THE PROMISE OF ADMINISTRATIVE DATA IN EDUCATION RESEARCH

David Figlio
(corresponding author)
Institute for Policy Research
Northwestern University
and NBER
Evanston, IL 60208
figlio@northwestern.edu

Krzysztof Karbownik
Institute for Policy Research
Northwestern University
Evanston, IL 60208
krzysztof.karbownik
@northwestern.edu

Kjell Salvanes
Norwegian School of
Economics
Bergen, Norway

Abstract

Thanks to extraordinary and exponential improvements in data storage and computing capacities, it is now possible to collect, manage, and analyze data in magnitudes and in manners that would have been inconceivable just a short time ago. As the world has developed this remarkable capacity to store and analyze data, so have the world's governments developed large-scale, comprehensive datafiles on tax programs, workforce information, benefit programs, health, and education. Although these data are collected for purely administrative purposes, they represent remarkable new opportunities for expanding our knowledge. We describe some of the benefits and challenges associated with the use of administrative data in education research.

doi:10.1162/EDFP_a_00229

INTRODUCTION

Thanks to extraordinary and exponential improvements in data storage and computing capacities, it is now possible to collect, manage, and analyze data in magnitudes and in manners that would have been inconceivable just a short time ago. And as the world has developed this remarkable capacity to store and analyze data, so have the world's governments developed large-scale, comprehensive datafiles on tax programs, workforce information, benefit programs, health, and education. Today, in many countries around the world, governments collect, maintain, and store an archive of information regarding a vast range of behaviors and outcomes over an individual's entire lifetime (Card et al. 2010). Governments have established statistics offices to maintain and use these data to produce official statistics about their populations. In the education sector, governments have invested large sums of funds to develop longitudinal data systems. The U.S. Department of Education alone has invested over $750 million to help states build, populate, and maintain these data systems.

Although these data are collected for purely administrative purposes, they represent remarkable new opportunities for expanding our knowledge and, through the conduct of analyses with more comprehensive data and better sources of exogenous variation than could typically be used in times past, challenging conventional wisdom in many areas based on previous research utilizing other sources (such as surveys). Administrative data also facilitate study of research questions that have heretofore not been possible to credibly study at all. Researchers who are able to access these data (especially those able to link data across administrative domains) have the ability to make extraordinary scientific advances by exploiting the population-wide datasets in combination with the increased opportunity for identification of causal effects through exogenous variation (by, for instance, policy changes, natural disasters, and other shocks that affect some groups of people but not others). In addition to natural experiments, these data can facilitate the conduct of field experiments, where the subjects of short-term experiments can be followed administratively for a longer period of time in manners that would have been impossible or prohibitively expensive to do absent large-scale administratively collected data. The new insights from these studies have extraordinary potential to inform education policy and practice, and we believe the massive growth in the quality and diversity of social science research on educational topics—on display over the past decade of the Association for Education Finance and Policy (AEFP)—is surely highly related to the increased availability of good administrative data. In this essay, we describe some of the benefits and challenges associated with the use of administrative data in education research.

BENEFITS ASSOCIATED WITH ADMINISTRATIVE EDUCATION DATA

There are many uses for designed survey data, but designed data collections are not a panacea. They offer great opportunities to ask questions in very specific ways, yet they are also expensive, necessarily have relatively modest sample sizes, are subject to attrition biases, and are not well-suited to prospectively studying policy and practice changes.

Administrative datasets offer a number of clear benefits for empirical research in education, supplementing designed datasets in some cases and supplanting them to

