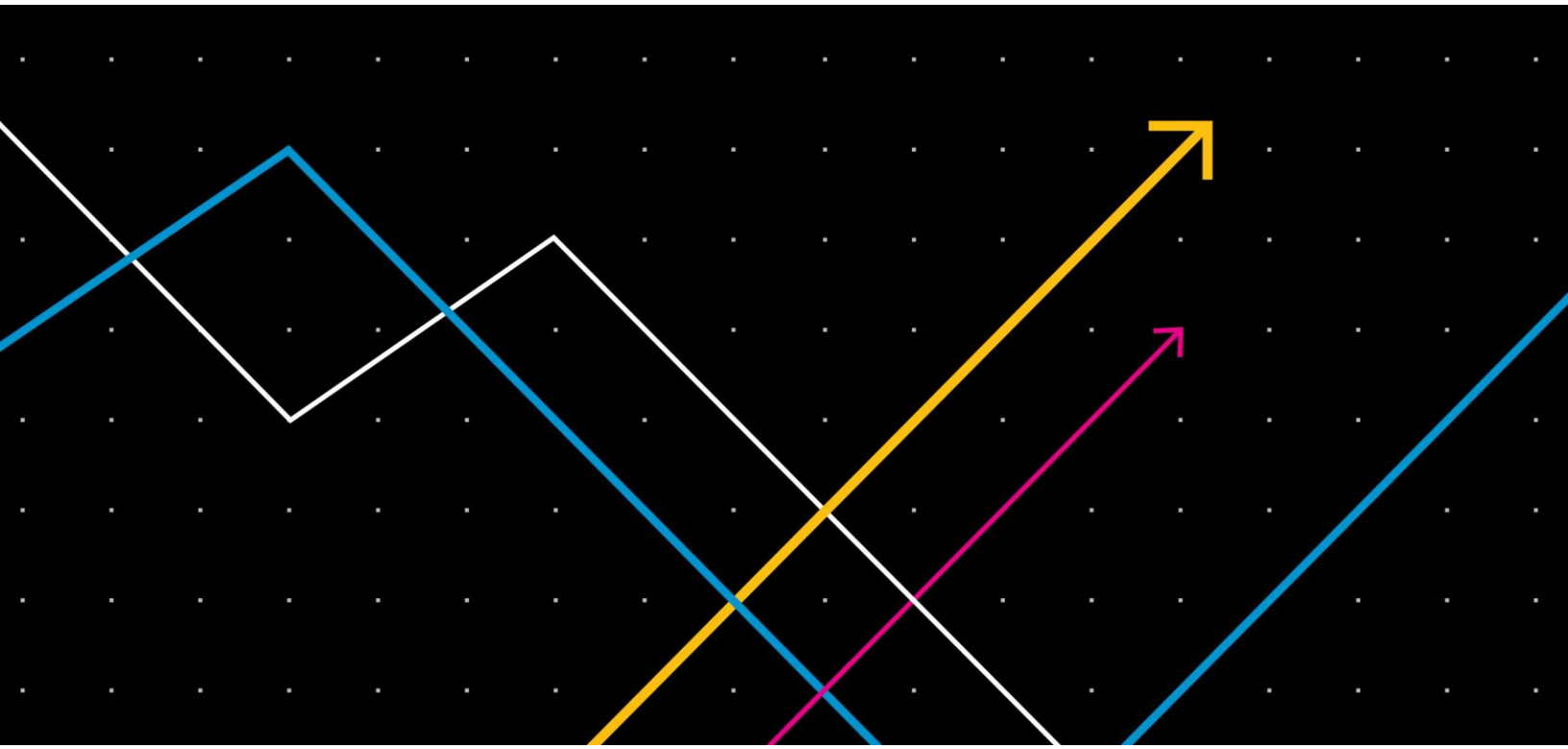


RESEARCH REPORT

The Feasibility of Program-Level Accountability in Higher Education

Guidance for Policymakers

Kristin Blagg
URBAN INSTITUTE

Erica Blom
URBAN INSTITUTE

Robert Kelchen
SETON HALL UNIVERSITY

Carina Chien
URBAN INSTITUTE

February 2021

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

This report was supported by Arnold Ventures. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors wish to thank David Hinson for editorial support and Matthew Chingos for reviewing earlier versions of this report.

Executive Summary

Some policymakers are considering federal accountability in higher education at the program level, in place of, or in addition to, accountability at the institution level. A "program" can be thought of as a major or a field of study, such as cosmetology or chemistry. A program-level accountability approach has some advantages. It may be easier to hold a poor-performing program accountable (e.g., by issuing sanctions or denying federal aid), rather than an institution. And accountability measures could be more tailored to the level of the credential and the labor market (e.g., the outcome metric for a graduate program could be higher than for an associate degree).

But program-level measures are difficult to develop. Program definitions must be large enough to be assessed but distinct enough to be meaningful, and any measure should be insulated against potential gaming. In this report, we assess the ways program-level data may be developed and assessed for accountability. We arrive at the following recommendations:

Set a minimum program size. Program-level metrics must assess the typical quality of a program. The outcomes of a small cohort of students—perhaps 10 or 15 graduates—may not be enough. A size of 30 (*n*-size) is typically cited as the goal for a measure that best represents the performance of a population (in this case, the population may be all program graduates over time). But smaller *n*-sizes could be used, especially if the accountability criteria depend on multiple "failures" (e.g., not meeting an earnings threshold across two or three consecutive cohorts).

Pool two years of data. One way to increase *n*-size is to pool years of data. We find that the share of programs and students included at a given *n*-size increases substantially when an additional year of data is added. Adding third and fourth years includes more programs but to a smaller degree. Given the trade-off between timeliness of data and development of sufficiently large cohorts, we recommend pooling two years of data.

Use iterative categorization of programs within credential levels. The most discrete level of a program is at the six-digit Classification of Instructional Programs (CIP) code level. Two- and four-digit CIP codes aggregate more specific programs and thus include more students. But relying on these broader program categories may hide certain programs' poor outcomes. We suggest using an iterative approach, rolling up programs within a credential level until a given *n*-size is achieved. This approach is not as useful for prospective students (because a program may be reported at the six-digit CIP code level at one institution and at the two-digit level at another), but this approach ensures that all student

outcomes are included in an accountability measure and allows for targeting of corrective measures, if needed.

Monitor accountability measures over time. Any accountability measure is subject to gaming. Institutions could avoid being subject to a given metric, for example, by reducing program sizes or by reclassifying failing programs. Policymakers should monitor their selected measures to ensure that institutions do not evade accountability.

The Feasibility of Program-Level Accountability in Higher Education

In recent years, policymakers have expressed increased interest in program-level, rather than institution-level, higher education accountability measures. Evidence shows that what you study matters as much as, if not more than, where you study (Altonji, Blom, and Meghir 2012; Webber 2016). Measures of higher education performance are beginning to reflect this reality. The College Scorecard, for example, recently published debt and earnings data by program of study (though these are not used for accountability), and the now-defunct Gainful Employment initiative aimed to assess program-level offerings.

But program-level accountability is not straightforward: How should a "program" be defined for accountability purposes? How might measures that hold an institution's program accountable (e.g., affect the institution's ability to receive Title IV funding) differ from informational measures used to guide student choices? In this report, we lay out criteria for defining a program, assess program definitions, and provide recommendations for policymakers seeking to enact program-level accountability metrics.

Criteria for a Valid Program-Level Metric

For a national program-level accountability regime to be successful, a "program" must meet certain criteria. In particular, the accountability regime must do the following:

- **Include all eligible students in the program measure.** Program-level accountability schemes must account for the outcomes of all eligible students (e.g., all students with federal aid or all students who have completed a degree).
- **Avoid statistical noise and allow for privacy suppression.** Depending on the metric and agency, current program-level measures are often suppressed for groups with fewer than 10 or 20 students for privacy reasons. Because these measures will be used to hold programs accountable, measurement must be stable over time. An "unlucky draw" of a handful of poor-performing students should not doom an otherwise strong small program. Statisticians often suggest a group of at least 30 members for stability, and this threshold has been supported by analyses of program-level graduation data in Virginia (Blagg and Rainer 2020).

- **Provide meaningful metrics.** A program should include students who are similar enough such that the metric represents the group's experiences. For example, students in an undergraduate certificate program should be reported separately from students pursuing other undergraduate or graduate credentials in the same field, and biology programs should be reported separately from foreign language programs.

