

## Education Research and Student Privacy

Jane Hannaway<sup>1</sup>

McCourt School of Public Policy

Georgetown University

Testimony before the Committee on Education and the Workforce

United States House of Representatives

March 22, 2016

Chairman Kline and Ranking Member Scott, and Members of the Committee. Thank you for inviting me to appear today to discuss education research and student privacy concerns.

I would like to make two main points in my comments. First, large-scale education data, especially individual level administrative data, make possible important new insights into policies and practices that promote student learning and longer term educational and employment outcomes. Second, provisions that protect the confidentiality of data about individual students are essential. With appropriate safeguards, I see no necessarily inherent conflict between research using individual

---

<sup>1</sup> Professor, McCourt School of Public Policy, Georgetown University; Institute Fellow, American Institutes for Research. Founding Director, National Center for Analysis of Longitudinal Data in Education Research (CALDER). The views expressed here are my own; they do not necessarily represent the views of Georgetown University or the American Institutes for Research, its funders or its Board of Directors. Dan Goldhaber and David Figlio provided helpful comments, but all errors are my own.

administrative data and protection of student privacy. Indeed, I would argue that appropriate safeguards foster a healthy environment for research productivity.

My comments are based on over a decade of experience conducting research using individual level longitudinal state administrative data and leading a highly productive national research center<sup>2</sup>, dedicated to working with such data, that includes some of the most accomplished and insightful empirical researchers in the country.

### **Research Advantages of Individual Level Longitudinal Data Systems**

Almost every state has developed an individual student level longitudinal administrative data system. These data systems have substantive and technical research advantages, as well as efficiency virtues.

Let me start with the *efficiency* virtues. Because the data are existing working files – created, maintained and used by the state for administrative purposes – they are readily available for approved research purposes. Researchers are, thus, not required to undertake costly and time-consuming data collection efforts. And, because the data are officially used files, data quality is high. Having data already in hand means the turnaround time for getting feedback on the results of new policies is short, allowing informed decision making about whether to discontinue, modify or continue particular policies and practices. Indeed, some decisions of interest can be made almost in real time.

The administrative files are also ‘census’ files that provide a wide range of possibilities in terms of *substantive* questions that can be addressed. The files include data on all students and all teachers in the state over a number of years. So data on students of interest for a particular intervention or for a particular study, say 8<sup>th</sup> graders, or high performing students, or disadvantaged students can be easily

---

<sup>2</sup> Using large-scale administrative data, CALDER has produced over 150 original studies, most of which have been published in peer-reviewed journals. [www.caldercenter.org](http://www.caldercenter.org)

selected. Similarly, subgroups of teachers, for example - teachers in high and low poverty schools or inexperienced and experienced teachers - can be compared in terms of a number of dimensions, such as their credentials and experience. Indeed, because teachers can be linked in the data to their students and students' test scores, teachers can also be compared in terms of their performance. Indeed, some of the most important findings from studies using longitudinal data have focused on teacher effectiveness.

Because longitudinal data systems extend for long periods of time, researchers can capture: difficult to study populations, such as highly mobile students; long term consequences of program or policy shifts, and the effect of students' past experiences on current performance. For example, an intervention at 8th grade may have relatively small short term academic results, but longer term large effects on, say, high school graduation and college attendance. Without these administrative files, conducting research by tracking students for long term outcomes would be nearly impossible or prohibitively expensive. And long term outcomes are what is of most central concern for education research. Indeed, a recent study showed that, beyond the immediate test performance of students, highly effective teachers have important longer term effects on students, including college attendance and higher earnings, than otherwise similar students with less effective teachers.<sup>3</sup> Such research was inconceivable a few years ago, and our understanding of the real value of great teachers was greatly underestimated.

The longitudinal nature of the data also provide a number of *analytic* advantages that strengthen the credibility of research findings and even allow identification of causal effects without requiring initial random assignment. For example, regression discontinuity designs can assess the effect of, say, receiving an award on subsequent behavior by comparing results for students just above and below the performance award threshold. Natural experiment effects created by policy shifts can also be assessed.

---

<sup>3</sup> R. Chetty, J. Friedman and J. Rockoff. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. **American Economic Review**, 2014.

The advantages in terms of policy insights of individual education data are also substantially expanded when linked to later individual measures in areas beyond education, such as labor market (employment and earnings), justice and health outcomes.

In addition to technical advantages and substantive policy contributions, the availability of individual level administrative education data has had important human capital effects on the research enterprise itself. Some of the most talented and technically well trained researchers in the country have been drawn into education research. The quality and versatility of the data, and the recognized importance of education outcomes for the economic well-being and social fabric of the country, have made education a “hot” area for research. It is rare to pick up a public policy, economics or other social science peer-reviewed journal without seeing the centrality of education to current evidence-based scientific inquiry.

### **Privacy Protections**

Student level data is considered private information protected by both state and federal laws. States also impose filters on requests for data for research purposes. For example, states review whether the proposed research questions are of interest to the state, whether the research plans and the proposed analytic strategies are appropriate, and whether the credentials of the researchers meet standards. All individuals working with the data, including research assistants, must be identified and sign data use agreements along with their home institution. If the data request is approved, the data must be only used by identified researchers and address only the research questions identified in the request. The data must also be destroyed at the end of the study period. In short, researchers use administrative data files under strict constraints, and individual researchers and their institutions are held accountable.

While researchers (including me) might complain about going through the necessary bureaucratic hoops to gain permission to use these valuable large-scale individual data files, I would argue that strong data security provisions are good, especially in the long run, for the field. Researchers want states and other data providers to trust us with the data that we value so highly for our work, and we are willing to do what is necessary to maintain that trust so that we may contribute analytically to the development of well-grounded and meaningful insights for education policy and practice.

I suspect that some of the public assumes that data on individuals in administrative education data files is easily available to clever data analysts -- for example, the kind who know how to target me after I shop for shoes on the internet. This is not the case. These administrative data files are highly restricted in terms of allowable users, data items, terms of use, and period of use. In fact, the data that researchers receive contain no individually identifiable information. The state anonymizes the data before researchers receive them. Each student is assigned a state-constructed unique student id (USI) that is used by researchers to link data for each student across years and schools. Identifying information, such as name, birthdate and social security number are stripped from the files before researchers receive them.

Unfounded public fears of disclosure of individual student level information, researchers worry, may lead to data restrictions that result in harmful effects for the research enterprise and, as a consequence, for the informed development of educational policy and practice. "Opting-out" provisions are a particular area of concern. Consider this theoretical possibility: Middle class parents, who consider the test scores of their adolescent sons to be low, attempt to restrict information on their sons in administrative data files. Excluding them from the data would result in a biased data set, one that underrepresents low performing middle class adolescent males and thereby limits our understanding of how policies and practices may (or may not) benefit those students.

To the best of my knowledge - as someone who has personally worked with these large individual level data sets, led a center of researchers focused on analyzing these data for over a decade, been in editorial positions for major journals and leadership positions in research associations - there has never been a violation of confidentiality by any researcher.

It is important that policy makers and the public understand the restricted conditions under which researchers work and the tremendous value these data hold for examining the effects education policies. I believe the research community is willing to work with policymakers to satisfy data confidentiality concerns but do not compromise the contribution of research to important national education efforts.