Federal Efforts Could Improve the Data Available to Drive Improvement in Higher Education

Jordan Matsudaira
September 2017

Enhancing the quality and accessibility of data on colleges and universities for students, institution leaders, and policymakers can improve college choices and advance evidence-based decisionmaking. Although efforts to repeal the student unit record ban such as the recently proposed College Transparency Act are important, this memo outlines several important and low-cost steps the US Department of Education could take today to improve the data available to drive improvement in higher education.[1] I argue that the value of existing data sources could be increased by (1) customizing information to students to be more relevant to their interests and backgrounds, (2) using existing platforms—such as the online Free Application for Federal Student Aid (FAFSA) application where students apply for federal financial aid—to push information to students, and (3) making data available to researchers in a secure environment to advance our understanding of higher education policy and improve program administration.

Background

These are strange times for higher education in the United States. On one hand, the share of Americans who say college is important has never been higher, college wage premia are near historic highs, and students are pursuing postsecondary education at near record rates. At the same time, there is rising skepticism over the value proposition of college, with scores of news articles posing some version of the question "Is college worth it?"

It is no surprise that the financial calculus of the college decision has come under new scrutiny. State disinvestment in public colleges and universities, a reflection of public ambivalence about higher education's value, has led to students paying a higher share of college costs and increasingly relying on loans to do so. More and more students are struggling to repay those loans once they leave college, and the news is filled with stories of sham degree programs taking students' dollars and dreams and leaving them with no marketable skills.

Although academic research finds high *average* returns on college investments, this masks substantial heterogeneity across students, the colleges they attend, and their programs of study. For example, new research points to low and even negative earnings gains for students attending for-profit institutions (Cellini and Turner 2016), and students who attended for-profit institutions accounted for 44 percent of all defaults in 2013, but only 11 percent of enrollment (Looney and Yannelis 2015).

The recognition that students' outcomes are greatly, and at times adversely, affected by their institutions and programs of study has fueled calls for better data in higher education to support better information for college choice, accountability, and performance management reforms. As I discuss below, data users differ in their data needs because of differences in their objectives, and the federal government should approach efforts to improve higher education data with this in mind.

Improving Data on College Costs for Prospective Students

Students choosing which college, or whether, they should attend need information on how much they will have to pay to attend and complete their intended program of study at each institution and how their lives are likely to differ along outcomes that are important to them.

The cost side of this equation seems straightforward, but despite recent efforts, accurate estimates of the price of college remain elusive or difficult for students to obtain. Federal government websites such as the College Scorecard list information on average net prices—that is, the price students will pay for tuition, fees, and room and board after grant and scholarship aid is deducted—for all students receiving aid and for the average student who receives federal aid (a grant or loan) within each of five family-income categories. These are based on data collected through the Integrated Postsecondary Education Data System (IPEDS), a compulsory survey for all institutions participating in federal student aid programs.

These data suffer from inaccuracies that have been documented elsewhere (Anthony, Page, and Seldin 2016; Kelchen, Goldrick-Rab, and Hosch, forthcoming; Levine 2014). Though many of the documented inaccuracies in net price information tend to understate net prices, many students still overestimate the price of attending college by focusing on sticker prices that do not account for grant aid (Hesel, Camara, and Kappler 2015).

Recommendation 1: Produce better data on net tuition using data already reported by colleges to the Internal Revenue Service (IRS) on Form 1098-T. The federal government could improve the price information at students' disposal with minor modifications to current data collection and dissemination procedures. Institutions already report student-level data on payments for tuition and related expenses (i.e., fees and course materials) as well as scholarship and grant awards to the IRS through Form 1098-T to allow tax filers to claim education-related credits. With simple modifications to reporting by institutions (e.g., requiring institutions to report on an award year in addition to the current tax-year basis and requiring the 1098-T to be filed even for students whose tuition expenses are completely paid through scholarship and grant aid), these data could be used to generate accurate estimates of the amount paid in tuition for each student for each year the student is enrolled. These data could be collected directly by the Department of Education, or the averages for various subgroups in each institution could be produced by the IRS and shared. Because expenses for room and board and other noneducational expenses are not eligible for tax credits, data on those components of the price of college would continue to be based on existing surveys (i.e., IPEDS).