We use these principles to guide our analyses and recommendations. To assess the best approach for program-level metrics, we look at two commonly used strategies: pooling similar programs together by subject and pooling years of data. We also discuss other considerations (e.g., the potential for gaming a given metric) that should be incorporated into decisionmaking around these program-level measures.

Pooling Programs by Subject Matter

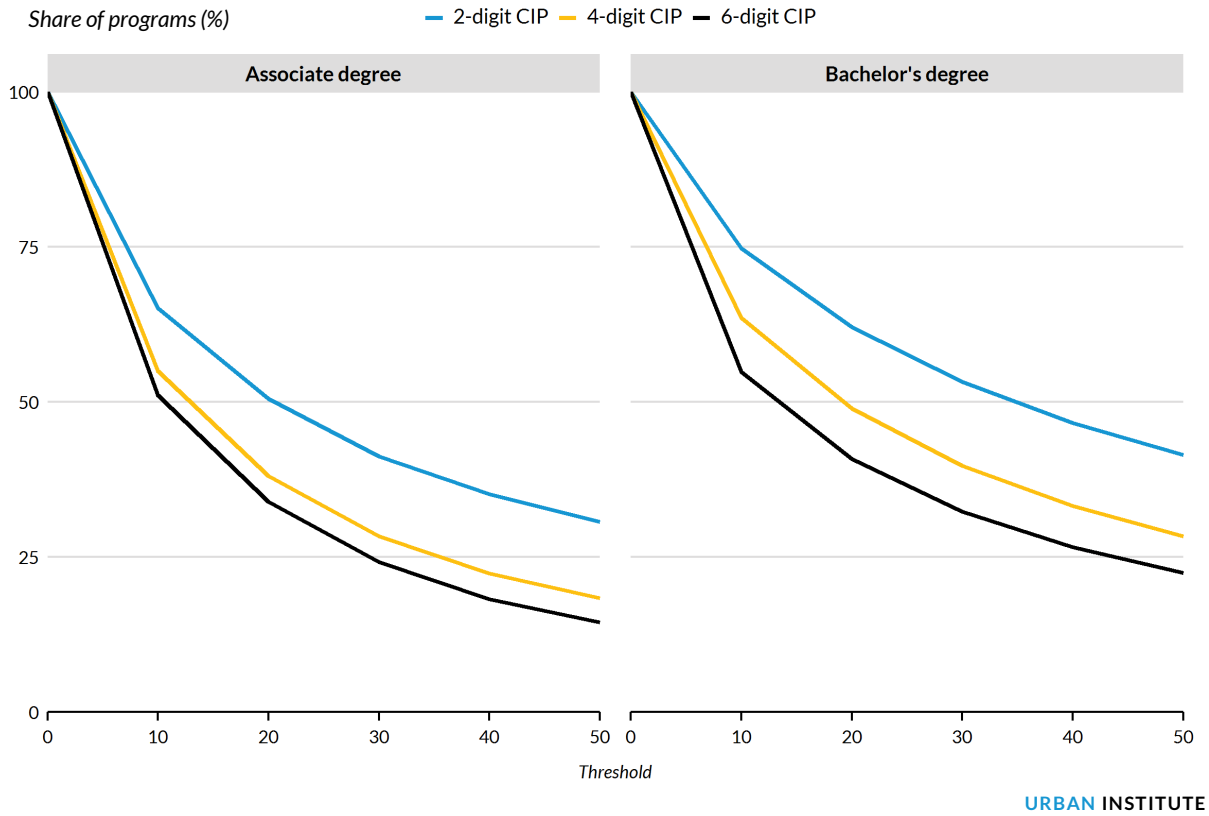
Higher education programs are classified by a taxonomic system called the Classification of Instructional Programs (CIP). Each academic program is assigned a CIP code. The most detailed classification is the six-digit code, which is often familiar to students as a "major" (e.g., entomology is CIP code 26.0702). These six-digit codes are grouped into broader four-digit codes (zoology/animal biology, 26.07) and two-digit codes (biological and biomedical sciences, 26). When we analyze higher education programs, we label each credential at each institution, within a given CIP code group, as a program (e.g., a bachelor's degree in zoology/animal biology and a research doctorate in zoology/animal biology, even if offered at the same institution, are distinct programs). The US Department of Education's Integrated Postsecondary Education Data System (IPEDS) tracks the number of graduates and their programs of study, regardless of whether students receive federal financial aid. IPEDS reports data on 253,093 programs at the six-digit CIP level within the 2017–18 and 2018–19 completions datasets. The number of programs listed in IPEDS completions data falls to 197,309 when measured at the four-digit CIP code level and to 99,274 at the two-digit CIP code level.

Each of these CIP code–level groupings provides a natural definition of "program," but each has trade-offs: six-digit CIP codes best distinguish individual programs (criterion 3) but are often too small to report stable and anonymous metrics (criterion 2). Conversely, two-digit CIP codes are generally large enough to report stable outcomes for most programs but may contain six-digit CIP codes that describe substantially different programs. For example, social sciences (CIP code 45) contains both anthropology (45.02) and economics (45.06).

FIGURE 1

Share of Programs with Certain Numbers of Graduates across Different CIP Rollup Levels

Combines 2017–18 and 2018–19 cohorts



Source: Integrated Postsecondary Education Data System completions data.
Notes: CIP = Classification of Instructional Programs. Combines the two most recent years of data available.

Figure 1 shows the share of programs measured at the six-, four-, and two-digit CIP code level that had at least 10, 20, 30, 40, or 50 graduates during the 2017–18 and 2018–19 academic years (combined) for associate and bachelor's degree programs. For each threshold sample size, the share of programs included in the metric increases when moving from narrow to broad program definitions. In general, the jump is larger when moving from four-digit CIP codes to two-digit CIP codes than from six-digit to four-digit codes. Just over half of associate and bachelor's degree programs had at least 10 graduates over two years, when measured at the six-digit CIP code level. At the two-digit CIP code level, 65 percent of associate degree programs and 75 percent of bachelor's degree programs meet the 10-person threshold. Using a 30-student threshold, only 24 percent of associate degree programs and 32 percent of bachelor's degree programs qualified at the six-digit level compared with 41 percent and 53 percent of programs, respectively, at the two-digit level.

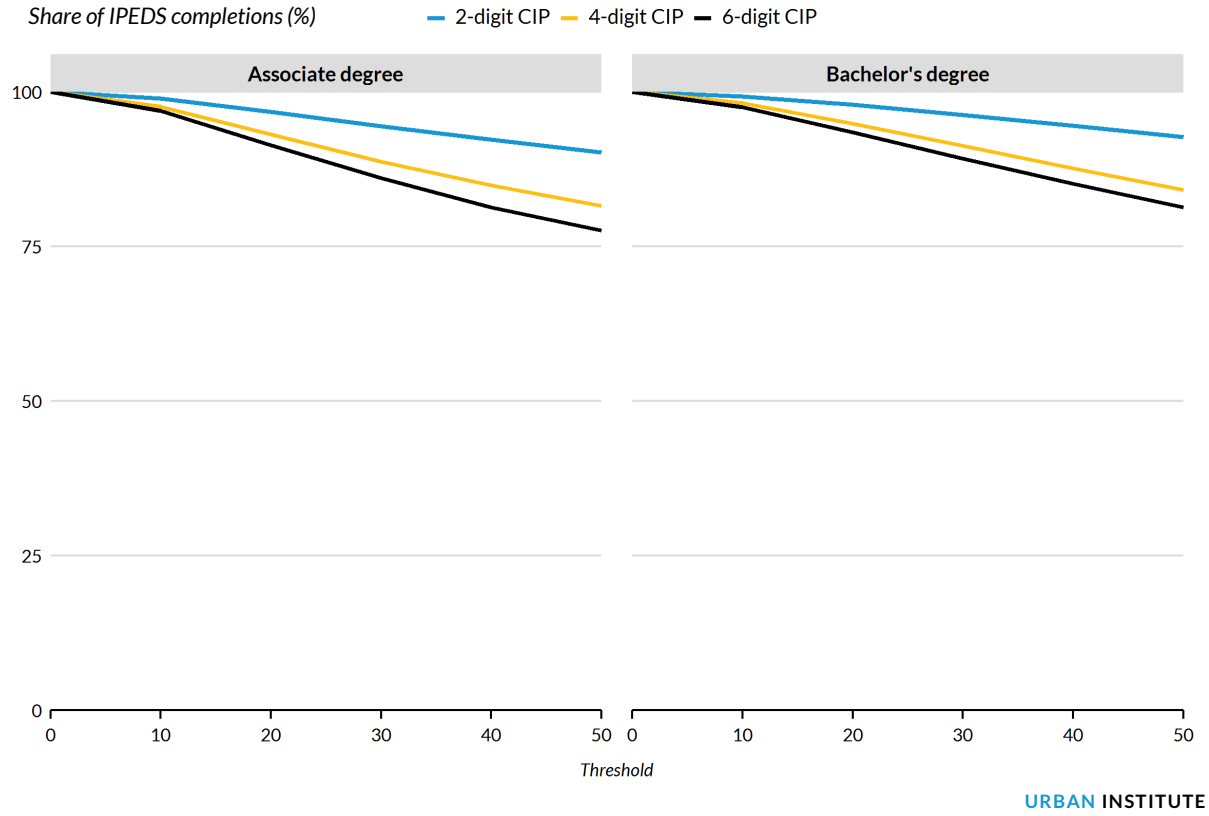
Other credentials show similar patterns (appendix figure A.1). At the six-digit CIP code level, most programs across all credential levels, except professional practice doctorates, have fewer than 20 graduates in a two-year period. The share of covered programs increases slightly when moving to a four-digit CIP code and then increases more substantially using a two-digit CIP code. But even at the two-digit CIP code level, only about half of programs for bachelor's degrees, master's degrees, and practice doctorates reach 30 graduates. And despite this aggregated two-digit level, coverage of individual graduate certificate and research doctorate programs remains below 30 percent.

