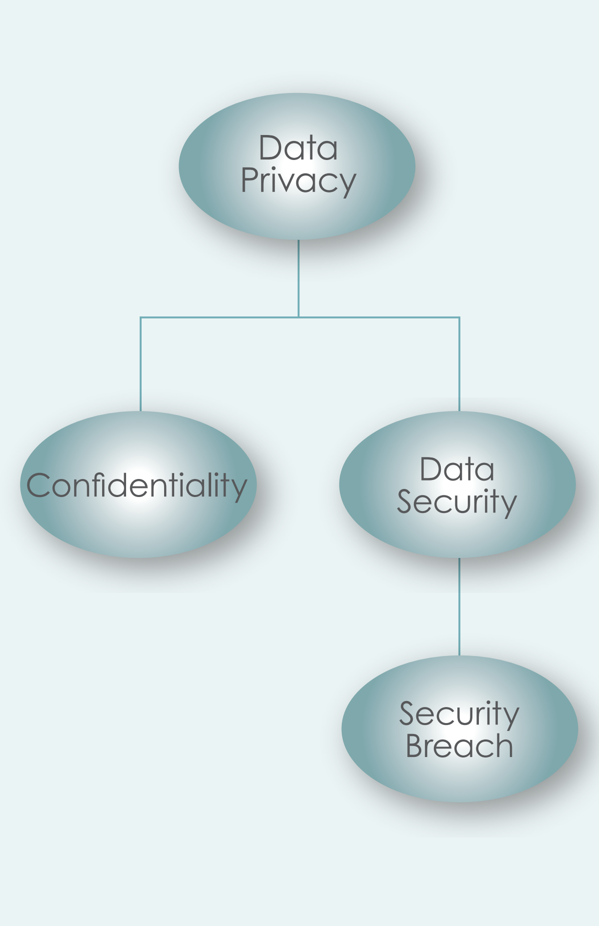
Administrative Data[8] are demographic, behavioral, and achievement data collected through schools, governmental agencies, and their contractors. Administrative data may consist of attendance records, test scores, transcripts, and surveys. Educational administrative data are collected over many participants, often longitudinally at prescribed, regular intervals (e.g., biannually or yearly). Examples of administrative data that are considered “big data” are census data, National Assessment of Educational Progress data, international test scores, state standardized test scores, and behavioral data such as those required to be maintained and collected for the U.S. Department of Education’s Civil Rights Data Collection.

Learning Process Data

Learning Process Data[9] are continuous or near-continuous, fine-grained records, usually of digital interactions of student behaviors to illuminate learning processes.[10] Learning process data are “big” data because they are “tall” (include many participants); “wide” (include a large number of variables about any one individual); “fine” (include multiple fine-grained observations taken across small time intervals); and “deep” (theoretically coded in a meaningful way).[11] Examples of learning process data are data collected in online assessments and courses (including massive open online courses [MOOCs]) or keystrokes and time latencies collected for interactive technologies for K-12 students in a school year.

What Is Data Privacy?

In the education context, concerns about data privacy often focus on student information falling into the wrong hands or being used for nefarious purposes. Others besides students and parents may be entitled to data privacy; depending on the nature of the data, this could include teachers, schools, districts, or colleges. Additionally, terms such as confidentiality and data security are often conflated. While confidentiality and security flow from privacy, their scope is dependent on privacy expectations.



DATA PRIVACY. According to the U.S. Department of Health and Human Services' Office for Human Research Protections (OHRP), privacy is "having control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others."[12] Data privacy, also called information privacy, addresses the ability of an organization or individual to determine what data about them in a computer system can be shared.

CONFIDENTIALITY. Once information is shared, 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."[13]

DATA SECURITY. It then follows that data security is essential to privacy as a means of "preventing unauthorized access to data and includes standards that can be followed to maintain proper access to data." [14]

SECURITY BREACH. A security breach is the loss, theft, or other unauthorized access to data containing sensitive personal information that results in the potential compromise of the confidentiality of the data. A security breach that violates existing laws can be intentional and unintentional.