Recommendation 2: Provide students individualized estimates of net prices paid by similar students at colleges of interest when they apply for federal aid. Information will only help students if they have access to it, and the federal government could do more to ensure students can easily access information on college prices. With the data on tuition and related expenses described above, the Department of Education could create highly individualized estimates of the net prices students would face at each institution and ensure students can access the information.

When students file their FAFSA forms online, they could instantly be given a range (e.g., the 10th, median, and 90th percentiles) of net tuition expenses charged to students *with the same family income background and other characteristics* at each of the institutions where they express interest. (Students currently choose up to 10 institutions to receive their FAFSA forms). If the FAFSA form were changed to collect information on academic background (e.g., SAT or ACT scores or high school grade point average), or if those data were gathered through other means, these net price predictions would more accurately reflect differences in merit aid availability across institutions. Moreover, estimates of net price in the second and later years at an institution could also be presented, and these data could further be disaggregated by program of study (which may become increasingly important as more institutions charge different tuition across majors). Noisier room and board estimates could be presented separately for different living situations based on IPEDS data, with links to campus housing information to allow students to better estimate those costs.

With the switch to the use of income information from an earlier year on the FAFSA form, all this information can be delivered to all students applying for federal aid early enough in the college search process for them to have good estimates of the relative prices of their options *before* applying to schools.

Improving Data for Prospective Students on College Benefits

Capturing the benefit side of the college choice decision is more complicated. Students have different reasons for attending college, and no data system can capture student success along relevant dimensions for every student. There are nonetheless many outcomes that students identify as key. Getting a better job, learning about areas of interest, getting training for a specific career, gaining an appreciation of ideas, making more money, and preparing for graduate school are all identified as "very important" college objectives by over 50 percent of students entering baccalaureate institutions (Eagan et al. 2016).

Even for these indicators, there are important conceptual gaps between students' objectives and what can be measured with available data. The federal government has high-quality data on employment and earnings from IRS reports from employers and individual tax filings, and these data were incorporated into the College Scorecard in 2015. Although this is valuable information, students may also wish to know how successful program alumni are in securing employment in a particular industry, occupation, or company, and such data are currently unavailable.

Completion rates have long been viewed as "the" measure of institution quality in higher education. Ideally, accreditation procedures would ensure that attaining a degree equates with some level of learning or mastery of skill for a particular career. But completion indicators seem only weakly correlated with labor market outcomes across colleges, especially in the two-year sector, casting doubt on whether they are good proxies for quality and useful for guiding college choices (Council of Economic Advisers 2015).

This uncertainty in their value notwithstanding, completion data are getting better. IPEDS completion rates will soon be expanded to include completion outcomes for both part-time and transfer students and be reported separately for Pell grant recipients. Moreover, the College Scorecard data files contain measures of cohort graduation rates that measure whether students transfer to and graduate from another institution over various time horizons,[2] similar to the outcome data produced by the Student Achievement Measure consortia. These data are accurate for students with federal loans, but will become accurate for all federally aided students as institution reporting to the National Student Loan Data System (NSLDS) improves.

Outside of earnings, employment, and completion measures, the federal government collects little information that tracks the success of students in higher education, but there is some scope for improvement. With small tweaks to processing online tax filing forms, employment in particular occupations and industries could be tracked.[3] Beyond labor market outcomes, more experimentation on how best to capture student success is needed. Although there have been calls for widespread measurement of learning based on assessments such as the Collegiate Learning Assessment, research suggests those indicators have poor predictive validity for labor market outcomes (Melguizo et al. 2015; Riehl, Saavedra, and Urquiola 2016). And because such tests are meant to measure skills valuable for employment rather than learning we might value per se, it seems ill advised for the federal government to advocate such measures unless and until better assessments are developed.

An area where more research might be useful is in measures of quality of life or life satisfaction. Such questions could reflect intangible benefits of higher education. And while some third-party data collectors have collected such data for institutions (e.g., Payscale and Gallup), the Department of Education could presumably field such surveys to consistent and well-defined groups of students across institutions, say, through surveys of borrowers at each institution measured years after they exit.

