Merging Data to Facilitate Analyses

Robert Kelchen
Associate Professor, Seton Hall University

robert.kelchen@shu.edu, @rkelchen

April 2020
About my presentation

• Based on an article in New Directions for Institutional Research in a special issue on national postsecondary data (link)
• Lots of great articles in that issue!
• Pre-publication copy is freely available on my personal website (link)
Why do institutional data merges?

• Benchmarking (the most common reason)
• Getting out in front of accountability pressures to demonstrate value
• Understand and model changes in college ranking/rating and performance funding systems
• Answer institutional and public policy questions
Outline of today’s talk

• Federal data sources
• State/local data sources
• Examples of answering questions using multiple datasets
• Tips, tricks, and lessons learned throughout!
Federal Student Aid-based data

• FSA collects data on the 70% or so of students who get federal financial aid for college
  – Lowest coverage at community colleges, highest coverage at for-profits
• Can provide information on student outcomes after leaving college
• But...the data don’t always play nicely with IPEDS!
UnitIDs versus OPEIDs

- IPEDS uses UnitIDs as the unit of analysis
- FSA uses OPEIDs—which are based on program participation agreements with ED
- Some systems all report under one main OPEID, while others have separate ones
- Parent/child reporting issues also a concern with IPEDS finance data
Identifying parent/child issues

- Look at the last two digits of the 8-digit OPEID
- Ends in 00: Parent institution (80% of colleges)
- Ends in 01-99: Child institution
- Especially a concern among for-profit colleges and certain public university systems
Parent/child examples

Table 5.1. Examples of UnitID-OPEID Crosswalks and Parent-Child Issues

<table>
<thead>
<tr>
<th>Name</th>
<th>OPEID</th>
<th>UnitID</th>
<th>Parent-Child Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio State University—Columbus</td>
<td>00309000</td>
<td>204796</td>
<td>Yes (parent)</td>
</tr>
<tr>
<td>Ohio State University—Lima</td>
<td>00309001</td>
<td>204671</td>
<td>Yes (child)</td>
</tr>
<tr>
<td>Ohio State University—Mansfield</td>
<td>00309002</td>
<td>204680</td>
<td>Yes (child)</td>
</tr>
<tr>
<td>Rutgers—New Brunswick</td>
<td>00262900</td>
<td>186380</td>
<td>Yes (parent)</td>
</tr>
<tr>
<td>Rutgers—Camden</td>
<td>00262901</td>
<td>186371</td>
<td>Yes (child)</td>
</tr>
<tr>
<td>Rutgers—Newark</td>
<td>00262902</td>
<td>186399</td>
<td>Yes (child)</td>
</tr>
<tr>
<td>Indiana University—Bloomington</td>
<td>00180900</td>
<td>151351</td>
<td>No</td>
</tr>
<tr>
<td>Indiana University—East</td>
<td>00181100</td>
<td>151388</td>
<td>No</td>
</tr>
<tr>
<td>Indiana University—Kokomo</td>
<td>00181400</td>
<td>151333</td>
<td>No</td>
</tr>
<tr>
<td>University of Wisconsin—Madison</td>
<td>00389500</td>
<td>240444</td>
<td>No</td>
</tr>
<tr>
<td>University of Wisconsin—Milwaukee</td>
<td>00389600</td>
<td>240453</td>
<td>No</td>
</tr>
<tr>
<td>University of Wisconsin—Green Bay</td>
<td>00389900</td>
<td>240277</td>
<td>No</td>
</tr>
</tbody>
</table>

Source: 2016 College Scorecard crosswalk file.
Linking UnitIDs and OPEIDs

• Challenge: Closures, mergers, acquisitions, and new branch campus make things messy
• Important to research the context of institutions you are studying
• College Scorecard provides crosswalks over time as a part of its data documentation
Solutions for parent/child issues

• Drop these colleges (especially if they aren’t of interest)
• Aggregate IPEDS data to the OPEID level
• Allocate FSA data based on per-FTE basis or IPEDS resource flag
• Drawback: Some data elements can’t be separated out
College Scorecard

- First announced by President Obama in 2012 as a consumer choice tool
- Modern version was launched in 2015 after a federal college ratings system failed
- Two portals:
  - Public-facing site ([link](#))
  - Data site ([link](#))
Scorecard data downloads

• Three potential options:
  – All data files
  – Most recent institution-level data
  – Most recent data by field of study (new program level data—released fall 2019)

• Caution: Downloads (and resulting Excel files) are quite large and may exceed computer/software package capacity!
Key institution-level metrics

• Lots of IPEDS data also pulled in for you
• Earnings data:
  – Earnings 6-10 years after starting college
  – Percent of students making more than $28,000
  – Broken down by gender, family income tercile, and dependency status
• If cell size is too small, “NULL” or “PrivacySuppressed” are reported
Key institution-level metrics

• Undergraduate federal student loan burdens upon leaving college
  – Mean, median, and percentile distributions
  – Debt by gender, family income tercile, dependency status, completion status (caution!), and first-generation status

• Excludes Parent PLUS, private loans, and grad school debt, so value is limited
Key institution-level metrics

• Repayment rate: defined as share of former borrowers repaying at least $1 in principal
  – Measured at 1, 3, 5, and 7 years after entering repayment
  – Also broken down by student subgroups