While unlawful security breaches are considered rare, a data incident that does not rise to the level of an illegal breach can be as harmful as an actual breach to the public perception of data collection and storage and consequently research. A data incident can violate security best practices but still be within the law or it can break trust and expectations. The public often does not distinguish between an unlawful breach and one that violates their trust.[15]

Stakeholders, including those who have a privacy interest, and data collectors and users, including states, districts, schools, vendors, and researchers, must work together to ensure the appropriate sharing of information to benefit educational outcomes and opportunities. To accomplish this they must work together to form a "social contract" with "negotiated norms of what information should be used and how" within the educational community.[16]

What Laws and Policies Govern Data Privacy and Educational Research?

Today, there are numerous places that a school, college, or researcher must look to for guidance on how to ensure that quality research using big data also maintains data privacy. There are several federal laws, varying state laws, Institutional Review Boards (IRBs; both at research institutions and partnering organizations), and local norms to address. Additionally, with new legislation being introduced and new forms of data created, it is an ever-changing landscape.

On the federal level, there are three key laws protecting student privacy.

- **Family Educational Rights and Privacy Act (FERPA),** the primary federal law, applies to all schools receiving federal funding and generally prohibits the disclosure of personally identifiable information from education records with exceptions for organizations conducting certain studies on behalf of the school.[17]
- **The Children's Online Privacy Protection Act (COPPA)** protects children's online privacy by largely placing parents and guardians in control of what information is collected from their children online.[18]
- **Protection of Pupil Rights Amendment (PPRA)** governs the administration to students of a survey, analysis, or evaluation that concerns one or more of eight protected areas, including political affiliations or beliefs, mental or psychological problems, sex behavior or attitudes, religious practices, and income. It requires, among other things, that local educational agencies provide information to parents about such surveys and the opportunity to opt their children out.[19]

There also have been federal attempts to further legislate the uses and availability of student data for research purposes. For example, members of the U.S. House of Representatives have called to amend FERPA to prohibit research that is not aimed at improving the instruction or testing of the specific students involved in the research, require parents to "opt-in" to all research, and require the deletion of student records when a student leaves an institution.[20]

Many states have state laws that address student data privacy. Initially, many such laws codified a version of FERPA. States, however, continue to expand on their original student data privacy laws. Since 2013, 49 states and the District of Columbia have introduced 410 bills addressing student data privacy,[21] and 36 states have passed 85 new education data privacy laws.[22] Additionally, since 2014, 19 states have passed laws that include at least one provision targeted at researchers.[23] While some of these recent laws have fairly reasonable requirements concerning the governance structures for the collection and storage of data, as well as transparency provisions aimed at ensuring better communications with parents, others set restrictions on the format of data that can be made available to researchers (e.g., in aggregated form only), insist on parental permission for any study using data (even data previously collected), or ban the collection of certain types of data (e.g., biometric data) or certain data uses (e.g., predictive analytics).[24] Penalties for privacy violations were also added to some of these new bills to enhance the accountability of data users, and in some cases researchers specifically.

In addition to state and federal laws, **Institutional Review Boards (IRBs)** ensure that researchers whose projects involve human subjects provide the appropriate protections and rights to their subjects, as well as get proper training for compliance purposes. IRBs are governed by the U.S. Department of Health and Human Services' OHRP and the Food and Drug Administration; the basic provisions are laid out by the Federal Policy for the Protection of Human Subjects, also known as the "Common Rule."[25] Typically, both the researchers and the school or institution must present information to their respective IRBs in order to initiate research or data sharing involving students. Given the nature of these new data, as well as the numerous data collectors (e.g., schools, states, private vendors, and researchers), it is important to engage in conversations about the possibly evolving roles of IRBs in addressing these data.