Federal data on completion and earnings have been criticized for not capturing the outcomes of all students at each institution. Earnings data in the College Scorecard, for example, cover only federally aided students, as do completion outcomes based on the NSLDS. Completion outcomes in IPEDS cover only first-time, full-time students, though this is scheduled to change next year.

This criticism lacks focus. Students' decisions about where to attend college are best informed by predictions of their outcomes at different colleges, or equivalently, estimates of the causal impact on some important outcome of attending School A relative to School B (or no school at all). Students' future earnings (or any other outcome) are influenced by many factors beyond where a student attends school and what he or she studies—family's income, academic background, interests, and so on—and the

composition of the student body differs across institutions. For this reason, much of the difference in average student outcomes across institutions is driven by differences in the composition of the students an institution enrolls.

These "composition effects" (an economist might call them "selection effects") compromise the value of data on "typical" (average) student outcomes for most purposes. And the greater the coverage of a statistic—that is, the more types of students a given metric covers at each institution—the more likely differences in the metric are to reflect composition effects, as opposed to the causal impact of attending one institution or the other for a given type of student. Average (or median) earnings differences in the College Scorecard, based only on federally aided students, are likely better consumer information than the average earnings of all students would be. The latter likely involves larger differences in family income across institutions, and thus, earnings differences will more likely be driven by composition effects.

The problem with current measures of institution or program outcome data for informing college choice is not that they do not include particular groups, such as non–federally aided students. Rather (1) the measures include too many types of students, so comparisons across schools reflect composition effects and school quality effects; and (2) data for some subgroups of students do not exist, so accurate predictions of their outcomes may be difficult to obtain.

How can the federal government improve the quality of available information on the potential benefits of different educational opportunities to students?

Recommendation 3: Provide students personalized predictions of the likelihood that they complete programs of interest and the earnings outcomes associated with these programs. Simple statistical models could generate institution- and program-specific predictions of each student's completion and labor market outcomes based on the information (e.g., family income, age, gender, parental education, and high school attended) students enter on the FAFSA form.[4] As with information on the price of attendance, these predictions would be improved with information (not currently collected at the federal level) on students' academic backgrounds. And as with data on net prices discussed earlier, minor tweaks to the FAFSA website could ensure all federal aid applicants get estimates of their outcomes at each of the schools they list on their FAFSA, leading to an immense improvement in the availability of such information relative to net price calculators currently available through individual college websites.[5]

Data constraints and the protection of information pose important technical challenges for progress in generating these data. The Department of Education can address these issues by fostering collaborations with states, institutional consortia, and the research community. For example, the federal government collects no individual-level information about academic achievement, and the Department of Education will not be able to access individual data for students not receiving federal aid unless Congress removes the ban on student unit record systems in the Higher Education Act. On the other hand, states and institutions often have data on students' academic history from K–12 into

(public) institutions of higher education and students' intended program of study that are not limited to aid recipients.

Recommendation 4: The federal government should work with states and institutions to match their data to employment and earnings (and other) outcomes from tax data to incorporate the advantages of information held outside the federal government. A handful of such innovative efforts have been accommodated. The University of Texas system arranged a match between its student data and earnings information in the Census Bureau's Longitudinal Household Employer Dataset's holdings of state unemployment insurance data to improve its SeekUT student information tool.[6] But the Department of Education could accelerate progress by, for example, creating competitive grants for innovative uses of federal, state, and institutional data to develop better information tools, data-driven advising and other student support services, and so on. At a minimum, the federal government should clarify and expedite the process for states and groups of institutions to pursue data matches. The federal government needs to make data available to, and work with, the research community to help ensure the data produced provide maximum benefits to students while ensuring privacy is not compromised.

The College Transparency Act and "Getting the Data Right"

There are important design considerations that affect the value of any information created for prospective students. The choice of which college to attend should depend on predictions for what students would pay and the outcomes they will likely realize at each institution under consideration. The limitation in using aggregate information—such as the average net price, completion, and earnings of all federal aid recipients—for this purpose is that differences in these averages across institutions are affected both by differences in outcomes for particular students *and* by differences in the students enrolling at each institution.