some degree in others. The ability to study population-level data offers a number of remarkable new possibilities that are extremely difficult to achieve with designed surveys and purpose-built datasets. Perhaps the most obvious involves statistical power—in contrast to datasets with hundreds or thousands of observations, administrative data sets with many times that number of observations mean that one can frequently detect modest but meaningful relationships with much greater precision than was previously possible. But there are at least two other distinct advantages of administrative data that are afforded by the large magnitudes of observations. One involves the ability to detect rare events that might be useful for identification: In administrative datasets, it is often possible to make twin comparisons or study children from three-child families; to investigate the effects of extremely rare climatic or seismic events that offer the opportunity for plausible identification of treatment effects; or to study specific economic events like plant closures (Roed and Raaum 2003; Card et al. 2010). In traditionally designed surveys, it is rare to have sufficient numbers of observations to be able to carry out analyses of these types. Another major advantage of having large-scale administrative data is the ability to study heterogeneous effects of educational policies and practice: With very large numbers of observations, it becomes possible to see whether the effects are similar across wildly different groups of individuals, and, if they differ, how they differ and for whom. Similarly, with population-level administrative data it is possible to study people at the extremes of the income distribution—say, those who are typically not well covered in traditionally designed surveys. Likewise, large-scale administrative data provide opportunities to use very rich nonparametric specifications when measuring effects, a valuable advantage in cases where relationships of interest may not be linear.

Another benefit of using administrative data for research purposes is that because data coverage is universal, it is possible to link administrative data from one domain (e.g., education) to data from another domain (e.g., workforce or health). This is obviously also possible in other non-administrative settings as well, but doing so is considerably more difficult because people would have to be purposefully longitudinally followed, and because a cross-section of educational data, for example, and a cross-section of health data may only include some of the same individuals by happenstance. Administrative data, by virtue of their population-level nature and the frequency of data observation, allow the researcher to follow individuals or entities over time so that there is a panel structure to the data.

Administrative datasets also provide novel types of variables typically not found in non-administrative data (Einav and Levin 2013). They can offer new opportunities, for instance, to look at measures of delinquency, of changing geographical location, of social networks, and of health instances that are nearly impossible to study in any other manner. The real-time nature of administrative data also provides new opportunities to study the effects of educational policies and practices that are very recent. It also offers the chance for researchers to make their scholarship much more relevant to the specific policy decisions that policy makers must make right away than are studies that make use of retrospective information (Einav and Levin 2013). And, of course, natural experiments need not be rare events to be better-studied using administrative datasets. Because natural experiments are unannounced, and often occur via chance or quirks, it is very difficult to set up a prospective study that will permit the evaluation

of a natural experiment. With administrative data that cover a population and that are recorded regularly, it is much more feasible to ex post identify and study these natural experiments (Roed and Raaum 2003).

Although not always the case, data quality is frequently better in administrative data than in retrospective data collection. Rather than asking people whether they participated in a given program twenty years ago, scholars who make use of administrative data can observe directly whether the individuals participated—according to the authorities who paid for the participation and therefore had a strong interest in correctly recording the occurrence! In addition, because of the mandatory nature of participation in the activities that generate administrative data, these data are much less likely to suffer from attrition problems or nonresponse problems than are data collected through voluntary means (Card et al. 2010). Likewise, administrative data are likely to be less subject to over-reporting or under-reporting of key variables than is the case with voluntarily collected data.

Administrative data also facilitate the study of intergenerational issues. It is possible, at least in some contexts, to match children's administrative records to that of their parents, and even grandparents. Whereas it is certainly possible to purposefully follow families longitudinally, the risk of attrition is surely greater when attempting to move from one generation to the next than if it is possible to directly match individuals using administrative means (Roed and Raaum 2003). And, in the case of questions that require a long amount of time to study in real time (e.g., intergenerational issues), the time horizon over which intergenerational questions may be studied can be shrunk considerably with administrative data.

Administrative data have major practical value for local policy as well. Different countries have extremely different policy environments, and so do different states. We absolutely can learn a lot from different contexts, but there are times when policy makers wish to act on data that are immediately relevant to their own populations. Developing, maintaining, and sharing administrative data with researchers creates new opportunities for knowledge creation that is directly tied to local policy, practice, and populations. And, of course, finding results from different settings increases the external validity and generalizability of research findings as well.