COMMUNICATING THE IMPORTANCE OF EDUCATIONAL RESEARCH

Although the importance of research is clear to researchers, it is not always evident to those sharing their personal information, to parents, and to other stakeholders. Moreover, privacy concerns by parents, students, and advocates about data collection are often not about the use of research specifically but include concerns such as the public release of their data or data used for academic tracking purposes (both short and long term).

The research community, in partnership with key stakeholders, must work to inform those sharing their personal information of the possibilities that their sharing provides, for them and others. High-quality research provides the evidence for practitioners and policy makers to make educational choices to help all children succeed. Big data allow researchers to identify interventions most promising for individual students, including those in high-risk populations. Education research is used for all types of classroom interventions as well as to support larger initiatives such as school lunch programs and early childhood education. Education researchers, however, need to better promote the uses of research to the nonresearch community. Persons are more likely to share their personal information when they see prior positive results and are told of the current uses of their data.[26]

CUPS

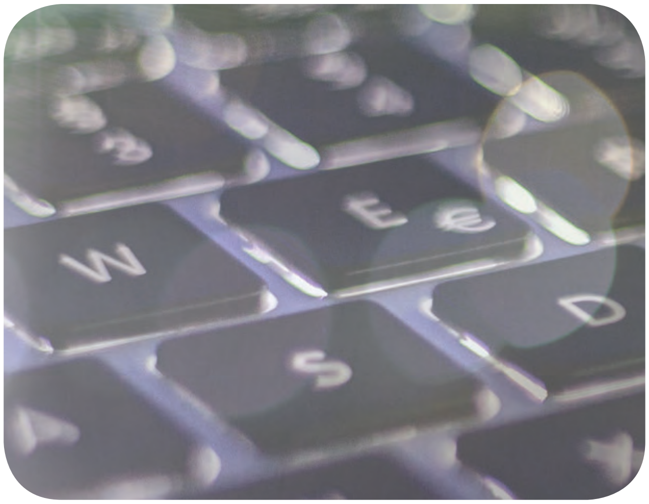
One useful framework for communicating a research project uses the acronym CUPS

COLLECTION	What is collected, by whom, and from whom?
USE	How will the data be used? What is the purpose of the research?
PROTECTION	What are the security protections for the data and how will access to the data be restricted?
SHARING	How and with whom will the results of the data be shared? Will the data be shared for other purposes?[27]

Source: Siegl, J. (2016, August). *Lessons Learned from Education Stakeholders* panel. Workshop on Big Data in Education: Balancing Research Needs and Student Privacy, Washington, DC. See also, Fitzgerald, B. (2017). *Lessons Learned from Education Stakeholders: Panel Summary.* Washington, DC: National Academy of Education.

Although researchers include such information in their research proposals, IRB proposals, and contracts with data-sharing entities, it would be helpful to also have this information available, in user-friendly language, for teachers, parents, and students. Depending on how the research is conducted, states, districts or schools could have the information on hand and posted on their websites, or the researcher can place it on his or her website. Such intentional transparency will be useful in allaying the privacy concerns of those sharing their information and in mobilizing them to action around the improvement of education.[28] Finally, researchers' voices need to be heard when laws and regulations addressing data use and collection are considered. Researchers must be present to communicate the importance of education research, the necessity of the data to carry it out, and the protections in place.[29]

ADVANCING HIGH-QUALITY EDUCATION & PROTECTING PRIVACY



The purpose of educational research is to understand and ultimately improve teaching and learning. It is important to ensure that privacy concerns are fully addressed, while collecting and analyzing data needed to inform practices and policy.

To ensure that personally identifiable data are not revealed, the data can be deidentified, and also aggregated to remove small subpopulations whose unique combination of data points, even without personal identifiers, may be combined with other data to identify individuals. Although some research is possible with such data, often such stripped data cannot replicate original findings and findings cannot be contextualized. Such data sets cannot be used to link administrative data to learning process data, precluding valuable analyses that can link learning processes with long-run outcomes.

