Workshop on Big Data in Education

Balancing the Benefits of Educational Research and Student Privacy

Role of Administrative and Survey Data in Education Research: Panel Summary

David Figlio, *Northwestern University*

NATIONAL
ACADEMY
of
EDUCATION

Role of Administrative and Survey Data in Education Research
Panel Summary

David Figlio, Northwestern University

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: Figlio, D. (2017). *Role of Administration and Survey Data in Education Research: Panel Summary*. Washington, DC: National Academy of Education.

Role of Administrative and Survey Data in Education Research: Panel Summary

David Figlio, Northwestern University

Panel:
David Figlio, Northwestern University (Panel Chair)
Sophia Rabe-Hesketh, University of California, Berkeley (Panel Moderator)
Chris Chapman, National Center for Education Statistics
Eric Hanushek, Stanford University
Larry Hedges, Northwestern University

INTRODUCTION

The broad purpose of this panel was to discuss the role of administrative and survey data in education research, to evaluate the benefits and costs associated with the application of these data, to touch on the ways privacy concerns and political factors might influence use of these data for research purposes, and to summarize the current federal efforts by the National Center for Education Statistics to integrate administrative records with survey data.

David Figlio, Director of the Institute for Policy Research and Orrington Lunt Professor of Education and Social Policy at Northwestern University, chaired the panel and led with an overview of the uses of administrative data in education research. While Figlio stipulated that administrative data are not a panacea, he emphasized the benefits of using population-level data for research. Eric Hanushek, Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University, followed with a discussion of the often poorly aligned objectives of researchers and politicians. Hanushek described how researchers and politicians often see different benefits and risks associated with the use of administrative data for research, and he listed a set of impediments to data access. Larry Hedges, Board of Trustees Professor of Statistics and Institute for Policy Research Q-Center Director at Northwestern University, continued the discussion by identifying the uses for which big data are particularly well suited, the uses for which they may be well suited, and the tasks for which big data are poorly suited. Hedges then presented case studies illustrating the degree to which disclosure control and information preservation are in tension. Chris Chapman, Associate Commissioner for the Sample Surveys Division of the National Center for Education Statistics (NCES), rounded out the panel by speaking about NCES's efforts to integrate administrative records and sample survey data. He discussed recent policy initiatives aimed at improving the use of official data, the administrative data currently being collected by the U.S. Department of Education, and the integration of nonfederal data from states and localities as well as private organizations such as the National Student Clearinghouse.

The panel focused on the uses of administrative and survey data and the different challenges faced by researchers and policy makers. Below is an outline of the key strengths and disadvantages of administrative data that emerged, the problems faced when protecting against disclosure (i.e., using deidentified data), and the tensions between researchers and policy makers surrounding the use of administrative data.

STRENGTHS OF ADMINISTRATIVE DATA

Support of Designed Studies

Administrative data can be instrumental in supporting the development and implementation of designed studies—surveys and field experiments. Figlio and Hedges both pointed out that administrative data can dramatically improve the efficiency of surveys and field experiments alike by reducing the tracking and follow-up costs. Administrative data often can provide important information that allows these studies to continue for longer or in richer dimensions than could ordinarily occur without considerable expense. Chapman described some of the ways the NCES is using administrative data, both nonfederal and federal alike, in order to enrich and deepen NCES's survey data collections. Nonfederal data examples include college entrance exam data from the ACT and SAT, alternative high school completion data from the GED, college attendance and completion data from the National Student Clearinghouse, and information about nonfederal financial aid. Federal data examples currently in the works include working with the Social Security Administration to obtain earnings and employment history through the Master Earnings File and with the Office of Veterans Affairs to match applications for veteran benefits. And, of course, NCES data collections are enhanced by merging administrative data collected from the U.S. Department of Education through the Office of Federal Student Aid, including information about grants, loan applications, and loan history, as well as data from the Office for Civil Rights. Indeed, as Figlio pointed out, administrative data can not only save scarce resources, but could in some cases improve data quality by reducing the likelihood for attrition or nonresponse problems frequently endemic in follow-up studies, as well as decreasing the risk of recall bias common when asking respondents to report retrospectively regarding past program participation or experiences. Administrative data often provide novel types of variables (e.g., measurement of delinquency, health instances, social networks, and changing geographies) that are often very difficult to study through other data collection methods.