In summary, administrative data are more comprehensive than are designed survey data, and can be collected with frequently far more accurate information. Furthermore, the costs of conducting research with administrative data are much lower as well, at least once the data systems are developed. Once data structures are established, linking and extracting more records from administrative data cost only the time of the programmer. Also, the marginal cost of adding more individuals or periods of data to the analytical sample is extremely small, suggesting remarkably large economies of scale associated with administrative data (Roed and Raaum 2003). There are obviously many important roles for purpose-built designed survey data—not least the fact that only with purpose-built data is it possible to study precisely the questions that one wishes to study in exactly the manner in which one wishes to study them—yet it is also evident that administrative data offer numerous new opportunities to conduct research on questions that were previously impossible to study, or at least to study so well. Indeed, administrative data and survey data, although sometimes substitutes, can frequently be considered complements, as when administrative data can reduce the set of questions

that needs to be answered via surveys, or when administrative data can be used to serve as a check on the reliability of retrospective information collected via surveys (Roed and Raaum 2003). Administrative data can also be thought of as complementary to the conduct of field experiments, as the costs of tracking and following up with field experiment participants are much lower, and the data frequently much better, when the field experiments can be linked with data collected by governments for administrative purposes (Card et al. 2010). Wide access to administrative data facilitates accountability of research as well, and helps to ensure increased research quality (reducing researcher "monopolies" over data, therefore, makes more and better research possible). For all of these reasons, having a high degree of access to administrative data makes a wide range of empirical studies in education more feasible and more believable.

CHALLENGES ASSOCIATED WITH THE USE OF ADMINISTRATIVE DATA

Although administrative data provide many exciting opportunities for improving education research and practice, there are some substantial limitations as well. For one, administrative datasets are collected for different reasons than research, and the types of variables that are captured in administrative data often do not comport with the types of variables that testing many educational and social science theories demands. For instance, administrative datasets provide very little information on cognitive skills other than certain measures of achievement and attainment, or social and behavioral skills, such as standardized test scores, attendance, and suspensions. Many research questions demand that we know generally unmeasured information, such as motivation, attitudes, and "big five" psychological traits. In some countries, some of these variables are occasionally measured but even then the data are only seen in limited circumstances, such as in military data for men for a limited number of years. But this does not foreclose the opportunities to collect these data in the future, either formally as part of an administrative data collection or as a supplemental purpose-built survey. For example, in Norway, just prior to the decision to attend high school, a subset of pupils was surveyed about "big five" traits, time use, and other variables typically only seen in designed studies. These students were recruited into lab experiments regarding willingness to compete, risk taking, patience, and other variables, which was then appended to administrative data (Almås et al. 2014). Approaches like this provide the opportunity to capitalize on the best attributes of both designed survey data and lab experiments (ability to measure the variables we most want to observe), and administrative data (efficiency in following people, accurate program participation data, and the like).

Another shortcoming of administrative data involves technical issues. One huge advantage of administrative data for research purposes is the potential ability to match data across domains (e.g., combining education data with health, workforce, crime, or other data). In some locales, such as the Nordic countries, all administrative registers maintain common identification numbers and laws for statistical research purposes, enabling researchers to use merged registers across administrative units, with de-identified and merged data made available through the national statistical offices. But in many locations a unique personal identification number either does not exist or there are legal restrictions to merging data with registers across administrative

units.[1] In the United States, for instance, only a few states have linked children's social security numbers, which are used for all labor market and benefits data, to their birth records, and in many states it is illegal to link social security numbers to education records. In cases like these, it becomes much more challenging to link data across administrative domains. Although many states in the United States are making strong progress in linking education and workforce data (thanks to the leadership of the Data Quality Campaign—among other allied groups—a national organization dedicated to promoting the development, implementation, and use of high-quality administrative education datasets), this is a difficult and slow-going process.

Attrition is less pronounced in the case of administrative data than in most cases of designed survey data, yet it still presents substantial challenges in administrative data applications. People often move to other countries to work or change citizenship, and this issue is compounded in countries like the United States where people move freely and often between states but individual states maintain their own birth records, health records, education records, and workforce data. In recent years, it has become possible in rare circumstances to match school records to tax data from the U.S. Internal Revenue Service in order to follow children living in one state to adult outcomes in another state (see, e.g., Chetty et al. 2014), and we are hopeful that more cases like this will occur in the future (although these cases are themselves still very limited).