Examples of Vetted Data Access

Some options for accessing useful data in privacy-vetted contexts exist for administrative data, particularly such data collected at the federal level. Deidentified and aggregated data are publicly available at numerous government websites. The U.S. Census Bureau has established 24 Federal Statistical Research Data Centers throughout the country that provide secure access to a range of federal statistical restricted-use microdata from numerous federal agencies to qualified researchers.[30] Some states provide similar access.

Virtual Private Networks (VPNs) are used by some federal agencies to allow their employees to access data remotely (e.g., to securely access an agency's private network via a public network such as the Internet while working at home). Additionally, other countries use such technology for researcher access.[31] Similarly, Databrary, a digital library, provides a platform for researchers to share and access video data. Some of the data, in deidentified form, are available to the public, and identifiable video data and other contextual information are available to authorized researchers.[32] Although the nonpublic portions of Databrary require informed consent and promises to best protect the identity of participants, it is not anonymous or deidentified.

Data infrastructures such as LearnSphere, sponsored by the National Science Foundation, are working to facilitate the storage and access of learning process data. LearnSphere builds on DataShop, which is the largest open repository of transactional data.[33] While the storage and sharing of deidentified data are a step forward, it does not allow for some of the contextual research necessary to address educational inequities, nor can it allow for the linking of administrative data.

The Data Continuum

There is a tradeoff between the fidelity of the data and the potential risks to privacy when sharing data. The more personally identifiable information enables greater generalizability and more accurate statistical analyzes; however, it increases the risks of reidentification as well as greater possible harm to individuals if the data are breached. As such, it is necessary for the research community to determine how to ensure that such data can be shared in non-deidentified forms while minimizing potential security breaches.

One approach would be to create a continuum in which one extreme is deidentified or aggregated data that are publicly available, and the other extreme is individual identified data that are highly protected, such as through the use of Data Centers, memoranda of understandings between data collectors/repositories and institutions, and VPN access.

It would be necessary to reach agreement among researchers, schools, parents, policy makers, and private companies to determine the parameters of the data access and use.

Click stream data from a MOOC, for instance, may be deidentified and shared with the public if they are not tied to demographic, academic record, or discussion board data (e.g., no mention of gender, age, race, prior or current classes, or number of posts on information in discussion board posts). These “missing” data, though, are crucial for researchers to provide context to any findings, as well as to generalize and link data across data sets. How to securely share data that contain possibly identifiable information is necessary to the advancement of teaching and learning. Using shared databases for deidentified data, common IRB processes for some data, and data centers or VPN access for other data sets with more personal information is one way to proceed.

> “How to securely share data that contain possibly identifiable information is necessary to the advancement of teaching and learning.”

BUILDING PARTNERSHIPS

Researchers need to build partnerships around the collection, use, protection, and storage of big data. Detailed, linked databases are the most promising option to date for understanding and addressing growing educational disparities; these require partnerships with researchers, states, school districts, parents, private-sector industries, and policy makers.

Such collaborations require trust and mutual interests. The mutual interest is often found in the improvement of educational opportunities, but at times the minutiae may not make this so obvious. As important, collaboration requires a network of trust for researchers to collaborate with each other as well as with others.[34]

The proliferation of the high-stakes uses of standardized tests, from in-grade retention to teacher performance evaluations, likely helped drive the movement to restrict the uses of student data and demand greater privacy protections. This is playing out in state legislatures around the country.[35] Although the use of education data for research is different than the use of education data for in-grade retention, these distinctions are not as easily identified in the eyes of wary parents, teachers, and other stakeholders. As noted above, researchers need to be key players in identifying the benefits of research and must also be transparent in their processes and uses of data. Using CUPS is important to gaining trust, especially from parents, advocates, and teachers.