Of course, not all programs that meet a given threshold have the same number of students. We find similar trends in coverage by credential type when we project the shares of *students* covered at each threshold, rather than shares of *programs* (see figure 2 for associate and bachelor's degrees and appendix figure A.2 for all other credentials). But because most students graduate from large programs, the share of students participating in a measured program is far higher than the share of programs. Given a threshold of 30 graduates, about 90 percent of students are included in programs measured at the six-digit CIP code level, for nearly all credentials. Approximately 95 percent of students graduate from programs with at least 30 students in the two-digit CIP code over a two-year period. Even though only a small share of graduate certificate and research doctorate programs have at least 30 graduates at the two-digit CIP code level, more than three-fourths of all graduates at these programs are covered. This analysis shows that the outcomes of most students would be easily covered by a program-level accountability framework at any CIP code level, even if many very small programs are excluded.

FIGURE 2

Share of Graduates in Programs with Certain Numbers of Graduates across Different CIP Rollup Levels

Combines 2017–18 and 2018–19 cohorts



Source: IPEDS completions data.

Notes: CIP = Classification of Instructional Programs; IPEDS = Integrated Postsecondary Education Data System. Combines the two most recent years of data available.

Understanding Information Loss When Pooling CIP Codes

Two- and four-digit CIP codes include more students, and more programs, relative to the more detailed six-digit CIP codes. But some information is lost when relying on broader program categories. For example, anthropology (45.02) and economics (45.06) are within the same two-digit CIP code. If they were combined, the resulting metric might not yield meaningful accountability if students in these programs have disparate outcomes. For example, if earnings are low (or debt levels are high) in one six-digit program but not the other, these outcomes could be masked by the better performance of other programs included in the two-digit CIP code. The median graduate of a bachelor's degree program in anthropology earned $23,900 one year after graduation, while the median graduate in economics earned $44,300—higher than the highest-earning anthropology program in College Scorecard data.

To explore the prevalence of this phenomenon, we examine two-digit CIP codes at institutions that host multiple four-digit CIP code programs. Using College Scorecard data that include all graduates who received federal Title IV financial aid, we look at reported median student debt and earnings for graduates of four-digit CIP code programs housed within the same institution, degree level, and broader two-digit CIP code. We calculate the range of median debt (and earnings) within each two-digit CIP code at each institution by subtracting the lowest debt (or earnings) from the highest. We then take the average across all institutions and two-digit CIP codes (table 1).

TABLE 1

Variation in Debt and Earnings among Four-Digit CIP Codes, by Credential Type

	Undergraduate certificate	Associate degree	Bachelor's degree	Master's degree	Professional doctorate
Median debt					
Mean value	11,097	16,066	22,966	42,725	164,700
Mean range within program	4,872	4,791	4,065	19,434	90,101
Number of observations	447	843	4,081	1,369	94
Median earnings					
Mean value	26,743	40,292	40,149	60,297	78,221
Mean range within program	9,398	15,409	12,091	21,657	44,037
Number of observations	567	843	3,323	1,291	86
Debt-to-earnings ratios					
Mean value	0.410	0.443	0.634	0.751	2.183
Mean range within program	0.141	0.193	0.231	0.397	1.763
Number of observations	387	644	3,191	1,125	86

Source: Authors' calculations using College Scorecard data.

Notes: CIP = Classification of Instructional Programs; OPEID = Office of Postsecondary Education Identification. The programs are counted as being offered at the OPEID level, even though some OPEID codes include offerings in the same CIP code across multiple unit IDs. The sample size refers to the number of two-digit CIP codes across all OPEID codes. These data include only 2015–16 graduating cohorts. Graduate certificates and research doctorates are excluded because of the small number of observations (fewer than 50 in most cases).

We find that the range in outcomes within two-digit CIP codes can be quite large. For example, although the typical median earnings for bachelor's degree holders (among two-digit CIP codes with multiple four-digit CIP codes) is about $40,000, the range of typical earnings at the four-digit level is about $12,000—nearly a third of the median earnings. Debt-to-earnings ratios vary considerably with two-digit CIP codes, with the largest amount of variation at the graduate level, where students can borrow up to the full cost of attendance. For example, the average master's degree program had median debt that was 75 percent of median earnings. But the variation within four-digit CIP codes was 40 percent, meaning that some four-digit CIP codes had much higher debt-to-earnings ratios than others. These results suggest that accountability metrics based on two-digit CIP codes rather than four-digit CIP codes could obscure meaningful variation across programs. (Note that this analysis includes only

programs and institutions with multiple four-digit CIP codes within a two-digit CIP code, which is 42 percent of all programs. See appendix table A.1 for more detail.)

The Accountability Measure Matters When Pooling CIP Codes

Aggregating programs at higher CIP code levels yields a more inclusive accountability measure but masks meaningful variation in program outcomes. How much the aggregation obscures poor-performing programs, however, would depend on the actual accountability metric used. For example, earnings metrics are not currently used for accountability but are reported on College Scorecard as medians to give potential students a sense of a given program's benefits. But an earnings accountability metric likely would not be based on mean or median earnings. Students might find the differences between a program that yields $60,000 earnings and one that yields $40,000 earnings to be meaningful differences (which they are), but those differences may not matter to policymakers, who might be interested only in whether earnings are "good enough" to justify the provision of federal student aid. Instead, an accountability measure might be based on threshold earnings (e.g., the share of students earning above 150 percent of the federal poverty level for a single adult). Our analysis cannot assess whether those values differ substantially within a two-digit CIP code.

Debt-to-earnings ratios have also been proposed for accountability purposes, namely to assess whether programs allow graduates to attain "gainful employment." The Obama administration's gainful employment regulations, which were rescinded by the Trump administration, used the six-digit CIP code as the unit of analysis for evaluating nearly all programs at for-profit colleges and certificate programs at public and private nonprofit colleges. These regulations required at least 10 students to report outcome data. With program-level College Scorecard data, we can also look at variation in debt-to-earnings outcomes, but we are limited to assessing sufficiently large four-digit CIP code programs aggregated within two-digit CIP codes. Even within this limited analysis, we find that aggregating to the two-digit CIP code obscures a great deal of variation in debt-to-earnings ratios. This variation is largest for graduate programs, where students tend to accumulate larger federal loans. This suggests that two-digit CIP codes might be too broad a measure to meaningfully distinguish differences between programs.