• Finally, percent of first-generation students is a useful new measure
About program-level data

- Includes graduate and undergraduate programs
- Graduates only—excludes dropouts
- Done at the four-digit CIP code, which can combine smaller programs into a larger one
- Cell size requirement is about 20 graduates receiving federal aid
Program-level metrics

• Median earnings measured approximately one year after graduation
• Debt at graduation (mean, median, estimated monthly payment)
• Longer-term earnings measures are planned
• Also expect to see a student loan repayment measure
Title IV volume reports

- FSA produces OPEID-level datasets of federal financial aid disbursements (link)
- Includes information on each type of federal grant, loan, and work-study program from 2001 to the present
- Data are presented quarterly for most programs—use the award year summary under Q4
Tips for using Title IV volume reports

- For loan data, use amount disbursed instead of amount originated
- Before 2009-10, combine loans disbursed from FFEL and Direct Loan programs
- Campus-based aid data include the federal award and the total amount disbursed with institutional matches
Federal accountability datasets

• FSA collects a host of accountability information ([link](#)), including the following:
  – Clery Act reports
  – Heightened cash monitoring
  – Cohort default rates
  – 90/10 rule
  – Letters of credit
  – Financial responsibility scores
  – Gainful employment—not updated
Military and veterans’ benefit data

- IPEDS has information on the Post-9/11 GI Bill and DOD Tuition Assistance program in the Student Financial Aid survey from 2013-14 forward
- Department of Veterans Affairs has a dataset ([link](#)) on GI Bill benefits going back to FY 2009
  - But they use a facility code that doesn’t align with UnitIDs or OPEIDs
Research productivity data

- NSF’s Higher Education Research and Development Survey ([link](#)): Information on sources and uses of research funding since 1972
- Survey of Earned Doctorates ([link](#)): Info on new doctoral recipients since 1958
- Go to “microdata”—now can get institution-level data
- Have to reshape data into wide format for analyses
Research productivity data

- Another cool data source: Center for Measuring University Performance at UMass ([link](#))
- Has data on annual giving, postdoctoral appointees, National Merit Scholars, faculty awards, and National Academies members
- But...no UnitIDs are provided
Opportunity Insights data

• Research team got access to IRS data on parental earnings, tuition payments to colleges, and student earnings (link)

• Covers students born 1980-1991 and tracked through 2014

• Focal measure: Social mobility rates by income quintile

• Can also track marriage rates by cohort—neat!
Opportunity Insights tips and tricks

• Data are reported by OPEID or super OPEID (“University of Wisconsin System” and “Certain Colorado Community Colleges”)
  – 2,143 unique UnitIDs
  – 222 OPEIDs
  – 96 super OPEID clusters
  – Table 11 of their dataset provides info—use the “multi” variable to identify super OPEID flags

• Will the dataset ever be updated again?
State and local data sources

• Important to provide context about how colleges operate
• Can help with identifying comparison institutions for benchmarking
• Can merge onto IPEDS using state/FIPS or county codes
State higher ed finance data sources

• State Higher Education Executive Officers Association’s SHEF survey ([link](#))
• National Association of State Student Grant and Aid Programs annual survey ([link](#))
• Can merge this onto other data, such as from the Census, to get measures of funding effort per young adult/adult over time
Other useful data sources

• Overarching state data source: Correlates of State Policy Project (link)
  – But only updated through 2016
• Unemployment rates (Bureau of Labor Statistics)
• Median household income (Bureau of Economic Analysis)
• Percentage of residents living in poverty (Census Bureau)
Other useful data sources

• Educational attainment rates, racial/ethnic makeup of state (Census Bureau)
• State political characteristics (National Conference of State Legislatures)
• Bureau of Labor Statistics and the American Community Survey have county-level data, but the ACS only goes back to 2005
Merging IPEDS with other sources

• Check merges each step of the way for errors, duplicate results, and missing data
• Make sure UnitID/OPEID issues have been resolved
• Be careful for college mergers, consolidations, and name changes
• Always check descriptive statistics!
Examples of IPEDS merges

• **Hillman (2015)** merged IPEDS and FSA data with an ED data on accreditors to look at factors associated with high cohort default rates

• **Klasik & Hutt (2018)** merged Center for American Progress data on accreditor actions with IPEDS and College Scorecard data
Merging your own data

• Some research questions require collecting your own data to merge onto IPEDS
• Data may be at the institution level, state level, or some other relevant unit of analysis
• My current project (link) collects data on state performance funding policies merged onto IPEDS and College Scorecard data and state-level control variables
Concluding remarks

• Researchers and practitioners have used IPEDS for years when conducting analyses across colleges
• But the field is moving quickly toward more sophisticated analyses
• Requires IPEDS to be supplemented with other sources, many of which are not as widely used...yet
Concluding remarks

• For many questions, program-level data will become increasingly important
  – College Scorecard data
  – Earnings data from the Census—becoming available for more colleges
  – State longitudinal data systems
  – Can even use data collected by sources like US News

• IPEDS data still play an important role for institutional context
Thank you!

• Feel free to contact me with any questions or for additional examples of data merges

• robert.kelchen@shu.edu