Administrative data can assist the design and interpretation of designed studies in other ways as well. Chapman pointed out the central role of administrative data regularly collected from schools in providing the sampling frames used in NCES data collections. Hedges described how administrative data can support both the design of targeted studies, helping to ensure that the studies have high external validity from their inception, as well as improving the ability of designed studies to be analyzed with external validity in mind.

Direct Descriptions of Populations

Relatedly, administrative data collected at the population-level means that researchers and policy makers can directly describe the characteristics of populations. Moreover, as Hedges pointed out, population-level data allow for direct descriptive comparisons of inputs and outcomes across different geographies and subpopulations, including relatively rare groups. This ability for deep description and meaningful characterization of subpopulations can help to shine light on important inequities or differences, and can also help developers of designed studies to validate the degree to which their samples are representative of broader populations.

Facilitate New Research Questions and More Credible Approaches to Previously Studied Questions

Another key theme involved ways administrative data could facilitate new and important research questions that would be difficult to study otherwise. Although analyses with administrative data will rarely have the internal validity that designed random assignment experiments would have, they do provide many new opportunities. Both Figlio and Hedges cited how the real-time nature of administrative data allows the study of very recent events and of situations where there may be continuous process improvement. Meanwhile, the archival nature of administrative data permits the study of events that are in the distant past, as well as facilitating the ability to conduct intergenerational studies. Whereas some educational questions can be well studied using designed studies, many others would be difficult to study because the introduction of policies and practices or changes in the educational environment are often not predictable. Figlio pointed out that administrative data allow researchers to study so-called "natural experiments" that would be impossible to anticipate and study prospectively. The population-level nature of administrative data is also ideal for detecting rare events that might be useful for causal identification or descriptive analysis alike, and is also excellent for facilitating the study of heterogeneous effects of educational policies and practices. It is impossible, on the other hand, to carry out evaluations with the same degree of internal validity using even the best-designed survey data, in many cases. And both Hedges and Figlio discussed how administrative data can help researchers to better identify structural parameters of human behavior that are useful both in making predictions about the effects of future policies and interventions as well as in the design of complex statistical studies.

CHALLENGES ASSOCIATED WITH ADMINISTRATIVE DATA

Need for a Large Data Set

Along with the strengths of the use of administrative data, the panelists identified some of the weaknesses associated with using administrative data for research. For instance, Hedges described how one needs a "surprisingly" large data set to ensure better inference than a probability sample would yield, a situation that is exacerbated by the fact that the precision of the construct measured in administrative data is quite likely to be poorer than would occur in the case of a designed study. In practice, this means that in many cases a designed study with probability sampling will likely dominate the use of administrative data—at least along the dimension of inference, if not necessarily along the dimension of cost-effectiveness. Figlio raised similar issues, pointing out that administrative data sets have considerably less flexibility and less information than is seen with designed studies. In addition, there are still many technical issues associated with the use of administrative data: sometimes the mechanisms that are necessary to link data across domains or follow individuals over time do not exist; attrition, while less of a problem than in most designed studies, is still prevalent, especially in open economies such as the United States; and administrative data sets are often not particularly well documented.

Cannot Fully Replace Random Assignment Experiments

As Hedges explained, there are controls and interventions used in random assignment experiments that cannot always be found or replicated in administrative data. While he recognized that quasi-experimental designs can be very instructive, the needed covariates are often not the ones that can be incidentally found in administrative data, and administrative data are by their nature not flexible in design regarding data collection. Of course, on the other hand, there are many research questions that are not well suited to random assignment experiments either, so an appropriate approach to thinking about random assignment experiments and quasi-experimental analysis using administrative data is to recognize that each approach is best for different types of research questions.