Research–practitioner partnerships are a useful way to ensure that important data are gathered, shared, and analyzed to support learning in the partnered environment. These partnerships are encouraged by organizations such as the Institute for Education Science (IES) and the Spencer Foundation, which are funding research–practitioner partnerships.[36]

Moreover, with the proliferation of digital learning tools and interactive technology (including MOOCs and personalized learning systems), partnerships between researchers and technology developers and vendors can potentially increase the value of the data collected. Under this model, researchers can iteratively adjust their research questions in collaboration with developers, who can then improve and adjust the kinds of data collected.

Educational agencies also need to be involved in these partnerships. They are typically in the best position to coordinate best practices and foster trust between the partners. They need to partner with researchers to make sure that students, parents, and teachers understand the collection, use, protection, and sharing of the research.[37]

Additionally, there are numerous stakeholders presenting information on data privacy. For instance, 41 organizations came together and drafted 10 foundational principles for using and safeguarding students' personal data.[38] The research community, however, was absent. Such opportunities for engagement should not be lost.

Finally, the involvement of policy makers as partners helps to strengthen the use of research in classrooms. At all levels of the government, partnering with policy makers helps to ensure that policy makers are invested in making evidence-informed decisions. For instance, building relationships at state educational agencies when using such data, keeping them informed of preliminary findings, and sharing understandable conclusions can assist in building trust for current and future projects, as well as to inform decisions. Similarly, such partnerships at the local level assist superintendents, principals, and teachers as they make instructional decisions. These partnerships help to build trust, which is foundational in using particular research in decision making.

PREPARING RESEARCHERS

In this ever-changing data environment, researchers should continue to educate themselves and others on how to use and protect data. The methodologies to protect the privacy concerns continue to advance, as do the potentials for hacking and misuse. According to panelist Amelia Vance, "no college of teacher education has a course on privacy."

The educational research community needs to hold itself and all of its members to standards for the knowledge they possess, including the terminology, laws (both federal and state as applicable), IRB requirements, and privacy concerns surrounding data collections. While it is critical that educational researchers understand their technical requirements, it is also necessary for them to have access to and make use of staff trained in data security. Universities require levels of security, such as secure networks, encryption, and deidentification. Researchers must know *what they do not know*, and make sure that they are working with information technology professionals who do know it.

Researchers should know:

- How to use and protect data
- Current laws and policies governing privacy
- Privacy concerns of students, parents, and stakeholders
- How to address privacy concerns
- How to communicate the benefits of research
- What they don't know, so that they can ask for help

In addition to comporting with the letter of the law, researchers must be trained on the importance of transparency and communication with all stakeholders. In addition to the legal requirements, there are ethical requirements to the use of data. Researchers must understand the privacy concerns of students, parents, teachers, school districts, and advocates and, as noted above, should be able to address their concerns, explain how the data will be protected, and describe the benefits of their and other educational research.

And importantly, as data become more fine grained, as the links between administrative data and learning data are forged, and as adaptive technologies are more utilized, researchers using such data must be continually trained and educated on advanced technology and methodologies to effectively process big data. It is through this continuing education of researchers that such data will be best used to advance teaching and learning.

NEXT STEPS

Research using big data can enhance understandings of student learning and how to close achievement gaps. When concerns about student data privacy are raised, the research community has not been highly visible in response. This workshop, as well as other such gatherings, is a first step in having the research community better understand the privacy concerns and work to address them. The following are a summary of next steps.

Adopt Common Terminology. Throughout this workshop, it was proposed that common terminology be adopted to ensure that those engaged in discussions about educational research, big data, and privacy are "speaking the same language." By understanding and adopting common terminology, researchers will be better positioned to ensure that stakeholders' concerns are being addressed.

