Building Better Longitudinal Surveys (on the cheap)

Through Links to Administrative Data

September 2014

Susan Dynarski
Professor of Economics, Education & Public Policy
University of Michigan

Abstract

I make four recommendations regarding the longitudinal surveys of the National Center for Education Statistics. I recommend that NCES:

(1) Supplement its surveys with administrative data, focusing survey efforts on collecting data that is not contained in administrative sources.
(2) Use administrative data to convert what are now cross-sectional surveys into longitudinal surveys.
(3) Support researcher-initiated requests to link existing NCES surveys to administrative data sources.
(4) Explore methods to make microdata more widely available to researchers, particularly by replicating successful practices at Census.

I use the example of student borrowing to explore the rationale for these recommendations, discussing the data needs of researchers and policymakers in this arena. I describe prospects for fulfilling those needs with a combination of NCES surveys and administrative data.

I. Introduction

The charge of this National Academy of Education workshop is to evaluate the current and potential uses of the longitudinal surveys fielded by the National Center for Education Statistics of the US Department of Education (ED). The present paper is focused on how administrative data can be used to improve these surveys. Other papers in this workshop have done a thorough job of describing in detail the current surveys and their contents. Rather than duplicate their work, I discuss the strengths and weaknesses of the existing NCES surveys in broad terms. I discuss in more detail the administrative data that could be used to increase the power, scope and utility of the NCES surveys.

The scale and scope of education data has changed dramatically over the past few decades. Until recently, surveys were the main source of information on student outcomes such as enrollment, test scores, skills, educational attainment, employment and earnings. Education researchers relied heavily Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS, of the Bureau of Labor Statistics). The longitudinal surveys of the National Center for Education Statistics (NCES) joined this set of surveys in the 1970s.[1] NCES tracked a cohort each decade through high school, college and for a few years of employment. Other NCES surveys tracked cohorts of college students and kindergarteners.

For education researchers interested in documenting the life cycle of human capital accumulation, these detailed longitudinal surveys were the only game in town. Their main drawbacks were their infrequency and small size. The infrequency of the surveys meant they could not be used to track short-term changes that might arise from (for example) a shift in

[1] The BLS longitudinal surveys include NLS Young Men, NLS Older Men, NLS Young Women and NLS Mature Women, all initiated in the 1960s, and the National Longitudinal Surveys of Youth initiated in 1979 and 1997. See http://www.bls.gov/nls/.

national policy. The small size meant that they could not be used to measure variation across states, including that which might arise from (for example) a shift in state policy.

In the last ten years large, state longitudinal data systems have emerged that track entire populations of students from kindergarten through elementary and secondary school, into college and (in some cases) into the labor force. These datasets are maintained by school districts and states. They were not built for the needs of researchers but rather as a response to federal reporting mandates. An example is the No Child Left Behind Act, which requires that students be tested periodically and that scores be reported by race, ethnicity, eligibility for subsidized lunch, and special-education status. To comply with NCLB, states improved the student-level data systems that contained these measures.

Also spurring the development of the development of student longitudinal data system was the promulgation of a standardized definition of high-school graduation and dropout. This new definition involves the tracking of the enrollment and attainment of students from ninth grade forward. States improved the longitudinal tracking of their students in order to generate these standardized graduation rates.

Yet another impetus to the growth of state longitudinal data system were reporting requirements attached to federal stimulus funding during the recent recession. States that received State Fiscal Stabilization Funds were required to report on the number of high school graduates who attended college, completed a year of credits, and took remedial courses in college. To comply with these requirements, states linked their longitudinal data on high school students to data from their postsecondary systems and/or the National Student Clearinghouse (NSC).

II. The Prospects for Improving NCES Surveys Using Administrative Data

Combined, these various datasets track tens of millions of students, covering the entire population of elementary and secondary students in public schools and (through NSC) 93% of students at colleges nationwide. Unlike the NCES surveys, these administrative datasets contain a very limited number of variables. But they hold great promise as a complement to the NCES surveys, creating the opportunity for NCES to track a subset of outcomes for their survey respondents longer, more cheaply and more reliably. [2]

For example: NSC could be used to track the postsecondary attainment of respondents. While a survey was still in the field, NSC could capture enrollment of non-respondents. After the survey had closed, NSC could be used to continue to track postsecondary attainment into respondents' thirties or even forties. Since students increasingly continue postsecondary attendance into their thirties (Turner, 2004), this approach would more accurately measure postsecondary attainment than does a survey that stops at a younger age.