Pooling Years of Data

Another approach to resolving the issue of small program sizes within a given year and six-digit CIP code is to expand the number of years included in a measure. Using graduation data from IPEDS, we look at how pooling up to four years of data would affect the share of six-digit CIP code programs that would meet certain *n*-size thresholds. For example, a program that graduates 15 students annually might have 30 students in a two-year pooled sample and 45 students in a three-year sample. Still, programs that graduate very small cohorts—perhaps two or three students a year—are not likely to reach this threshold, even when pooling four years of data.

Additional Years of Data Tend to Have Diminishing Returns

When we pool years of data together, we find that the share of programs included rises substantially for each *n*-size threshold (see figure 3 for bachelor's and associate degrees, and see appendix figure A.3 for all credentials). But even with four years of data, the number of included programs can still be far from the full number of available programs.

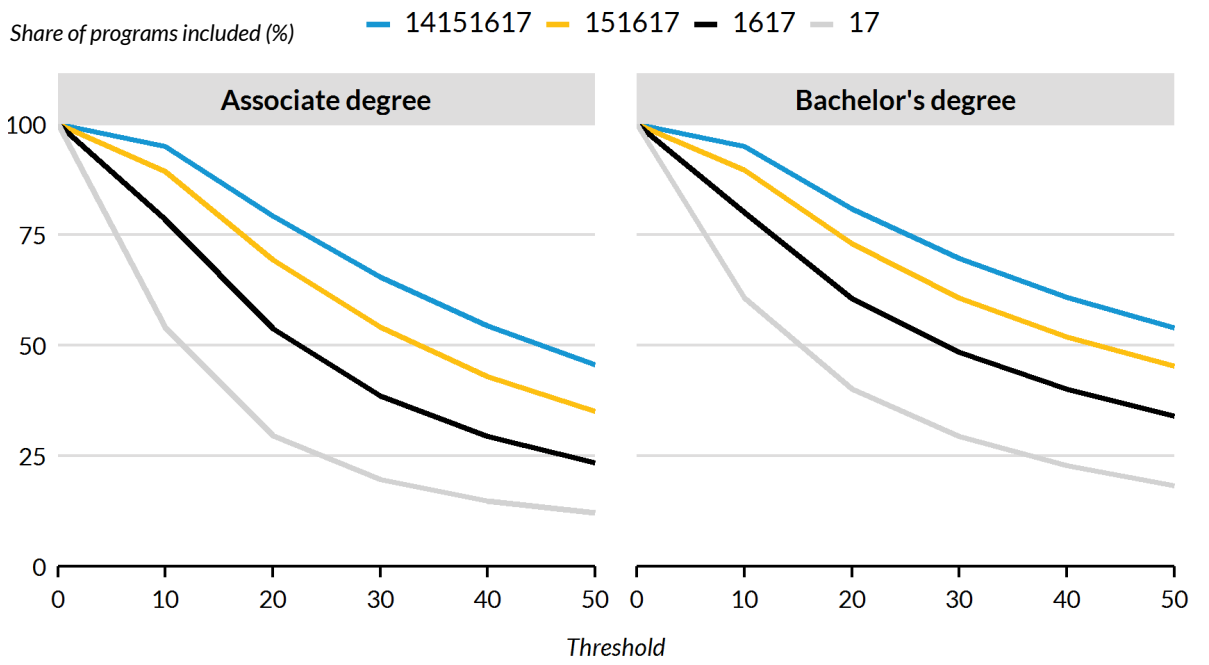
A sample size of 30 is commonly cited as the threshold at which a measure might stabilize or provide an estimate consistent with a program's performance over time. In previous work with Virginia data, we found that estimates of program graduation rates tend stabilize at an *n*-size of around 30 students (Blagg and Rainer 2020). Assessing a single year of 2017 completions data against a threshold of 30 graduates, we found that just 29 percent of six-digit CIP code bachelor's degree programs and 20 percent of six-digit CIP code associate degree programs would meet this threshold. In fact, only programs that offer practice doctorates (typically, programs offering an MD or JD) are large enough to include more than 50 percent of programs offered in 2017.

When we add a second year of data (2016), the share of included six-digit CIP code programs nearly doubles for many completion levels. Forty-eight percent of bachelor's degree programs and 39 percent of associate degree programs are included. Adding a third year (2015) and fourth year (2014) of data increases the share of programs available for measurement at an *n*-size of 30 but in incrementally smaller amounts (61 percent of bachelor's degree programs at three years and 70 percent at four years; 54 percent of associate degree programs at three years and 66 percent at four years).

FIGURE 3

Share of Programs, by Rule, for Each Credential Level

Six-digit CIP codes



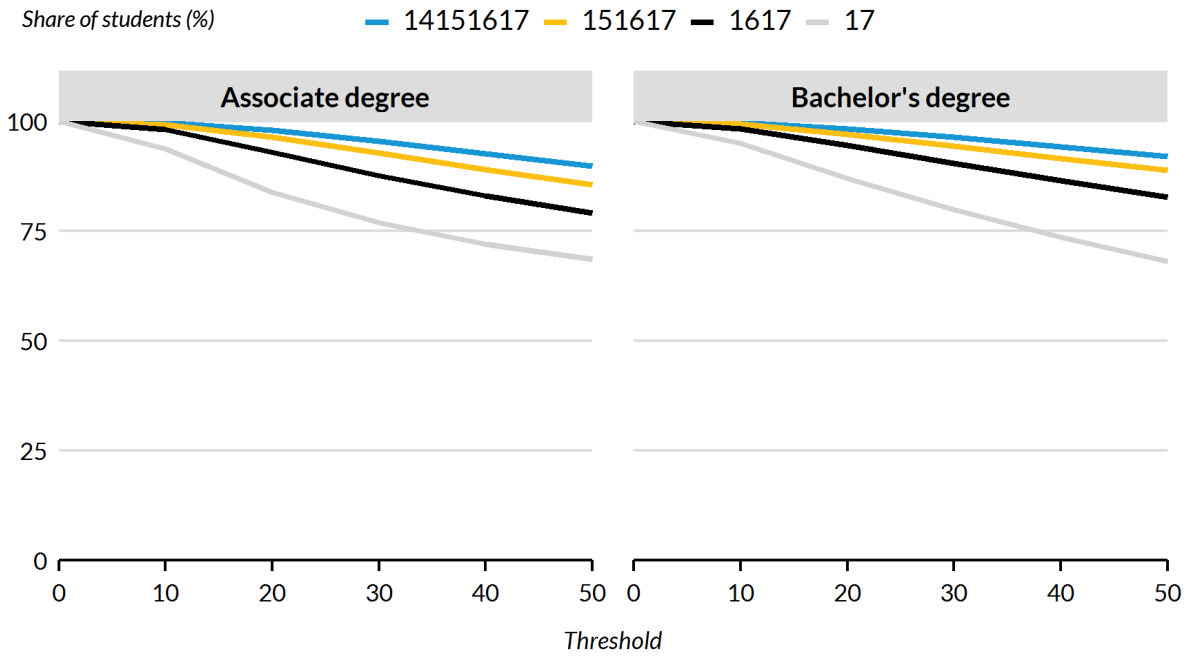
Source: Integrated Postsecondary Education Data System completions data.
Notes: CIP = Classification of Instructional Programs.

But similar to our CIP code analysis, we find that although a substantial share of *programs* are not included—even in a measure that rolls up four years of data—most students are included in these year-pooled measures (figure 4 and appendix figure A.4). Because the excluded programs graduate few students by definition, the share of students overall who are not included at an *n*-size of 30 is low, especially as years are pooled together. For example, 80 percent of bachelor's degree graduates and 77 percent of associate degree graduates are included in a single year of data. This share rises with two years of pooled data (90 percent of bachelor's degree graduates and 88 percent of associate degree graduates), three years (94 percent and 93 percent), and four years (96 percent and 96 percent).

FIGURE 4

Share of Students, by Rule, for Each Credential Level

Six-digit CIP codes



Source: Integrated Postsecondary Education Data System completions data.
Note: CIP = Classification of Instructional Programs.

Pooling years of data does have disadvantages, especially for developing a performance metric for accountability. First, although most programs are included when we pool four years of data, not all programs are included. Even with four years of data, graduate certificate programs and research doctorates (typically, academic PhD programs) still do not have large shares of programs or students that would meet the 30-graduate threshold. And a four-year measure means this metric would be slow to change, as any improvements would be averaged into older data.

