Workshop on Big Data in Education

Balancing the Benefits of Educational Research and Student Privacy

Lessons Learned from Other Fields: Panel Summary

Amy Berman, *National Academy of Education*

NATIONAL ACADEMY of EDUCATION

Lessons Learned from Other Fields
Panel Summary

Amy Berman, National Academy 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: Berman, A. (2017). *Lessons Learned from Other Fields: Panel Summary*. Washington, DC: National Academy of Education.

Lessons Learned from Other Fields: Panel Summary

Amy Berman

Panel:

Amy Berman, National Academy of Education (Panel Moderator)
Stanley Crosley, Drinker Biddle & Reath LLP and Indiana University Center for Law, Ethics, and Applied Research (CLEAR) in Health Information
John Friedman, Brown University
Brian Harris-Kojetin, Committee on National Statistics, National Academies of Sciences, Engineering, and Medicine
Camilla Stoltenberg, Norwegian Institute of Public Health

INTRODUCTION

Many fields, in addition to education, have experienced significant transformations in the gathering and use of large-scale data sets. Like education researchers, researchers in other fields are balancing the benefits of ensuring that large data sets are used to make evidence-informed policy decisions with privacy concerns of individuals and institutions. This panel gathered experts who use big data in noneducational contexts—including health, census, tax, labor, and retirement—to learn from their experiences in striking this balance.

DIFFICULT TO ACCESS AND LINK DATA FOR RESEARCH ANALYSIS

Currently, U.S. federal agencies collect and acquire a wealth of practical and useful administrative data; however, accessing and linking the data to ensure better informed policy making provide significant challenges for researchers, as well as government employees. Federal agencies do not have uniform policies to share data among agencies and in some instances are banned by law from doing so. Similarly, standard practices do not exist for sharing with nongovernment researchers.

As Harris-Kojetin discussed, about 12 federal agencies have primary responsibilities to gather statistical information, typically in the form of surveys and censuses. Many though have fragmented ways for researchers to gain access to some of the data. For instance, the U.S. Department of Education's National Center for Education Statistics provides some online analysis tools, allowing statistical output data to be exposed. Individual agencies provide for limited access to data based on licensing agreements with institutions. The U.S. Census Bureau provides researchers access to federal census microdata in secure locations. Today there are 23 Federal Statistical Research Data Centers (FSRDCs) that provide researchers access to Census data as well as data from other agencies such as the U.S. Department of Health and Human Services and the Bureau of Labor Statistics. Although this is a step in the right direction, the FSRDCs are a closed system and consequently researchers can access the data at only the FSRDC locations and are prevented from linking data sets not within the system. Another research roadblock arises from the masking and removal of personally identifiable information in many of these data sets, with limited exceptions such as when researchers are permitted to bring in outside data and merge it with the existing data. In addition to having to obtain approval to access the data, researchers also need approval to remove the data, even the regression coefficients, which is only accomplished after human review. This not only limits the utility of the data, but also significantly increases the time before the resulting analyses are available. As Friedman noted, this cordoning off of the data makes the normal iterative process of research difficult to accomplish and significantly limits collaboration. While shared data in the FSRDC is useful, the continued fragmentation of the data sets, the inability to link certain data sets as well as external data sets, the need to be onsite, and the difficulties collaborating continue to create obstacles in research data analysis and restrain the flexibility of many research projects.

Some agencies, including the U.S. Internal Revenue Service and the U.S. Census Bureau, also provide public access to synthetic or masked data sets (i.e., structurally similar data but not the actual data) so that privacy is protected and individual, personally identifiable data cannot be recovered. As Friedman pointed out, while such data sets are useful for various summary statistics, they typically do not work for quasi-experimental identification because the masking often alters the original data in a fundamental way that prevents the examination of

specific variation. As Harris-Kojetin noted, some agencies such as the U.S. Census Bureau allow researchers, after using the synthetic data set, to request that the agency run its analysis on the actual data to ensure the fidelity of the results. This can be a time consuming and cumbersome process, and even with such accommodations, masked data sets continue to pose problems for quasi-experimental work given that true variation in the actual data may never be identified.