For example, the median annual earnings of students who attended Cornell University is $14,700 higher than the median earnings of students who attended the State University of New York (SUNY) Binghamton ($72,100 versus $57,400) 10 years after students first enroll. Part of this difference might be attributable to the relative quality of education offered by Cornell, but some of the difference owes to differences in the students enrolled at each institution: Cornell students have family incomes more than $22,000 higher, have average math and reading SAT scores about 125 points higher, and are more predisposed to higher-earning fields (e.g., 18 percent of degrees are awarded in engineering versus 10 percent at SUNY Binghamton). For a given student, the difference in earnings they should expect from choosing between these colleges is likely less than they would be led to believe by the average (or median) difference among all students. In general, this inaccuracy is likely to bias upward students' assessments of the benefit of attending schools that are more selective, that enroll more affluent students, and that focus on programs that are more rewarded in the labor market.

Like the data behind the College Scorecard, the College Transparency Act proposes calculating averages of important outcomes for different student subgroups, such as race and ethnicity, gender, age, first-generation status, Pell receipt, and program of study. Although well intentioned, this type of

data disaggregation is not likely to help students, who occupy multiple statuses at once. An Asian 22-year-old male who is a first-generation Pell recipient interested in majoring in biology would presumably compare *six* averages (i.e., the average for Asians, the average for 22-year-olds, etc.) for each outcome across the schools he was considering, with no clear way of combining that information to make a principled guess about the school where his outcome is likely to be highest. Instead, better predictions could be offered to students by (1) using statistical models to provide individualized predictions based on student characteristics, or (2) generating average outcome data for a set of mutually exclusive student types (e.g., one type might be "traditional-age males interested in science, technology, engineering, and math"). Data scientists and other researchers should be enlisted to identify the discrete student types (or best statistical model) that best captures variation in student outcomes within colleges, while ensuring the privacy of individuals' information.

The draft of the College Transparency Act contains a provision that would limit its ability to provide accurate information to students. In particular, it prohibits the Department of Education from incorporating data on student academic preparation (e.g., SAT or ACT results or high school grade point average). This is an important restriction, because academic background is an important predictor of future outcomes for students enrolling in the same institution (Cunha and Miller 2014). The consequence of not accounting for academic preparation would likely be to give students an inflated sense of the benefit of attending more academically selective schools, risking steering students to "low-value" schools.

Data for Policy Research and Evaluation

Another area where the federal government could improve the higher education sector is in expanding the use of its administrative data for research and evaluation to improve programs and support evidenced-based policymaking. The release of data on average earnings, completion, and borrowing outcomes for detailed subgroups in the College Scorecard was a dramatic improvement in the data available to researchers to monitor the performance of colleges and universities and assess the broad impact of policies. For a broad set of pressing policy issues, however, individual-level data are required to shed light on how to best design public policies.

Suppose one wanted to identify institutions where increasing Pell awards would have the greatest impact on students' chances of success to, for example, offer "Pell bonuses" to students attending such institutions. Such an analysis requires identifying students within each institution that differ in the amount of Pell they receive, but are otherwise similar to one another, to calculate the independent (i.e., causal) effect of Pell on their outcomes (e.g., graduation rates, postenrollment earnings). Although researchers have used state-level datasets combined with sophisticated quasi-experimental methods (comparing students with family incomes that are similar but who nonetheless experience large differences in their Pell awards because of discrete jumps in the Pell grant formula) to do this for state higher education systems, federal data would allow researchers to compute such information for the near universe of institutions (Carruthers and Welch 2017; Denning, Marx, and Turner 2017). Unfortunately, data on individual federal aid recipients are currently not made available to researchers.

As another example, suppose a policymaker wanted to understand which students might be affected by a policy that converts state tuition grants to loans if students work out of state after attending a state school, as does New York's Excelsior Scholarship. Linking state data with student demographics and financial aid information to federal tax data would permit a calculation of how many students in each of various income and racial groups work out of state at each institution and what fraction of those students might have income below some level where the conversion to debt would be greater cause for concern. Future policy researchers could use the same data to assess whether such a policy affected students' mobility and economic success and evaluate whether there were unintended consequences that should be addressed.