Administrative data could also be used to turn what are primarily *cross-sectional* datasets into *longitudinal* datasets. For example, the National Assessment of Educational Progress (NAEP) is a set of student-level cross-sections. Data from state longitudinal data systems could be attached, allowing for the measurement of grade progression, high school graduation, and standardized test scores in order to examine the relationship between the nationally normed NAEP (taken in grades 4, 8 and 12) and these outcomes.

[2] NCES has long supplemented its surveys with data from schools: several of the longitudinal surveys include high school transcripts, for example. What has changed is that this information is now collected, standardized and stored at the state (and sometimes national) level, which potentially reduces the time costs of collecting and harmonizing these data.

These linkages can be initiated both from the field (by researchers) and by NCES staff. Easing the process of making such linkages should be a priority for NCES. So, too, should be the development of expanded, secure channels for researchers to access NCES microdata. The NCES surveys are unusual among social science datasets in that accessing microdata of any sort requires a restricted-use data license, which in turn demands technological resources that many researchers do not have (e.g., a separate computer devoted solely to NCES data). NCES should look for inspiration to other data agencies (in particular, Census) that have made their microdata freely available to researchers.

III. Recommendations

In light of these opportunities, it's time to reconsider NCES's approach to survey design and collection. I have four recommendations

1) Supplement NCES surveys with administrative data. Focus NCES surveys on collecting information that is not contained in administrative data sources.

2) Use administrative data to convert cross-sectional surveys into longitudinal surveys.

3) Support researcher-initiated requests to link existing NCES surveys to other data sources.

4) Explore ways to make microdata more widely available to researchers, particularly by looking to successful initiatives at Census.

I describe the rationale for these four recommendations in the next section of this paper. In the last section of the paper I use a case study to explore the rationale for these recommendations. The case study is student debt; I discuss the data needs of researchers and policymakers in this arena and prospects for filling those needs with a combination of NCES surveys and administrative data.

Recommendation 1: Supplement NCES surveys with administrative data, focusing survey efforts on information not contained in administrative data

When there are administrative data available, NCES should rely on them, rather than duplicate efforts with a survey. For example, NSC contains information on college attendance spells and the identity of those colleges. The IRS holds similar data in the form of the information returns (1098-Ts) that colleges file for every student. This information can also be obtained from the majority of states that now have information on college attendance in their longitudinal data systems.[3]

Data from these administrative sources can therefore be used to measure spells of college attendance and the identity of the college attended, rather than surveys. At the very least, these data can be used to prepopulate surveys and respondent asked to confirm their accuracy.

Similarly, every state now tracks the enrollment and grade progression of its students. These data can be used to capture the identity of the school attended, grade attainment, and high school graduation. The data also include information on special education, English learner and subsidized lunch status. These data can be used to either replace or confirm survey responses on these topics.

IRS holds extensive data on the income and other characteristics of households. The Census Bureau has successfully negotiated access to these data for at least one of its surveys, the Survey of Income and Program Participation (SIPP). SIPP microdata are publicly available. Appended to the SIPP microdata are individual-level data from IRS and the Social Security Administration (SSA). In this dataset, the values of the IRS and SSA variables have been

[3] The disadvantage of the state data is that (unless the state has purchased NSC data or formed a data consortium with neighboring states) it will miss enrollment spells that occur outside of a state.

"perturbed" to protect the privacy of the individual respondents. This is referred to as the SIPP Synthetic Beta.[4]

Researchers who have used these public data to develop their statistical models can upload their code to run their analyses on the original, unperturbed SIPP/IRS/SSA data. After being checked for compliance with privacy protocols (e.g., no cells below a certain size), the analytic results are returned to the researcher.

Similarly, perturbed versions of IRS and SSA data could be linked (for example) to the NPSAS. Researchers could conduct initial analyses using these perturbed data and subsequently upload code to run on the complete, unperturbed data. Such an approach could overcome the legal, organizational and political barriers to a linkage of NCES data with other sources.