In contrast to the somewhat fragmented U.S. system, other countries, including the Nordic countries, are forerunners in data access in terms of their broader availability of administrative data and better integrated data infrastructure. Stoltenberg described the national registries of countries such as Norway, which, through the use of personal identification numbers, track every individual on numerous data points throughout their lives. While the data are thin (i.e., numerous variables but not extensive), the data are comprehensive (i.e., everyone in the country is included) and permits—while not always in the most efficient manner—the linking of other data sets to provide rich longitudinal data. Additionally, in Norway, there are extensive longitudinal cohorts imbedded within the registry focusing on more specific issues. One such striking example is the cohort focusing on Norwegian children and their parents with more than 300,000 participants, including biological samples, survey data, and clinical observations that can also be combined with randomized trials.

The data in such countries are often easily obtainable and unlike the burdensome and fragmented system in the United States, researchers can far more easily simultaneously access data across different categories (e.g., education and health). Moreover, as Friedman described, instead of a closed data system, these countries provide Virtual Private Networks (VPNs) that permit secure access to the data stored on private systems across a public network, allowing researchers to access the data at their institutions but not maintain it on their personal or institutional systems. This approach protects privacy without creating some of the research barriers found in the United States. Additionally, researchers can access aggregated results such as regression coefficients promptly (e.g., within minutes) through an automated review that is subject to a subsequent human audit. This is in sharp contrast to the human review required in the United States. As Friedman noted, such abilities are driving researchers in education- and noneducation-related fields to use data from European countries to answer their research questions because the U.S. data are either unavailable or very difficult to access and link to other data.

These research obstacles that result from the lack of accessible data infrastructure not only have an impact on domestic U.S. policy making, but also have a direct effect on international and global data access, and thus international policy and decision making. Migration patterns, global climate change, international health care concerns (such as global epidemics), and transnational education levels are just some of the areas affected by the shortcomings in data access.

OVER-REGULATION OF RESEARCHER USE OF DATA

One of the fundamental and recurring concerns about allowing access to data sets is rooted in individual privacy considerations. Indeed, securing privacy in research is often stated as a critical goal. However, over-regulation of data access may not be the solution to these privacy concerns. As Crosley noted, privacy breaches are often caused by human error; almost anyone with an email account can create an intentional or unintentional privacy breach. Addressing human error through more regulations, however, may cause more harm than good. Over-reliance on regulations tends to cause collected data to be under-utilized by researchers and

consequently for evidence-based policy making. While researchers are often swept up in regulations addressing privacy concerns, there is little evidence of researcher-related breaches causing harm. While researchers may fail to follow protocols such as locking doors, which they should be admonished for doing, resulting harms have not been substantiated. Stoltenberg stated, "Research is regulated as particularly risky, in my view, for privacy while most of the risk probably lies in other use of the exact same data." As Crosley asserted, however, once researchers are governed by stringent privacy regulations, historically it has been difficult to become less regulated—legislators typically do not want to be the ones fighting for less individual privacy.

In the health field, Crosley described how "privacy" was often the rationale provided for regulating access to data, even when actual meaningful privacy concerns were not implicated. One such example Crosley provided related to the protection of proprietary algorithms. In order to protect the proprietary nature of the information (which was the true goal of the institution or the organization), privacy concerns were used as a shield to data access.

Similarly, the over-reliance of deidentification to comply with laws and regulations harms research and in particular important and necessary longitudinal analysis. Crosley opined that the current overlapping array of health field laws, coupled with strong public sentiment concerned with privacy, is limiting necessary access to important health care data.

In addition to being subject to specific regulations (such as the Health Insurance Portability and Accountability Act of 1996 [HIPAA]), researchers also fall prey to regulations that do not appropriately carve out special data uses for researchers to realize gains from data analytics. For instance, as Harris-Kojetin pointed out, federal statistical agencies typically collect or acquire data under a pledge of confidentiality and for exclusively statistical purposes. However, state laws and regulations similarly impact federal agencies' abilities to collect and use data. Some federally obtained data are collected and owned by states. Consequently, the federal government may have 50 different memoranda of understanding governing the access and use of the data and often they do not provide access for broad statistical purposes.