Although we look only at the number of students included in a potential pooled measure, policymakers who pool years of data will have to make additional choices about how individual cohorts' outcomes are combined. For example, how should a program that graduates 5 students in one year, 10 the next year, 20 the next, and 40 the next have their four-year pooled estimate compared with a program that graduates 15 students a year? If data on earnings, for example, are pooled together without regard for cohort timing, the former program might have an advantage if the economy has improved (or have a disadvantage if the 40-person cohort graduated into a recession). Policymakers

therefore need to decide whether to weight cohorts by the number of students in each cohort—which might be preferable absent dramatic swings in the economy—or to weight each cohort.

Other Concerns for Program-Level Accountability

By pooling years and CIP codes in some combination, federal policymakers will likely be able to develop accountability measures that include most programs and that describe the outcomes of nearly all students who graduate from higher education. But policymakers should also account for other factors when deciding how and whether to implement accountability at the program level.

Program-Level Metrics Often Fail to Include Students Who Do Not Complete Credentials

A fundamental issue with program-level accountability measures is that they are typically reported only for completers. This is out of necessity: many students who leave school before attaining a degree may not have declared a program of study. Blagg and Rainer (2020) show that this is more of a problem at four-year institutions (where students can remain undecided for multiple years and switch majors easily) than at two-year or less-than-two-year institutions (where students typically enroll in a particular program). At community and technical colleges, therefore, program-level metrics that include noncompleters likely could be developed for vocationally oriented fields of study. They are also typically possible at the graduate level, where students are admitted to particular programs.

The extent to which omitting noncompleters biases or colors a program-level measure depends on the metric. Conceptually, an institution should be responsible for its students' outcomes in proportion to the amount of time a student spends there. With some amendments, current metrics could achieve this goal. For example, the cohort default rate (CDR) is based on the share of students who default, and the amount a student borrows is roughly in proportion to the amount of time a student is enrolled. But the likelihood of default does not increase with the amount borrowed, as noncompletion tends to increase delinquency and default. But a CDR metric that is more scalable to time enrolled might instead be the share of dollars defaulted on (rather than share of borrowers who default). Similarly, an equivalent to the repayment rate (the share of students who have paid at least one dollar of their principal) might be the share of total dollars repaid.

In the case of earnings, it is not straightforward to determine whether or how noncompleters should be included. Consider a program that initially enrolls 100 students. Thirty complete the program

and earn $100,000, and 70 drop out (or switch majors) and make the minimum wage. The College Scorecard would report the earnings for such a program as $100,000, even though most students earn only the minimum wage. This is not only misleading for prospective students but uninformative for federal policymakers deciding whether students in a program should continue to receive federal student aid.

If earnings for completers and noncompleters do not differ that much, reporting earnings for completers only is a reasonable approach. But we know this is not the case. Our analysis of data from the Beginning Postsecondary Students Longitudinal Study indicates that earnings differ substantially depending on whether a student completed their program. Table 2 shows 2017 earnings for selected majors based on intended majors in 2011 ("All") and completed majors ("Completers").

TABLE 2

Earnings for All Enrollees Compared with Completers

	Associate Degree			Bachelor's Degree		
	All	Completers	Diff.	All	Completers	Diff.
Computer and information sciences	28,000	24,000	-14%	34,000	56,500	66%
Engineering and engineering technology	30,000	33,600	12%	46,000	62,400	36%
History	‡	‡		24,000	30,000	25%
Personal and consumer services	23,163	27,000	17%	27,300	28,000	3%
Manufacturing, construction, repair, transportation	36,000	40,000	11%	‡	‡	
Health care fields	23,040	26,000	13%	28,000	39,000	39%
Business	24,000	25,000	4%	36,000	42,000	17%
Education	21,320	‡		30,450	31,500	3%
Public administration and human services	20,000	‡		26,000	24,500	-6%
Design and applied arts	23,400	22,000	-6%	25,000	32,000	28%

Source: US Department of Education, National Center for Education Statistics, 2012/17 Beginning Postsecondary Students Longitudinal Study.
Note: Daggers indicate insufficient sample size.

For example, median earnings for computer and information sciences bachelor's degree completers were about $56,500, whereas earnings for all students who intended to study computer and information sciences were only $34,000. This indicates that lower-earning students switched out or stopped out of the major (or that students with higher earnings potential switched in).

The substantial differences in earnings between prospective and actual majors, in some cases, indicates that the question of whether to focus only on completers in an earnings metric is important for policymakers to consider. An institution should be accountable in proportion to how much time a

student spends there. It is reasonable to expect that an institution should be accountable for its graduates' earnings, and perhaps it should be largely accountable for students who complete all but one credit needed for a particular program. After all, the student has completed most program requirements. At the opposite extreme, a student who enrolls for only one course should probably not be included in an earnings metric for that program.

One potential approach is for a program to be accountable for every student's earnings, weighted in proportion to the share of the program that the student completed. For example, earnings for a student who was present for only one year of a program that is typically completed in four years might receive only a quarter of the weight of a completer's earnings. But this approach is somewhat opaque. A more straightforward approach might be to require a minimum level of enrollment necessary to be included in an earnings metric (e.g., a certain number of credits completed). Another approach would be to incorporate program-level completion rates into an earnings accountability metric. For example, a measure could assign noncompleters an assumed basic earnings level for a given credential (e.g., $10,000 for those dropping out of bachelor's degree programs) that would be averaged with the earnings of completers.[1] This would mechanically down-weight earnings in proportion to the share of noncompleters.

Any Program-Level Metric Could Be Subject to Gaming

As with prior federal accountability measures, a new program-level accountability metric may be subject to gaming, where institutions sidestep accountability by changing in response to a new metric. Campbell's Law suggests that any metric will induce gaming as colleges seek to improve their performance without making substantial operational changes (Campbell 1979). For example, researchers have found evidence of gaming with the cohort default rate, performance-based funding systems, and college and law school rankings (Kelchen 2018).

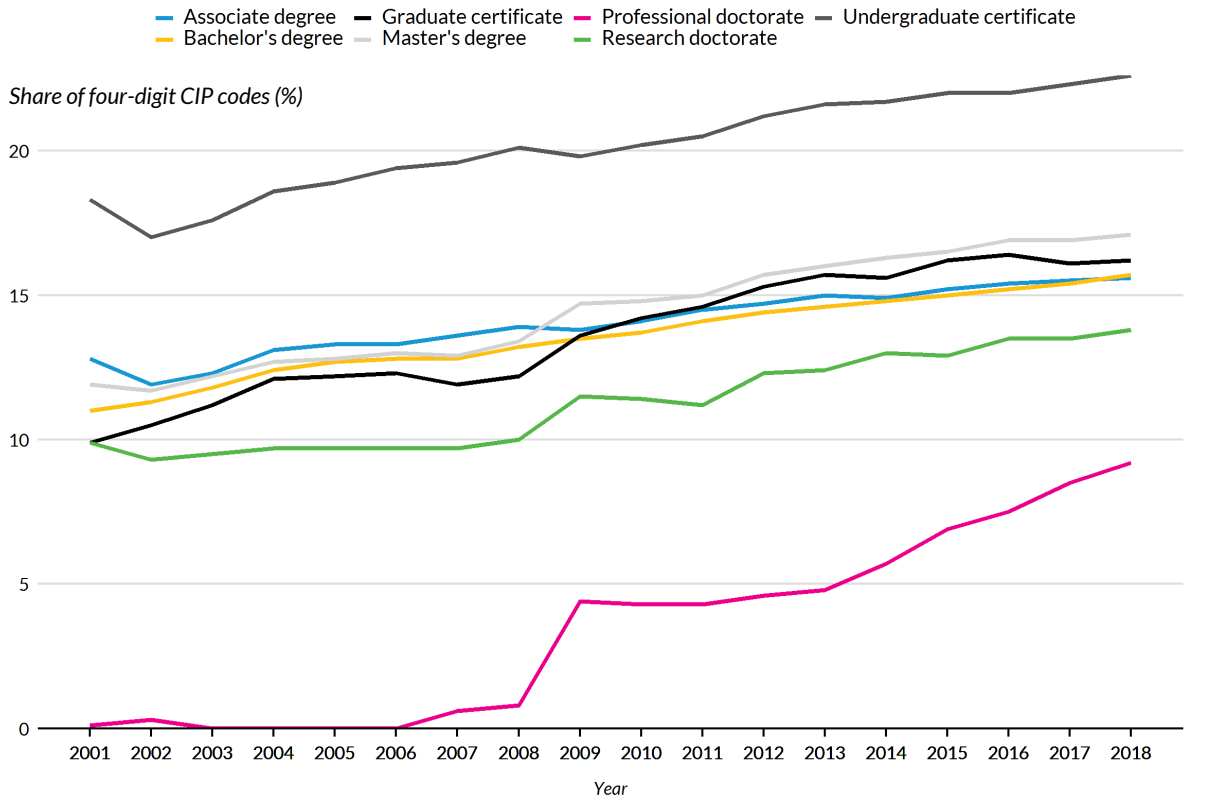
Gaming of program-level accountability metrics is likely to occur. For example, if data are reported only for programs of at least 30 students, schools might cap programs at 29 students or divide large programs into smaller ones under different CIP codes. But this is a less precise effort when it is based on the number of program completers instead of the number of entrants because colleges cannot perfectly predict completion rates.