Communicate the Importance of Educational Research. Education researchers need to communicate the importance of education research, both with respect to the stakeholders involved in the research as well as more broadly. Education researchers need to be prepared to present evidence as state and federal legislators, as well as local school boards, look to expand student privacy laws and regulations. The research community needs to stand prepared to present evidence of the benefits of educational research and the existing safeguards to student privacy. No information was presented at the workshop to encourage more regulation; the regulation in place, particularly in the form of FERPA, is sufficient to protect interests, and the research community must continue to work vigorously to comply with it. Education researchers must be prepared to discuss the precautions taken to protect the privacy of the data and, importantly, the benefits flowing from research.

Build Strong Partnerships and Models to Ensure the Sharing of Data. The research community, in partnership with stakeholders, governments, and private entities, needs to collaborate to ensure that data can be combined in important and useful ways. It is with the sharing and linking of administrative and learning process data that learning and teaching can be best enhanced. These stakeholders must come together to determine ways to provide access to the personally identifiable data needed to link such data sets, while ensuring that the data are not breached. Expanding the opportunities for LearnSphere and data centers will likely be crucial to this endeavor. It is also important for researchers to identify successful research and private partnerships to emulate, along with critical documents explaining the nature and scope of the partnerships and the measures taken to address privacy concerns. The research community needs to be engaged in creating a continuum of data sharing, with non–personally identifiable data being widely available and sensitive data available in a more protected setting.

Continue to Educate Researchers and Universities on Privacy Issues. While reaching out to the broader community to form partnerships and communicate the importance of educational research, simultaneously researchers must be properly prepared. Researchers must comply with numerous privacy requirements as well as be communication agents within communities. First, the educational research community could agree on model IRB procedures and informed-consent models to present to universities which address the needs of educational big data. Additionally, developing guidelines for researchers concerning both best security practices and engagement with stakeholders would help ensure that researchers understand the privacy concerns as well as how to best communicate the protections being taken.

ENDNOTES

[1] Ujifusa, A. (2014). State Lawmakers Ramp Up Attention to Data Privacy. *Education Week*. Available at http://www.edweek.org/ew/articles/2014/04/16/28data.h33.html?qs=data+privacy.

[2] Vance, A. (2016). *Panel Handout: Lessons Learned from Education Stakeholders*. National Association of State Boards of Education. ("Vance Handout").

[3] Vance, A. (2016). Trends in Student Data Privacy Bills in 2016. *Policy Update, National Association of State Boards of Education*, Vol. 23, No. 13. Available at http://www.nasbe.org/wp-content/uploads/Vance_2016-State-Final.pdf.

[4] See National Academy of Education. (2013). *Adaptive Educational Technologies: Tools for Learning, and for Learning About Learning*, G. Natriello (Ed.). Washington, DC: Author (providing examples of educational research using big data from adaptive educational technologies).

[5] 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.

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

[7] See, e.g., Perry, J. & Klopfer, E. (2014). UbiqBio: Adoptions and Outcomes of Mobile Biology Games in the Ecology of School. *Computers in the Schools*, Vol. 31, pp. 43-64; O'Rourke, E., Chen, Y., Ballweber, C., & Popovic, Z. (2016). *Personalized Learning and Its Behavioral Impact on the Classroom Ecosystem* (in submission).

[8] See Figlio, D. (2017). *Rule of Administration Data in Education Research: Panel Summary*. Washington, DC: National Academy of Education ("Panel 1 Summary").

[9] See Steinkuehler, C. (2017). *Learning Process Data in Education Research: Panel Summary*. Washington, DC: National Academy of Education ("Panel 2 Summary").

[10] Ho Workshop Paper.

[11] See Panel 2 Summary.

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

[13] The IRB Guidebook.

[14] Bienkowski, M. (2017). *Implications of Privacy Concerns for Using Student Data for Research: Panel Summary*. Washington, DC: National Academy of Education ("Panel 5 Summary").

[15] Fitzgerald, B. (2017). *Lessons Learned from Education Stakeholders: Panel Summary*. Washington, DC: National Academy of Education ("Panel 4 Summary").

[16] See Panel 5 Summary (quoting Kirsten Martin).