PROBLEMS WITH DISCLOSURE CONTROLS

Furthermore, Hedges described the issues associated with balancing statistical disclosure control with correct inference in administrative data. Statistical disclosure control requires that (1) individuals in the protected data cannot be reidentified and (2) the analyses of the protected data provide the same answer as the analyses of the original data. As Hedges noted, it is difficult to accomplish both of these goals. He presented examples from state longitudinal data to demonstrate how methods that mask disclosure by creating artificial "fuzziness" in the data in units with few observations, or deleting these cells entirely from the data set—both very common approaches to disclosure control—often lead to substantial bias in inference, because the patterns observed in these masked data can deviate dramatically from the equivalent patterns observed in the population as a whole. Put differently, although these masked data are likely to pose little inferential threat when describing a large population, this masking might produce extremely nonrandom population subsamples and, therefore, potentially very faulty inference when focusing on small subsamples. Given that, as mentioned above, the ability to study heterogeneous policy effects or rare or unusual events is a major advantage of using administrative data for research, this suggests that disclosure control methods could considerably reduce these benefits, at least when carried out in a naïve manner. Chapman also acknowledged issues with masking when making data sets available to researchers. One solution Hedges posited was the use of secured data centers for the handling of the data.

TENSIONS BETWEEN RESEARCHERS AND POLICY MAKERS

A final major theme of the panel involved the frequent tension between the objectives of researchers and policy makers. Hanushek presented a schematic of the researcher's "value function" and the politician's "value function" and described these objectives as being at odds with one another. The researcher, to Hanushek, wishes to optimize results for a given level of expenditure, and administrative data permit the evaluation of program results. The value of administrative data, therefore, is the degree of improved outcomes associated with the research or, alternatively, the cost savings associated with achieving the same outcomes. Hanushek argued that the researcher is interested in comparing this value to the expected losses associated with a privacy breach. The politician, to Hanushek, wants institutions to perform better, all else equal, but is constrained in ways that the researcher is not: there are political ramifications to learning that a politician's chosen policies do not work, for instance, and politicians are sensitive to a wide range of stakeholders' preferences. To the politician, researchers underestimate both the probability of a data security breach and the losses associated with that breach; that is, researchers underestimate both the politics and the estimated costs associated with making available administrative data to researchers. Politicians control access to administrative data, and Hanushek listed some of the ways some politicians limit research, for instance, through variable interpretations of the Family Educational Rights and Privacy Act (FERPA), state laws that go beyond the requirements of FERPA, implementation of bureaucratic hurdles, or onerous demands on researchers, such as required indemnification of the state against possible loss, establishment of extremely high price tags for data access, or control over the release of research results. Hanushek also described ways politicians might degrade the data, for example, by encouraging parents to opt out of assessments, or to not allow their children's data to be included in research databases. Hanushek also identified legitimate concerns by politicians and policy makers, including the time and expense to make data available as well as the consequences of bad research. As Hedges pointed out, state departments of education have limited resources and using them to make data available to researchers is not always their top priority or most important job. At the same time, Hanushek and others on the panel described cases in which policy makers are particularly and increasingly interested in knowing what the evidence is regarding particular policies or practices, suggesting that policy maker and researcher objectives, while not in complete alignment, are not entirely at odds with one another.

CONCLUSION

In all, the panelists described a situation in which administrative data present extraordinary new opportunities for scientific breakthroughs that have the possibility of dramatically improving educational policy and practice—while pointing out that administrative data in themselves are not a panacea, and that there are still many cases in which administrative data may not be extremely helpful. But there are many more cases in which administrative data can open up new research questions that have not been adequately studied in the past, if at all, as well as reducing the costs and increasing the benefits of designed studies like randomized control experiments and surveys. It is clear from the panel discussion that there are major political obstacles to widespread administrative data use for research purposes, and this is partially due to the fact that researcher and politician objectives are rarely aligned. But for all of the reasons mentioned in this summary, Hanushek summarized the explosion of access to administrative data as "revolutionary," with dramatic changes in what we know and what we do not know, and argued that "the future of the United States depends on getting [research access to administrative data] right."