A key drawback is that NCES does not control these data sources and has to negotiate access. This drawback must be weighed against the cost of NCES duplicating the efforts of other agencies. By obtaining key variables from administrative sources, time and money are freed up that could be used to increase sample sizes, increase the length of follow-ups, and field more frequent surveys (e.g., a high school survey every five years rather than every ten years).

Further, these administrative data can be used to examine the characteristics of non-respondents. A key set of variables can continue to be collected even for those who leave the sample. These variables can be used to check the accuracy of sample weights that adjust for non-response and attrition. They can be used to answer these questions: Does the longitudinal behavior of non-respondents align with that of respondents who are observationally identical at baseline (the assumption underlying the construction of weights that account for attrition)? Do

[4] See http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html for more information.

the current sample weights, when used to reweight respondents, generate the same results as an administratively supplemented dataset that includes both respondents and non-respondents?

Recommendation 2: Use administrative data to convert cross-sectional surveys to longitudinal surveys

Administrative data can be used to turn what are currently cross-sectional datasets into longitudinal datasets. For example, the periodic National Assessments of Educational Progress (NAEP) data are student-level cross-sections. Data from states and NSC could be attached to NAEP in order to capture the relationship between the nationally normed NAEP (taken in grades 4, 8 and 12) and educational attainment. As discussed above, data from IRS or SSA on earnings could be combined in a secure setting, creating a dataset that allows us to measure whether the skills tested in NAEP translate into labor market success. These patterns could be examined separately by the variables gathered in NAEP: e.g., school type and student demographics.

If the waivers signed by participants in past NAEP do not allow for mergers at the individual level, group-level merges can be conducted. For example, lists of students in state-year groups could be created and sent to the relevant data-holder, who could then return cell means and variances.[5] This is how ED has obtained data from SSA on the earnings of college graduates in order to comply with the Congressionally required "gainful earnings" reporting.[6] It is also how researchers typically obtain merges of surveys such as the decennial Census with SSA earnings data (see, for example Angrist, Chen and Song, 2011). Properly defined cell means

[5] As long as they remain large enough, the groups could be sliced more finely: state-year-gender, for example, or state-year-gender-race.
[6] See US Department of Education (2012).

could then be used to evaluate the effect of state-specific policies on NAEP scores (as is currently done) as well as educational attainment and earnings.

This new resource would massively expand the data available to researchers to who seek to identify causal connections between state policies and educational attainment. Note that, in NCES data, such patterns can currently be tracked only in the relatively small and infrequent longitudinal surveys (every ten years), which is too infrequent for capturing the effects of short-term shifts in policy.

Recommendation 3: Support researcher-initiated requests to link the longitudinal surveys to other data sources

Going forward, NCES should create the conditions that allow their surveys to be regularly linked to the resources described above. However, NCES had a rich collection of existing surveys that could potentially be linked to such data. NCES should support researchers in linking the other data sources (e.g., NSC or state data) to the longitudinal surveys. This will, at low cost to NCES, improve the scope and quality of the surveys and make them more useful for researchers.

The NCES surveys are not currently amenable to such links. In older surveys, respondents did not sign waivers that would allow such administrative links; going forward this should be the standard protocol. For surveys in which these waivers have been signed, researchers need a process for requesting, making and paying for such links. Without such an institutionalized process, the creation of such links relies on individual relationships and the entrepreneurship of individual NCES staff. These efforts will falter as NCES staff shift jobs or retire.

Below I describe one possible, institutionalized process for data requests. There are many feasible models. What matters most is that the process is publicized and understood by the research community. NCES should not be discouraged if take-up is limited in the first few years that the process is put in place; it takes years for information about new data opportunities to disseminate through and embraced by the research community.

An online portal will be created for the initiation of proposals for data linkages. IES will review the scientific merit of proposed data links on a rolling basis. An NCES staff member would be tasked with working with researchers to get the pertinent identifying information transferred to the organization that holds the target data. In the case of NSC, for example, the names and dates of birth would be securely transmitted from NCES to NSC, which would match on data about college attendance and securely transmit it back to NCES.