[17] U.S. Department of Education. (2016). Family Education Rights and Privacy Act (FERPA). Available at http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

[18] Federal Trade Commission. (2016). Complying with COPPA: Frequently Asked Questions. Available at https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions.

[19] Family Policy Compliance Office. (2016). Protection of Pupil Rights Amendment (PPRA). Available at http://familypolicy.ed.gov/ppra.

[20] Vance Handout.

[21] Data Quality Campaign. (2016a). Student Data Privacy Legislation: A Summary of 2016 State Legislation. Available at http://2pido73em67o3eytaq1cp8au.wpengine.netdna-cdn.com/wp-content/uploads/2016/09/DQC-Legislative-summary-09302016.pdf.

[22] Vance Handout.

[23] Vance Handout.

[24] See Panel 4 Summary and Data Quality Campaign (2016a).

[25] Federal Policy for the Protection of Human Subjects ("Common Rule"). Available at https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

[26] Data Quality Campaign. (2016b). Turning Data into Information: The Vital Role of Education Research in Improving Education. Available at http://2pido73em67o3eytaq1cp8au.wpengine.netdna-cdn.com/wp-content/uploads/2016/08/DQC-Why-research-matters-09142016.pdf.

[27] See Panel 4 Summary (citing Jim Siegl).

[28] See Panel 4 Summary.

[29] See Berman, A. (2017). *Lessons Learned from Other Fields: Panel Summary*. Washington, DC: National Academy of Education ("Panel 3 Summary").

[30] The 24 Federal Statistical Research Data Centers are listed on the U.S. Census Bureau website. Available at http://www.census.gov/about/adrm/fsrdc/about.html.

[31] See Panel 3 Summary.

[32] See the Databrary website: https://nyu.databrary.org.

[33] See the LearnSphere website: http://learnsphere.org.

[34] See Panel 5 Summary.

[35] See Panel 4 Summary.

[36] See Ho Workshop Paper for examples of IES-funded partnerships, and for Spencer Foundation projects see http://www.spencer.org/research-practice-partnership-program-statement.

[37] See Panel 5 Summary.

[38] Student Data Principles: 10 Foundational Principles for Using and Safeguarding Students' Performance Information. Available at http://studentdataprinciples.org/the-principles.

Steering Committee

Susan Fuhrman (Co-Chair)
Teachers College, Columbia University

Elizabeth Buchanan
University of Wisconsin–Stout

Chris Dede
Harvard Graduate School of Education

Louis Gomez
University of California, Los Angeles

P. David Pearson (Co-Chair)
University of California, Berkeley

Andrew Ho
Harvard Graduate School of Education

Sophia Rabe-Hesketh
University of California, Berkeley

National Academy of Education Staff Amy Berman, Deputy Director

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

Printed in the United States of America

Suggested citation: National Academy of Education. (2017).
Big Data in Education: Balancing the Benefits of Educational Research and Student Privacy: Workshop Summary. Washington, DC: National Academy of Education.

The National Academy of Education (NAEd) held a two-day workshop on August 9-10, 2016, to address a fundamental tension faced by the education research community: how to balance the benefits of access to comprehensive (“big”) data with the potential risks to privacy. The NAEd website contains a commissioned background paper, summaries of each workshop panel, handouts and presentations, and video recordings. This publication synthesizes the significant points raised and provides a blueprint for future action.

For more information visit: naeducation.org/bigdata

The project was funded by the Ewing Marion Kauffman Foundation.
The contents of this publication are solely the responsibility of the NAEd.

NATIONAL ACADEMY *of* EDUCATION

The National Academy of Education (NAEd) advances high-quality research to improve education policy and practice. Founded in 1965, the NAEd consists of U.S. members and foreign associates who are elected on the basis of outstanding scholarship related to education. The NAEd undertakes research studies to address pressing issues in education and administers professional development programs to enhance the preparation of the next generation of education scholars.