Finally, designing effective accountability systems requires information that is currently nonexistent on how students and institutions will respond to the incentives created. For example, the impact of a risk-sharing proposal will depend, inter alia, on how much institutions increase their tuition, whether and how institutions change their admissions policies in response to changes in net revenue associated with different students, whether institutions continue to operate, how students respond to changes in tuition (i.e., if they choose not to attend college or to attend elsewhere), and whether and where students attending institutions that close reenroll. Not all these quantities can be known ex ante, but researchers can and have used historical administrative data to shed light on these issues. Because all these quantities are causal relationships, detailed individual-level data provide the best opportunity to develop sensible estimates of how students and institutions will respond to, and estimate the costs and benefits of, any accountability scheme.

There are little data available to inform the questions described above.

Recommendation 5: The federal government should make its administrative data more readily available to policy analysts and the research community. A staged approach to data access would maximize the benefits to releasing data while minimizing the risk of privacy disclosure. First, a web-based tool (similar to the PowerStats application that the National Center for Education Statistics, or NCES, uses to facilitate simple analyses of its various survey datasets) to calculate simple descriptive statistics (e.g., cohort completion rates by institution or by income category and sector or loan repayment rates for independent students by program of study) based on data in the National Student Loan Data System would likely serve the needs of many policy analysts, institutional research office staff conducting benchmarking studies, and journalists. For more nuanced, but still broad, policy questions—for example, estimating the effect of eligibility for an income-based repayment plan on the likelihood of default—a public-use individual-level dataset could be developed based on a probability sample of the NSLDS data, with all personal identifying information removed. This dataset would serve the needs of more advanced policy researchers and support more sophisticated analyses and evaluations. Similar data are already made available to researchers by the Department of Education through the NCES, sometimes under a restricted data-use agreement to provide an extra level of data security.

Finally, the Department of Education should expand the ability for researchers to use the raw data in the NSLDS (and associated data) for projects that require the universe of aid recipients, for example,

to study small subgroups or to address questions that might involve merging data from outside the department. Just as the IRS has developed procedures for researchers to submit proposals and conduct research on highly sensitive tax data to inform tax policy, the Department of Education could allow researchers access to its data. The Office of Federal Student Aid's new Data Warehouse is a good platform to facilitate this work for projects involving confidential student- and institution-level data, but the Department of Education can leverage the utility of its data to policy researchers even more by facilitating merges with critical outside data, such as students' postenrollment employment and earnings. One option for doing so would be to share extracts of its Data Warehouse files (based on the NSLDS) with the Census Bureau's Center for Administrative Records Research and Applications, where researchers could work with these data, potentially matched to other administrative data, in a secure environment.

Conclusion

Many observers have suggested reasons the gains associated with improved data might be low. But the costs of improving data are low. Most of the recommendations in this memo involve minor modifications to current practice and could improve decisionmaking and the efficiency of government programs for many actors. Improving the data available on postsecondary institutions is important, but the federal government should not lose sight of its resources that could be deployed in ensuring these data are seen and used by prospective students, institutions, and policymakers to make better decisions. Taking advantage of federal websites accessed by most students to apply for financial aid to deliver personalized data is one example, but the federal government should direct more effort toward creating tools and incentives for other stakeholders to use existing data to drive improvements in the sector.

Notes

1. For a comprehensive overview of federal data on higher education, important gaps in the system, and recommendations to address these gaps, see the Institute for Higher Education Policy's collection of papers on the National Postsecondary Data Infrastructure at "Envisioning the National Postsecondary Infrastructure in the 21st Century," Institute for Higher Education Policy, Postsecondary Data Community, accessed July 19, 2017, http://www.ihep.org/postsecdata/mapping-data-landscape/national-postsecondary-data-infrastructure.

2. These data are based on information in the National Student Loan Data System, an administrative dataset used by the Office of Federal Student Aid at the Department of Education to manage federal aid, especially loan repayment.

3. Occupation is already asked on IRS tax forms, but is not processed in a way that allows it to be easily categorized. State earnings information collected through the unemployment insurance system typically includes industry identifiers for employers, but to my knowledge, this information is not used by any higher education information system. Minor tweaks to information collection (e.g., modifying tax preparation software to collect standardized occupation codes) could make these data more readily available.