NCES would merge the new data onto the longitudinal survey and release it to the researcher for analysis. The resulting match would be available to only the researcher who paid for it for a specified period of time (one year is reasonable) and would then be available to other researchers. The resulting data could be distributed through the current channels (CDs sent to those who hold restricted-use licenses). In the next section I describe some alternative models for distribution of NCES data, including those matched to administrative data.

Recommendation 4: Explore new ways to safely make more detailed microdata widely available to researchers

NCES releases public-use versions of its surveys, as well as restricted-use versions. I do not know of a resource that clearly delineates exactly what additional information is available in the restricted-use versions. Broadly, they contain more detailed information about geography, as

well as more finer-grained information about (for example) income.[7] NCES sends compact disks of its data to researchers who have completed restricted-use data licenses.

The intent of the restricted-use model is to prevent disclosure of respondent identities. The risk of disclosure rises as more researchers use it: it sits on more computers, for example, increasing the risk of a hack or stolen data. And the more researchers hold the restricted-use data, the more expensive it is for NCES to supervise its use. This suggests (to me, at least) that NCES should make the public-use versions as useful as possible in order to reduce demand for the restricted-use versions.

The restricted-data license model is a substantial roadblock for many researchers who want to use NCES microdata. Getting access requires completing a restricted-use data application. The speed of this application process has improved considerably, but it is still discourages many who want to make use of the data. Many policy analysts and Congressional staff, for example, can't set up a standalone computer for NCES data. Yet often they need to do analyses that depend on (for example), fine measures of income or geography. How to get this information into their hands, so that policy can be informed by the rich information NCES has gathered in its surveys?

NCES is more restrictive in what it includes in its public-use datasets than are other statistical agencies. The other surveys I use regularly as an education researcher are freely distributed as microdata, and they contain variables that NCES limits to the restricted-use

[7] I do not know of a document that details which variables are in the public-use datasets vs. the restricted-use datasets. It may well exist, but it's not easily available on the webpages that describe each data set (e.g., http://nces.ed.gov/surveys/nels88/data_products.asp), which is how researchers learn about NCES data products. Making the distinctions between variables available in each datasets would help casual users know whether they truly needed the restricted-use version or could make do with the public version.

versions. The American Community Survey, Current Population Survey, Survey of Income and Program Participation, Panel Study of Income Dynamics and the National Longitudinal Surveys are freely available for download by any researcher. The public versions of these data meet the needs of most researchers, because they include detailed measures of variables such as geography (state, sometimes metropolitan area) and income (sometimes top-coded, but always continuous rather than in brackets). Restricted-use versions contain even more information (e.g. county and school identifiers in the NLSY and PSID). The few researchers who need this additional information apply for restricted-use licenses. Everyone else can freely use the widely available microdata that does not include this information.

Census has been particularly aggressive and creative in both matching administrative data to its surveys and finding new ways to distribute those data to researchers. Public-use, microdata versions of their data products (e.g., American Community Survey, Survey of Income and Program Participation) are freely available for download by any researcher who wants access. Versions of the American Community Survey with very fine measures of geography (e.g., block) can be used in the Census Research Data Centers by researchers who have successfully applied for these data. A version of SIPP that is linked to data from the Internal Revenue Service is also available in the Research Data Centers. The Research Data Center model, however, is not one I would recommend to NCES, since they create their own barriers for researcher use.[8]

A promising route for NCES is that taken by Census for the distribution of exceptionally detailed versions of the SIPP. A version of SIPP that is linked to data from the Internal Revenue

[8] Many researchers do not live close to an RDC, and the process of getting a project approved is quite lengthy. There is also a fee associated with the use of the RDCs, though some universities cover this fee for researchers.

Service is available online in the form of the SIPP Synthetic Beta.[9] The SIPP Synthetic Beta is a promising model for NCES to handle the distribution of versions of its data that include variables sufficiently detailed that they threaten to reveal individual identities.

Census publicly releases a microdata version of its SIPP-SSA-IRS match that is "perturbed," with some variables statistically blurred to prevent identification. An unperturbed version of the matched data sits on the Census servers. Researchers run and refine their statistical models on the publicly available data, on their own computers. They can then upload the resulting code to the Census servers, where it is run on the original, unperturbed data. Results are returned to the researcher after being checked for compliance with data standards (e.g., minimum cell sizes).