The principal risk of gaming in program-level accountability is by redefining programs to make them smaller to avoid scrutiny. Figure 5 shows trends in the share of four-digit CIP codes by credential level that contain multiple six-digit CIP codes. There has certainly been a trend in this direction over the past

few decades, with the share of programs with multiple six-digit CIP codes rising from 13 percent in 2001–02 to 17 percent in 2018–19. This increase is likely caused by the increase in more specialized credential programs because colleges have not faced an incentive to create small programs to evade accountability. But the general trend of colleges creating new programs of study (which may be reversed by the pandemic) threatens to create further measurement issues unless some outcomes are aggregated to a higher CIP code.

FIGURE 5

Share of Four-Digit CIP Codes Containing Two or More Six-Digit CIP Codes



Source: Integrated Postsecondary Education Data System data.
Note: CIP = Classification of Instructional Programs.

Aggregating some outcomes to a two-digit or four-digit CIP code would also guard against colleges rebranding programs under a new six-digit CIP code in an attempt to continue operating. It is important to scrutinize programs that fail accountability systems to make sure they do not change CIP codes to try to continue operating.

A cursory analysis of program size in relation to earnings and debt outcomes suggests that poorer outcomes at the four-digit and two-digit CIP code level are generally not more present in larger or smaller programs, on average. Moreover, our proposed schema—described in the following section—includes accountability for all students, so that programs too small for program-level accountability are nonetheless held accountable as part of a larger program or at the institution level. Further, not all institutions are "bad actors," and accountability regimes can and often do work: for example, Harvard University shuttered its theater program after it failed gainful employment regulations, and research shows that low-performing programs were more likely to close following the release of the gainful employment data (Kelchen and Liu 2019).[2]

Recommendations for Program-Level Accountability

The preceding sections describe how aggregating CIP codes and years of data into different program definitions can increase the coverage of programs and students at minimum *n*-size thresholds. Rolling six- or four-digit CIP codes up into two-digit CIP codes covers more programs and students but at the expense of combining disparate programs' outcomes. Similarly, rolling up multiple years of data allows more programs and students to be included but at the expense of recent relevant information. And any measure carries additional risks of gaming by institutions, which should be carefully guarded against. Here, we discuss our findings and provide recommendations for policymakers.

Setting Minimum Program Sizes Reduces Volatility in Accountability Measures

Figuring out the minimum number of students needed to report out data is difficult. Allowing for a smaller number of students creates volatile data that could inadvertently identify a program as being low performing if just a few students had poor outcomes. But requiring a larger number of students to assess program-level results would cover fewer programs or aggregate less-similar programs together.

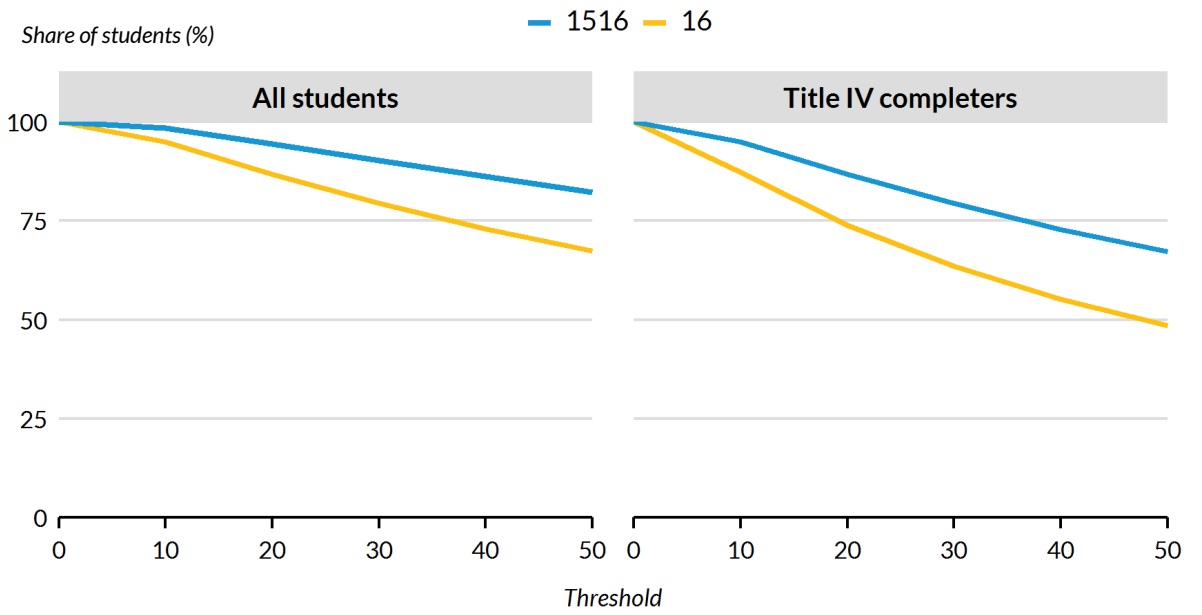
Most federal program-level data sources in higher education use thresholds of 10 or 20 students to report outcomes information for consumers. These *n*-sizes are lower than the recommended 30-person threshold for more stable estimates, but they allow for more programs to be reported upon. A lower threshold size may be feasible for high-stakes accountability purposes if sanctions require failing in multiple years, as this rule reduces the likelihood of a program being sanctioned because of a small number of students having anomalously poor outcomes.

One way to more easily achieve these minimum sample sizes at Title IV–eligible institutions could be to collect data for all students, regardless of Title IV status (typically only students receiving Title IV aid are included in accountability metrics and other metrics reported in the College Scorecard).[3] In our analyses of IPEDS data (figures 1 through 4), we included all students, slightly inflating the share of students and programs that could be included at different thresholds, relative to Title IV students only. Figure 6 shows what the share of students included in a measure might look like for bachelor's degree programs if only Title IV students are included, based on six-digit CIP codes and pooling 2015–16 data. These results are imperfect because we do not know the number of Title IV completers at the six-digit CIP code level. We assume that Title IV students are evenly distributed across all programs.

FIGURE 6

Share of Students, by Inclusion Rule, Rolling 2015 and 2016

Six-digit CIP codes



Source: Integrated Postsecondary Education Data System completions data.
Note: CIP = Classification of Instructional Programs.

Pooling Two Years of Data Increases Coverage of Programs

Regardless of final threshold size, pooling years of data is likely necessary to build program-level accountability measures. As our analyses indicate, there is a large gain in terms of programs and students covered when going from a single year to two years but smaller gains from adding subsequent

years. As such, we recommend two years of data for each metric, as the College Scorecard already does. This also provides a good balance between measure timeliness and program size.

We considered more complicated alternatives. What if, for example, we use a single year of data for large programs but multiple years for smaller programs, until the minimum sample size is reached? The main difficulty with this approach is that large economic changes, such as a sudden recession, might make it difficult to compare a program's single year of data with another program's multiple years of data. Although our focus is not on comparing programs per se, program-level metrics must be compared with some benchmark, and even though it is possible to have multiple benchmarks for different rollups, this is needlessly opaque and complicated. In short, the trade-offs between additional coverage and complexity did not make this alternative plausible.