This desire to legislate concerning privacy, which has led to conflicting regulatory schemes between states, is not limited to between state differences. At the federal level, while the Confidential Information Protection and Statistical Efficiency Act governs all federal agencies, many agencies have additional statutes governing the collection and use of data. As a result, not all federal statistical agencies can share data among themselves.

As noted, whether it be by complete lack of access, closed systems, or additional regulations and requirements, researchers often have significantly less access and flexibility to examine data to which thousands of public employees have access. Even though the current regulatory framework provides that federal employees and their agents—which includes researchers accessing agency data—obtain training, submit to confidentiality pledges, and be subject to significant potential penalties for disclosures, researchers do not obtain similar access as the employees.

Data regulations have created barriers leading to a decentralized statistical system, with every state and federal agency having its own access limitations and agreements. Given the important contributions made by researchers, regulations must ensure security is met while at the same time permit access.

IMPORTANCE OF RESEARCH

For both education and noneducation research, researchers must ensure that policy makers, as well as the citizenry, understand the benefits of research using big data. As Stoletenberg noted, we need to work toward a knowledge system both nationally and globally. To do so, researchers must be prepared to justify the role of research in evidence-based policy making.

As Crosley noted, trust is critical to ensure the efficient and effective exchange of data. If the citizenry has confidence not only that personally identifiable information is being protected, but also that the research is important, access to data can improve. As trust is under assault, researchers must address these concerns with lawmakers and with individuals who share their data. Researchers need to garner trust in order to further research.

Additionally, researchers should support the sharing of data not only for research but also for the operation of government. As Friedman noted, this is often foremost in the minds of lawmakers and agency employees working to open data access. Often data are already collected by one agency but not shared with another that would find them useful. For instance, the U.S. Census Bureau seeks to determine information on new hires in the United States; however, the National Directory of New Hires (NDNH) collects these data for child support and wage garnishment purposes. The NDNH though is not available to the U.S. Census Bureau and thus cannot be used for employment-related evaluations or information. Consequently, researcher and agency interests can be combined to promote better and more efficient access to data.

Steps have already been taken to build this trust and better balance privacy concerns with the need for data access for researchers. The Evidence-Based Policymaking Commission Act of 2016 established a commission charged with determining how to integrate administrative data to make them more available to facilitate research while protecting privacy interests. For federal agencies, the conclusions and guidance of this commission could assist with removing some of the silos. The commission hopefully will provide a forum and an opportunity for researchers to voice their concerns and provide statistical and anecdotal evidence that supports the need for enhanced access to data sets while at the same time providing privacy protections. Researchers need to justify the role of research and political actions are required to lessen the burden of obtaining data access permissions.

PROPOSED RECOMMENDATIONS

Based on their experiences accessing and working with large data sets, the panelists provided numerous recommendations for furthering research and addressing privacy concerns. Their recommendations are aimed both at ensuring that privacy measures are implemented and that researchers have access to important data to further evidence-based policy decision making.

- **Unified statistical data system:** While the panelists agreed that the United States is far from having a unified data system, it is important to continue to work toward making more federal agency data, as well as other data sets, more easily linked and accessible.
- **Integrating advanced technology to improve efficiency and security:** Data privacy and data access would benefit from integrating advanced technology to improve efficiency and security. As discussed above, virtual technology such as VPN has significantly changed how researchers access data and how agencies protect individual data privacy in many European countries. Similarly, some federal agency employees, such as those

at the U.S. Census Bureau, can use VPN access to work from home, while still accessing secure Census data. Such virtual access is suggested widely, but in the shorter term, it would seem that researchers who have been bound by similar agreements or oaths as employees of agencies such as the U.S. Census Bureau could similarly access the data as employees through VPNs. Moreover, instead of human review of all released data and summary information, systematic electronic monitoring with human auditing functions will not only ensure data integrity but also will largely preserve research flexibility.

- **Greater transparency:** Privacy should not be a cover to transparency. It is necessary to have transparency of not only the data but also of algorithms to address a multitude of concerns, including validity and fairness.
- **Expressing the value of research:** Researchers need to be a part of the local, state, and federal conversations about privacy regulations and the uses of data. Researchers need to demonstrate the value of research and the need for data to continue to inform evidence-based policy making.