This model could be used as the standard for merges of NCES surveys with sensitive data from IRS, SSA or the states. Agencies reluctant to allow their data to be released to researchers may well cooperate when the SIPP Synthetic model is used, with public versions being statistically perturbed. This approach appears to have worked with IRS and SSA, agencies that are notoriously protective of their data. I recommend that NCES consult closely with the Census staff who have successfully negotiated these data merges and releases.

[9] See http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html for more information.

IV. Case Study: Understanding student debt using NCES surveys and administrative data

In order to illuminate the promise of administrative data for improving these surveys, I focus on a particular topic in education: student loans. Student loans are the subject of heated discussion in Congress, the Department of Education, think tanks, the media, and the academic press. How have the longitudinal surveys informed this discussion? Where have they fallen short? How could administrative data be used to improve their utility? My intent is to use this particular case study to draw out lessons and insights about the longitudinal surveys that are relevant to multiple policy domains, not just postsecondary education.

The Policy Context

Student-loan debt has mounted to $1 trillion, now surpassing credit cards as the third-largest form of consumer debt.[10] This has triggered a national conversation about the cost of college and the appropriate level of student borrowing. With seven million student loans in default many ask, "Is there a student-debt crisis?"[11] At the heart of this question is a concern that borrowing is out of line with the value of college.

Economists point to the high payoff to college to put student borrowing in perspective. Over a lifetime, the typical holder of a bachelor's degree earns several hundred thousand dollars

[10] The top two forms of debt are home mortgages and car loans (Lee, 2013). Note that these widely cited statistics on loan debt are obtained not from NCES or the Department of Education but from credit records purchased by the Federal Reserve Bank of New York. The Federal Reserve Banks have been charged with improving tracking of student debt. Unable to obtain data from the Department of Education, they have resorted to buying data on federal loans from the private sector (i.e., credit agencies such as EquiFax).

[11] There are 6.5 million borrowers in default as of the third quarter of 2013. See http://studentaid.ed.gov/sites/default/files/fsawg/datacenter/library/PortfoliobyLoanStatus.xls, accessed October 2013.

more than a high school graduate (Barrow and Rouse, 2013). Even those with only some college see lifetime gains of about $100,000 (Greenstone and Looney 2013).

How do these figures compare to student borrowing? 69% of undergraduates borrow less than $10,000 and 98% borrow less than $50,000. About half of BA recipients borrow for college, with a debt of around $27,000 (College Board, 2012). While those who borrow would surely prefer not to be in debt, to most these debt loads look reasonable when compared to the financial gains of college.

The recent spike in defaults on student loans is worrisome. Defaults are not driven by the small fraction of borrowers with large loans: borrowers with typical levels of student debt struggle with their payments. Those who default have borrowed less than others: the average loan in default is $14,000 while the average loan in good standing is $22,000.[12] This pattern of defaults is consistent with two scenarios with very different implications for policy.

One scenario is that defaulters have *temporarily* low earnings and their loans fall into distress during these unusual bad times. At low cost to government, an income-based repayment (IBR) program would insure borrowers against these temporary downturns by automatically reducing their payments. If *lifetime* earnings are sufficient to pay off the loans, this system can be self-funding.

An alternative scenario is that those who default have *permanently* low earnings that cannot support even moderate debt loads. An IBR plan would still help these borrowers, but the ultimate cost to government would be much higher, since many of these loans will ultimately be

[12] Calculated from data in spreadsheet "Direct Loan and Federal Family Education Loan Portfolio by Loan Status," accessed October 2013. http://studentaid.ed.gov/sites/default/files/fsawg/datacenter/library/PortfoliobyLoanStatus.xls

forgiven. The cost of making, servicing and forgiving these loans could be so high that a grant program could be cheaper for taxpayers.

The Data Needed for Research

Distinguishing between these two scenarios requires individual-level, longitudinal data on student borrowing that follows former students for twenty-five years after college.

Why twenty-five years? IBR plans have students paying back their loans 20 to 25 years, when any remaining balance is forgiven. Costing out these programs therefore requires tracking earnings for decades.

Why individual-level, longitudinal data? Individual-level data are needed to capture the shocks to income that IBR programs insure against. The payments required of borrowers with different earnings paths cannot be backed out from group averages. Any analysis that relies on averages will smooth away the within-person shocks that are needed to estimate the benefits and costs of IBR.