CIP Codes Should Be Combined Iteratively

Because we have eliminated the possibility of reaching a minimum sample size for each program through rolling up multiple years, we turn to CIP code rollups. The gaming of accountability metrics by institutions is a real concern. Any accountability regime must cover *all* programs so that institutions do not have an incentive to keep programs below a threshold. But even the most aggregated CIP codes—two digits—do not cover all students at a sufficiently large *n*-size and, moreover, could obscure important differences across four- and six-digit programs.

To resolve this measurement issue, we recommend that CIP codes be rolled up iteratively beginning with six-digit CIP codes. If two years of program data are below the *n*-size threshold at the six-digit CIP code level, this program is rolled up to the four-digit level. If the number of students is too small at the four-digit level, the program is rolled up to the two-digit level. Ultimately, any remaining programs can be rolled up to the credential level, and then to the institution level, if all remaining programs are otherwise too small at the two-digit CIP code level.

This may lead to situations where one program is large enough that reporting it separately leaves one that is too small, so we recommend that programs be rolled up in such a way that no program is excluded for being too small. Table 3 provides an example.

TABLE 3

Example of Iterative Rollups

	School A	School B	School C
12 Personal and culinary services	178	73	63
12.04 Cosmetology and related personal grooming services	113	33	13
12.0406 Makeup artist or specialist	41	31	10
12.0410 Nail technician or specialist and manicurist	30	2	2
12.0407 Hair styling or stylist and hair design	22	0	1
12.0404 Electrolysis or electrology technician	10	0	0
12.0411 Permanent cosmetics or makeup and tattooing	7	0	0
12.0412 Salon or beauty salon management	3	0	0
12.05 Culinary arts and related services	65	40	50
12.0502 Bartending	33	40	50
12.0501 Baking and pastry arts, baker, or pastry chef	12	0	0
12.0504 Restaurant, culinary, and catering management	10	0	0
12.0510 Wine steward or sommelier	8	0	0
12.0506 Meat cutting	2	0	0

We assume that the size requirement is 30 students. In School A, the makeup artist, nail technician, and bartending programs are all large enough to report at the six-digit level. Remaining programs in the 12.04 and 12.05 four-digit codes are rolled up, respectively. In School B, the makeup artist program is large enough to report on its own, but that would leave the nail technician program unreported, so it is rolled up together with the makeup artist program. Bartending is reported on its own. In School C, all four programs are rolled up to the two-digit level, personal and culinary services, because even though bartending was large enough to report at the six-digit level, the three programs under 12.04 were not.

This means that programs will be rolled up in different ways depending on the institution and will therefore not be directly comparable across institutions, but the goal of this rollup scheme is to enable accountability metrics, not to provide information for potential students. This also means that an institution's program definitions might change as program sizes change, but this is a possibility with any program definition. Another challenge to consider is that there may be different sample sizes available for different outcomes. For example, earnings outcomes are typically available for all students receiving federal financial aid, while debt metrics are available only for students who take out federal student loans. This means that students who receive Pell grants and not student loans are in the earnings data, not the debt data.

How Should Program-Level and Institution-Level Accountability Be Combined?

A program-level accountability regime would likely supplement, rather than supplant, an institution-level regime, such as the system currently based on the cohort default rate. This supplementation might work in two ways:

- The same metric could be collected at both the institution and program level but with different thresholds for each. For example, the institution-level CDR threshold might be set at 30 percent, while the program-level rate might be set at a less stringent level, such as 35 percent. This would allow a buffer for small programs that have a few students experience poor outcomes in a certain year.
- Different metrics could be collected at the institution and program levels. For example, because earnings make most sense for completers, an earnings threshold (or debt-to-earnings ratio, similar to the Gainful Employment initiative) could be set for programs, while metrics that can be easily determined for noncompleters could be used at the institution level (e.g., the CDR, the repayment rate, or the share of loan repaid).

Other Recommendations

We recommend that policymakers use a two-year pooled program-level estimate, aggregated at a sufficient minimum *n*-size (e.g., 20 or 30 students) by CIP code level, or at the credential or institution level if programs are not large enough. But our examination of program-level accountability also raised other measurement issues, such as how to measure the outcomes of students typically excluded from program measures and how to account for gaming.

FACTOR GRADUATION RATES INTO ANY PROGRAM-LEVEL ACCOUNTABILITY MEASURE

To address the issue of earnings being typically well defined only for completers, for example, we recommend graduation rates be included with any metric that is otherwise provided only for completers. At the bachelor's degree level, institution-wide graduation rates may be used in place of program-level graduation rates (as these are typically not calculable). At the associate degree level, a combination of institution-level graduation rates and transfer rates could be used for transfer-oriented programs.

SCRUTINIZE PROGRAMS WITH SHARP CHANGES IN PROGRAM DEFINITION OR SIZE

In addition, we recommend that programs or institutions be subject to heightened scrutiny should their program definition or size change too rapidly. Examples include a large change (e.g., greater than 20 percent) in the number of students from one year to the next, a change from an *n*-size above a critical threshold to just below the threshold, or the sudden introduction or removal of a six-digit CIP code within a two-digit CIP code. This is to mitigate gaming that might occur through redefining programs to change the way programs are rolled up.

SECURE STUDENTS' PRIVACY

It is crucial that individual students not be identifiable through consumer information data and program-level accountability data. As policymakers select metrics, they must make sure that individual outcomes cannot be determined. The larger an *n*-size threshold for the development of program-level data, the less likely it is that a single student's outcome could be determined. Opting for larger samples from which to draw accountability measures thus both increases measurement stability and protects student privacy.

Program-level accountability measures can produce targeted improvements in higher education outcomes and accountability. But these accountability measures should incorporate enough participants to accurately capture typical outcomes. By pooling years of data, rolling up program definitions within credentials, and incorporating other safeguards where appropriate, a program-level accountability metric is possible.

Appendix. Methodology

Inclusion Rules

All programs we refer to are institution-specific programs, so that the same program at different institutions can be counted multiple times. A program is "included" if the number of students that graduate from this program reaches some threshold of *n* students. We call this threshold the inclusion rule and define this rule at *n* = 10, 20, 30, 40, and 50 students. For example, at rule *n* = 20, any program that graduates 20 or more students is large enough. Generally, the larger the rule threshold, the more stable and reliable any program-level measure will be. But because increasing the rule threshold will also exclude more programs, having a rule threshold that is too large can limit the scope of programs included. To understand how much increasing the rule threshold will limit the number of programs that can be included, we collapse the number of program-institution observations to the credential-year level and calculate the share of programs included. To increase the number of programs included without compromising the stability provided by a larger inclusion rule, an alternate approach was to roll years together.