Moreover, because administrative datasets are not designed for research in the first place, they are often not particularly well documented. As a consequence, researchers using administrative data must often undertake large time and effort investments relative to those who make use of well-designed and well-documented surveys. Additionally, as part of the very nature of administrative data collection, it is not uncommon for variable definitions to change as the things being measured change, and these changes are not always easily available to outsiders of the administrative units collecting the data. Lastly, of course, it is always possible that the administrative datasets are incomplete or contain errors, since their purpose was never for research quality but rather for recording activities such as governmental program participation and compliance. As a consequence, these data may not have been subjected to the same type of quality assurance/quality control that is standard in the case of datasets collected specifically for research purposes.

One other highly important issue with using administrative registers is that, because of security and confidentiality concerns, they cannot be made available publicly. Given that unique personal identifiers exist, most countries with these datasets have developed secure systems for making them available to researchers or research groups. In the Nordic countries (with the exception of Finland, where a slightly different system is in place), very similar systems have been developed over time where statistical agencies play an important role in merging and de-identifying data for researchers, generally through research centers that have been through a quite extensive application procedure with data authorities, owners of data, and the national statistical offices. In all situations, governments must balance the costs associated with potential security breaches against the very large benefits of making data available to a wide range of researchers, who have

1. Another legal restriction was in place, until recently, in Finland where the possibility of indirect identification of individuals in small groups prevented researchers from using registry data in full capacity.

insights and expertise in a larger set of substantive research issues. It is imperative that researchers work diligently to ensure that they treat administrative data with care and maintain a high degree of security so that justifiably worried stewards of administrative data can feel more confident that sharing data with scholars provides high benefits to citizens with extremely low risks of security breaches or other forms of negligent behavior. Moreover, it is essential that researchers build trust relationships with partners in government so that officials are confident that the data they are entrusted to collect and protect are only used for the stated purposes, that they are not used to score political points, and that they are treated with care and the highest levels of protection.

CONCLUSION

We are learning more than ever before about education finance, practice, and policy as a direct consequence of the widespread use of administrative data in education research. Administrative data open up new questions that could not previously have been studied, allow us to reevaluate existing questions with new and more compelling empirical approaches and identification strategies, and permit analysis of questions of specific interest to particular localities. Therefore, although administrative datasets will never eliminate the need for or utility of purpose-built designed survey data, they can make these surveys more efficient and effective by concentrating more of the energy on the types of designed data collections not seen in administrative data, while supplementing surveys with administrative data when possible. The same, of course, is also true for field and lab experiments, which can be much more efficient when merged with administrative data.

For administrative data to reach their full potential for research, policy, and practice, however, more work needs to be done. There are practical and technical issues that impede the utility of these data, and governmental officials have good reason to be hesitant to widely share the data that they are entrusted to collect and protect. An important role of the membership of AEFP should be to show the value of thoughtful and mutually beneficial collaborations between researchers, practitioners, and policy makers. The more we engage in these types of partnerships and build trust relationships, the more likely it will be that widespread use of administrative data will continue to revolutionize and improve education research, practice, and policy in the United States and around the world.

ACKNOWLEDGMENTS

David Figlio acknowledges financial support from the National Science Foundation (award 1244752) for this work.

REFERENCES

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen, and Bertil Tungodden. 2014. Willingness to compete: Family matters. Bergen, Norway: NHH Department of Economics Discussion Paper No. 03/2014.

Card, David, Raj Chetty, Martin Feldstein, and Emmanuel Saez. 2010. Expanding access to administrative data for research in the United States. Washington, DC: National Science Foundation White Paper No. 10–069.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Quarterly Journal of Economics* 129(4):1553–1623. doi:10.1093/qje/qju022.

Einav, Liran, and Jonathan D. Levin. 2013. The data revolution and economic analysis. NBER Working Paper No. 19035.

Roed, Knut, and Oddbjørn Raaum. 2003. Administrative registers—Unexplored reservoirs of scientific knowledge? *Economic Journal (Oxford)* 113(488):F258–F281. doi:10.1111/1468-0297.00134.