Understanding the relationship between earnings and borrowing is critical for designing sound aid policy. If many former students are carrying debt beyond their capacity to repay, we need to reconsider the parameters of student borrowing, such as loan limits, loan forgiveness, and repayment structures. All of these topics are currently under discussion in Washington, with little data to inform the debate.

Data Prospects: Earnings

Multiple data sources contain information on lifetime earnings: The National Longitudinal Surveys (fielded by the Department of Labor) and the Panel Study of Income

Dynamics are well-known examples. The NCES longitudinal surveys, at present, do not contain data on lifetime earnings. The decadal cohort surveys, which follow a high school class every ten years, do not follow respondents beyond early adulthood, typically stopping when respondents are in their twenties.[13] The postsecondary surveys do not go much later, with the last surveys waves fielded ten years after the start of college.[14]

This leaves a major gap for researchers. Multiple studies have now shown that educational interventions do not demonstrate their full effects until students are well into adulthood (e.g., Kemple, 2008; Chetty *et al*., 2011; Dynarski, *et al*., 2013). The NCES surveys therefore miss many of the potentially positive effects of education policy. In the context of student loans, the NCES surveys stop before many students have finished paying off their education loans, which hampers the analysis of loan policy and the estimation of the long-term effects of college.

There have recently been advances in making administrative, longitudinal data on earnings available to researchers. As discussed earlier in the paper, the Survey of Income and Program Participation has obtained earnings data from IRS and the Social Security Administration for its samples.

---

[13] The National Longitudinal Survey of the High School Class of 1972 surveyed students until 1986, when they were about 32. High School and Beyond, which includes the high school class of 1982, stopped surveying students when they were in their twenties (in 1986, four years after high school). So did the National Education Longitudinal Study of 1988, which stopped surveying in 2000 (when respondents were about eight years out high school). The surveys currently in the field (Education Longitudinal Study of 2002 and the High School Longitudinal Study of 2009) are not planned to survey any later in life than their predecessors. See http://nces.ed.gov/surveys/hsb.

[14] The Baccalaureate and Beyond has varied in how long it tracks students. Graduates who started college in 1993 were followed for ten years, which would yield a typical exit age of late twenties. See http://nces.ed.gov/surveys/b&b/about.asp.

Data Prospects: Borrowing and Debt

The postsecondary surveys contain data on student borrowing and debt. Some of this information is drawn from respondent surveys (e.g., private student loans), but most of it is from the administrative system of the federal loan programs, the National Student Loan Data System (NSLDS). NSLDS contains information on the universe of federal borrowers.

NCES matches NPSAS respondents to NSLDS. These matches appear to be done at baseline and when the respondents are re-interviewed (it is unclear from the documentation). They are not regularly refreshed with updated data from NSLDS. In principle, these data could be updated each academic year.

This NPSAS-NSLDS match (and the matches with NSLDS of the datasets that are drawn from NPSAS) constitutes the most comprehensive survey data on student borrowing. The data are used and cited widely. The College Board uses NPSAS data in its popular *Trends in Student Aid* series. To my knowledge, these matches constitute the only dataset in which, at the individual level, detailed data are available on both demographics and borrowing.

The drawbacks of NPSAS are its frequency and size. New samples are started every four years (and its students surveyed every two years) and then data are released with a lag of a year or two. NPSAS therefore cannot be used to track annual fluctuations in student borrowing. And while NPSAS is large (100,000 students) it is not representative at the state level, so it can't be used to tabulate state-level estimates of student borrowing.[15]

[15] NCES has made substantial progress on this last point, with the recent NPSAS surveys including representative samples for some larger states.

Prospects for Improving Data on Borrowing and Earnings

As described above, many data sources contain individual-level, long-term earnings, while a smaller number contains information on student borrowing. The key data sources are listed in the Appendix, which serves to show that no single dataset contains the required information on both borrowing and income and to identify the datasets that could be combined. I next describe a scenario for linking these datasets.

The National Postsecondary Student Aid Survey (NPSAS) is the authoritative source for information about student borrowing (see Appendix). Information in NPSAS about individual students' federal borrowing is drawn from the administrative data that ED uses to run the loan programs (NSLDS). These data should be updated annually and updates should continue after the survey has "concluded." National Student Clearinghouse data can be used to update student postsecondary attendance.