4. Information on program is not currently, but will soon be, available in the NSLDS, allowing information to be reported by the program a student graduates from. It would be better to define student subgroups for reporting purposes based on ex ante student plans over planned program of study to mitigate concerns that differential dropout rates across programs could bias results. Doing so, however, would require modifying data collection procedures to elicit such plans from students before they enroll.

5. Although websites such as CollegeAbacus.org that allow students to access information from multiple institutions' net price calculators at once help students in the current environment, the Department of Education could provide this information more accurately and automatically to students.

6. See, for example, the University of Texas System, "UT System partners with US Census Bureau to provide salary and jobs data of UT graduates across the nation," news release, September 22, 2016, https://www.utsystem.edu/news/2016/09/22/ut-system-partners-us-census-bureau-provide-salary-and-jobs-data-ut-graduates-across.

References

Anthony, Aaron M., Lindsay C. Page, and Abigail Seldin. 2016. "In the Right Ballpark? Assessing the Accuracy of Net Price Calculators." *Journal of Student Financial Aid* 46 (2): 25–50.

Cellini, Stephanie Riegg and Nicholas Turner. 2016. "Gainfully Employed? Assessing the Employment and Earnings of For-Profit College Students Using Administrative Data." Working Paper 22287. Cambridge, MA: National Bureau of Economic Research.

Council of Economic Advisers. 2015. *Using Federal Data to Measure and Improve the Performance of US Institutions of Higher Education*. Washington, DC: Council of Economic Advisers.

Carruthers, Celeste K., and Jilleah G. Welch. 2017. "Not Whether, but Where? Pell Grants and College Choices." Knoxville: University of Tennessee.

Cunha, Jesse M., and Trey Miller. 2014. "Measuring Value-Added in Higher Education: Possibilities and Limitations in the Use of Administrative Data." *Economics of Education Review* 42:64–77.

Denning, Jeffrey T., Benjamin M. Marx, and Lesley J. Turner. 2017. "ProPelled: The Effects of Grants on Graduation and Earnings." Unpublished working paper.

Eagan, Kevin, Ellen Bara Stolzenberg, Abigail K. Bates, Melissa C. Aragon, Maria Ramirez Suchard, and Cecilia Rios-Aguilar. 2016. *The American Freshman: National Norms Fall 2015*. Los Angeles: University of California, Los Angeles, Higher Education Research Institute, Cooperative Institutional Research Program.

Hesel, Richard A., Wayne Camara, and Steve Kappler. 2015. "Student Perceptions on Price, Aid, and Debt Provide an Extraordinary Opportunity for Colleges and Universities." *StudentPOLL* 11 (1): 1–10.

Kelchen, Robert, Sara Goldrick-Rab, and Braden Hosch. Forthcoming. "The Costs of College Attendance: Examining Variation and Consistency in Institutional Living Cost Allowances." *Journal of Higher Education*.

Levine, Phillip B. 2014. "Transparency in College Costs." Washington, DC: Brookings Institution.

Looney, Adam, and Constantine Yannelis. 2015. "A Crisis in Student Loans? How Changes in the Characteristics of Borrowers and in the Institutions They Attended Contributed to Rising Loan Defaults." Washington, DC: Brookings Institution.

Melguizo, Tatiana, Gema Zamarro, Tatiana Velasco, and Fabio J. Sánchez. 2015. "How Can We Accurately Measure whether Students Are Gaining Relevant Outcomes in Higher Education." Los Angeles: University of Southern California.

Riehl, Evan, Juan E. Saavedra, and Miguel Urquiola. 2016. "Learning and Earning: An Approximation to College Value Added in Two Dimensions." Working Paper 22725. Cambridge, MA: National Bureau of Economic Research.

Jordan Matsudaira is a nonresident fellow at the Urban Institute and assistant professor of public policy and economics at Cornell University. He researches the impact of education and labor market policies and institutions on the mobility of low-income Americans. From 2013 to 2015, he was senior and then chief economist on President Obama's Council of Economic Advisers. Matsudaira led the interagency team that

developed the College Scorecard, a web-based database designed to make institution-specific information on prices and student loan, completion, and labor market outcomes available to students and policymakers.

Acknowledgments

URBAN
INSTITUTE

2100 M Street NW
Washington, DC 20037

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.