A key weakness of NPSAS is its earnings data. For those who earn a BA, self-reported earnings data are collected every few years for just ten years after graduation. For those who do not earn a BA, earnings data are collected for just six years after college entry. The Social Security Administration and IRS hold detailed data on earnings and income. These data could be merged with the NPSAS data, just as SIPP data are currently merged with these data sources. These merges could continue well after the survey has left the field, and it can occur between surveys.

As is done with SIPP, perturbed versions of the NPSAS match with IRS and SSA data could be made publicly available. Researchers could conduct initial analyses using these perturbed data and subsequently upload code to run on the complete, unperturbed data.

V. Conclusion

There are no obvious conceptual, logistical or technological barriers to linking data from the NCES surveys to administrative data on earnings and borrowing. As I see it, the barriers are lack of communication across agencies and perceived legal barriers. When agencies work together, and legal restrictions more carefully inspected, creative solutions can emerge.

For example, there were (and are) organizational challenges in simplifying the process for applying for federal student aid (Dynarski and Scott-Clayton, 2006, 2007; Dynarski and Wiederspan, 2012). Conceptually, simplification was straightforward. A key element of aid simplification was linking IRS data to the online aid application (the FAFSA). The IRS resisted this linkage, citing restrictions to sharing data with other agencies. A creative solution that emerged from dozens of meetings allowed the linkage, enabling applicants to automatically populate their aid applications with their tax information.[16]

A similarly creative approach would allow for more timely and complete data on borrowing and earnings. More broadly, a commitment to linking the NCES longitudinal surveys to administrative data would broaden their scope, increase their accuracy and enhance their usefulness to researchers and policymakers.

[16] The workaround is that IRS (technically) provides data not to ED but to the applicant. While completing their FAFSA, applicants are prompted to log into the IRS servers. The IRS server passes the tax data to the applicant's web browser, and the applicant approves the transfer of the data from the browser to the aid application.

Appendix: List of Key Data Sources for Borrowing and Earnings

1) Administrative & Survey Data on Individual Borrowing

    a. National Student Loan Data System (NSLDS): This is a census of all federal student loans, used by the U.S. Department of Education's (ED) Federal Student Aid office to administer the loan program. It does not contain information on private loans. There is no research version of NLSDS and only limited summary statistics are released. No information on earnings is available in these data.

    b. National Postsecondary Student Aid Survey (NPSAS): this is a nationally-representative, longitudinal survey of college students run by ED's National Center for Education Statistics (NCES). Information about students' federal borrowing is drawn from NSLDS and merged onto NPSAS. For those who earn a BA, self-reported earnings data are collected for ten years after graduation. For those who do not earn a BA, earnings data are collected for just six years after college entry.

    c. Federal Reserve Bank of New York (FRBNY): the FRBNY has recently created a panel of credit reports that includes information about student debt (Lee, 2013). It does not contain information about earnings.

2) Administrative Data on Individual Earnings[17]

    a. Social Security Administration (SSA): SSA maintains longitudinally-linked records of individual earnings. These records are used to compute Social Security benefits, which are a function of lifetime earnings. Researchers have successfully linked these data to surveys, including the Census (Angrist, Chen, and Song, 2011). ED has used these data to calculate median earnings of graduates from career-training programs (ED, 2012). Since ED has already accessed these data for regulatory purposes, it is a particularly promising prospect for linking to ED data on loans.

    b. Internal Revenue Service (IRS): IRS maintains household-level records of income-tax returns and the informational returns that are used in the calculation of taxes. These data include information on college attendance, in the form of the 1098-T, which colleges send to the IRS to document tuition payments. In recent years, versions of

[17] I limit this list to earnings data that follow workers across state lines. Every state maintains longitudinally linked, individual earnings records for the purposes of administering unemployment insurance and workers' compensation. However, these records do not follow workers across state lines and so do not provide a comprehensive profile of lifetime earnings (especially for the college-educated, who are the most mobile).

these data have become available to outside researchers (e.g., Chetty et al., 2011). Treasury employees can also conduct research with these data, and outside researchers have coauthored with them on studies (e.g., Manoli and Turner, 2014).

c. Unemployment Insurance Earnings Records: states maintain records of earnings that occur within each state. The Bureau of Labor Statistics pools these records in order to generate local employment and earnings estimates. In theory these data are available for researchers within the Census's Research Data Centers, but a requirement that states opt into each proposed research project has hobbled their use.

Bibliography

Angrist, Joshua D., Stacey H. Chen and Jae Song. 2011. "Long-Term Consequences of Vietnam-Era Conscription: New Estimates Using Social Security Data." *American Economic Review* 101(3): 334-38.

Barrow, Lisa, and Rouse, Cecilia Elena. 2005. "Does College Still Pay?" *The Economists' Voice* Volume 2, Issue 4 2005 Article 3.

Chetty, Raj, Friedman, John N., Hilger, Nathaniel, Saez, Emmanuel, Whitmore-Schanzenbach, Diane, and Yagan, Danny. 2011. "How does your Kindergarten Classroom affect your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* Vol. CXXVI November 2011, Issue 4.

College Board. 2012. "Trends in Student Aid 2012." College Board Trends in Higher Education Series. Accessed at http://trends.collegeboard.org/sites/default/files/student-aid-2012-full-report-130201.pdf

Cunningham, Alisa F., and Kienzl, Gregory S. 2011. "Delinquency: the Untold Story of Student Loan Borrowing." Institute for Higher Education Policy (IHEP), Washington, DC. Available at http://www.asa.org/pdfs/corporate/delinquency_the_untold_story.pdf

Dynarski, Susan M., Joshua Hyman and Dian Schanzenbach (2013). "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion." *Journal of Policy Analysis and Management* 32:4, pp. 692-717

Dynarski, Susan M. and Scott-Clayton, Judith. 2007. "College Grants on a Postcard: A Proposal for Simple and Predictable Federal Student Aid." Hamilton Project Discussion Paper, 2007-01. Accessed at http://www.hamiltonproject.org/papers/college_grants_on_a_postcard_a_proposal_for_simple_and_predictable_fed/

Dynarski, Susan M. and Wiederspan, Mark. 2012. "Student Aid Simplification: Looking Back and Looking Ahead." *National Tax Journal* 65:1, pp. 211-234.

Dynarski, Susan M. and Scott-Clayton, Judith. 2006. "The Cost of Complexity in Federal Student Aid: Lessons from Optimal Tax Theory and Behavioral Economics." *National Tax Journal* 59:2, pp. 319-356.

Greenstone, Michael, Looney, Adam, Patashnik, Jeremy, and Yu, Muxin. 2013. "Thirteen Economic Facts about Social Mobility and the Role of Education." The Hamilton Project Policy Memo. June 2013. Accessed at http://www.hamiltonproject.org/files/downloads_and_links/THP_13EconFacts_FINAL.pdf

Kemple, James (2008). "Career academies: Long-term impacts on labor market outcomes,

educational attainment, and transitions to adulthood." Unpublished manuscript.

Lee, Donghoon. 2013. "Household Debt and Credit: Student Debt." Federal Reserve Bank of New York. February 28, 2013. Accessed at http://www.newyorkfed.org/newsevents/mediaadvisory/2013/Lee022813.pdf

Manoli, Dayanand S. and Turner, Nicholas. 2014. "Cash-on-Hand and College Enrollment: Evidence from Population Tax Data and Policy Nonlinearities." NBER Working Paper 19836.

Turner, Sarah. (2004). "Going to college and finishing college. Explaining different educational outcomes," in *College choices: The economics of where to go, when to go, and how to pay for it* University of Chicago Press, pp. 13-62.

U.S. Department of Education, Federal Student Aid. 2012. "Gainful Employment Operations Manual." Accessed at http://www.ifap.ed.gov/GainfulEmploymentOperationsManual/attachments/GainfulEmploymentOperationsManualMasterFile.pdf

U.S. Department of Education, National Student Loan Data System. 2013. Data in "Direct Loan Portfolio by Loan Status and Federal Family Education Loan Portfolio by Loan Status." Accessed at http://studentaid.ed.gov/sites/default/files/fsawg/datacenter/library/PortfoliobyLoanStatus.xls