Rolling Up Years

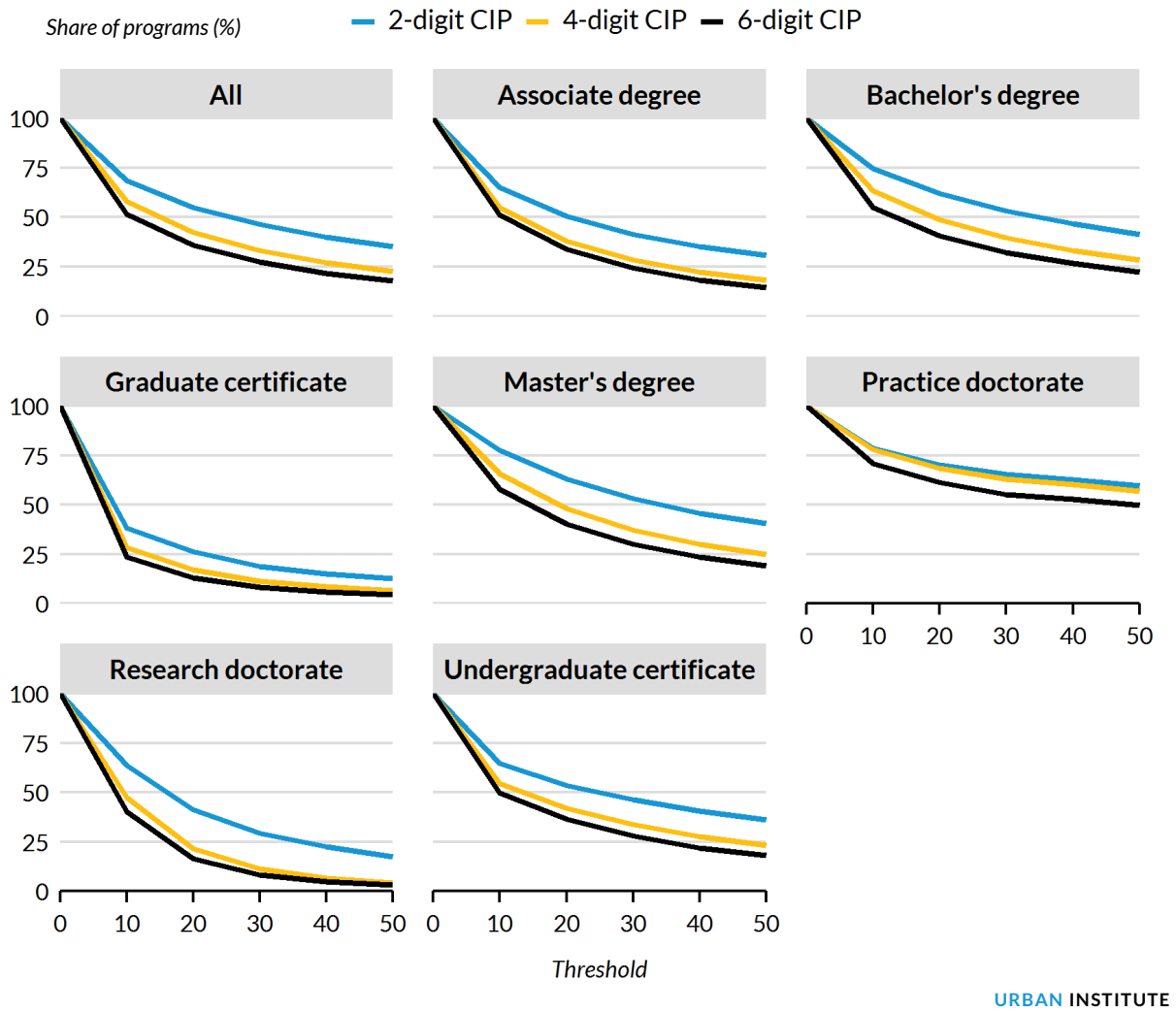
To find the point at which we have a reliable within-program measure and a large enough share of programs across the board, we roll together two years (2016–17 and 2017–18), three years (2015–16, 2016–17, and 2017–18), and then four years (2014–15, 2015–16, 2016–17, and 2017–18) and compare them with the year 2017–18. For two years, this would mean that we sum the number of students that graduated from a certain program in one year and the number that graduated from the same program in a second consecutive year. We treat the resulting total of students, which would be larger than either year, as if they occurred in one year and compare this new total with the inclusion rules above. If a program fails the rule threshold *n* = 50 in two separate years and pooling the years results in a total number of students greater than 50, the program is included. Regardless of whether a program exists one year and not the next, it will exist in the pooled year total and be included if the number of students that graduate in the one year is larger than the rule threshold *n*. Because pooling only two years together may not change the outcome for some programs, we test how increasing the number of pooled years will affect the share of programs included.

To understand how pooling years coincides with the rule thresholds, we collapse the number of program-institution observations to the credential–rolled year level and calculate the share of programs included. Again, the more years are rolled together, the more programs will be included, but rolling up too many years together would similarly have drawbacks, including a loss of relevance, slower visible changes in the data over time, and older data not necessarily reflecting current institutional practices and student characteristics.

FIGURE A.1

Share of Programs with Certain Numbers of Graduates across Different CIP Rollup Levels

Combines 2017–18 and 2018–19 cohorts



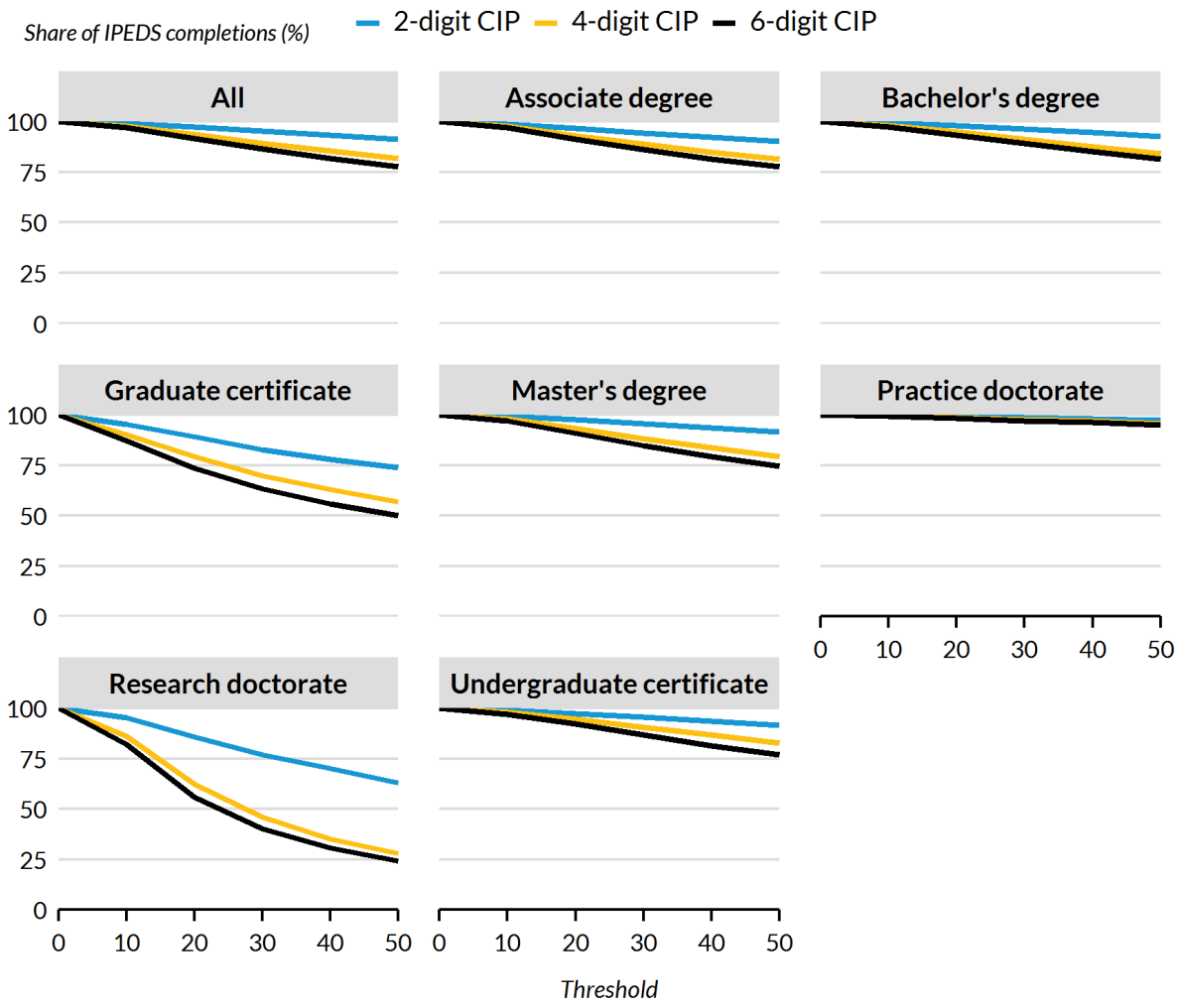
Source: Integrated Postsecondary Education Data System completions data.

Notes: CIP = Classification of Instructional Programs. Combines the two most recent years of data available.

FIGURE A.2

Share of Graduates in Programs with Certain Numbers of Graduates across Different CIP Rollup Levels

Combines 2017–18 and 2018–19 cohorts



URBAN INSTITUTE

Source: IPEDS completions data.

Notes: CIP = Classification of Instructional Programs; IPEDS = Integrated Postsecondary Education Data System. Combines the two most recent years of data available.

TABLE A.1

Unique College Scorecard CIP Codes, 2015–16 Cohorts

Share of two-digit CIP codes containing two or more four-digit CIP codes

	Median debt	Median earnings
Undergraduate certificate	29.4%	27.2%
Associate degree	35.6%	36.2%
Bachelor's degree	43.5%	47.1%
Graduate certificate	33.3%	35.0%
Master's degree	49.0%	51.0%
Research doctorate	31.7%	30.2%
Professional doctorate	25.1%	24.4%
Total (%)	41.2%	42.7%
Total (*N*)	24,619	22,148

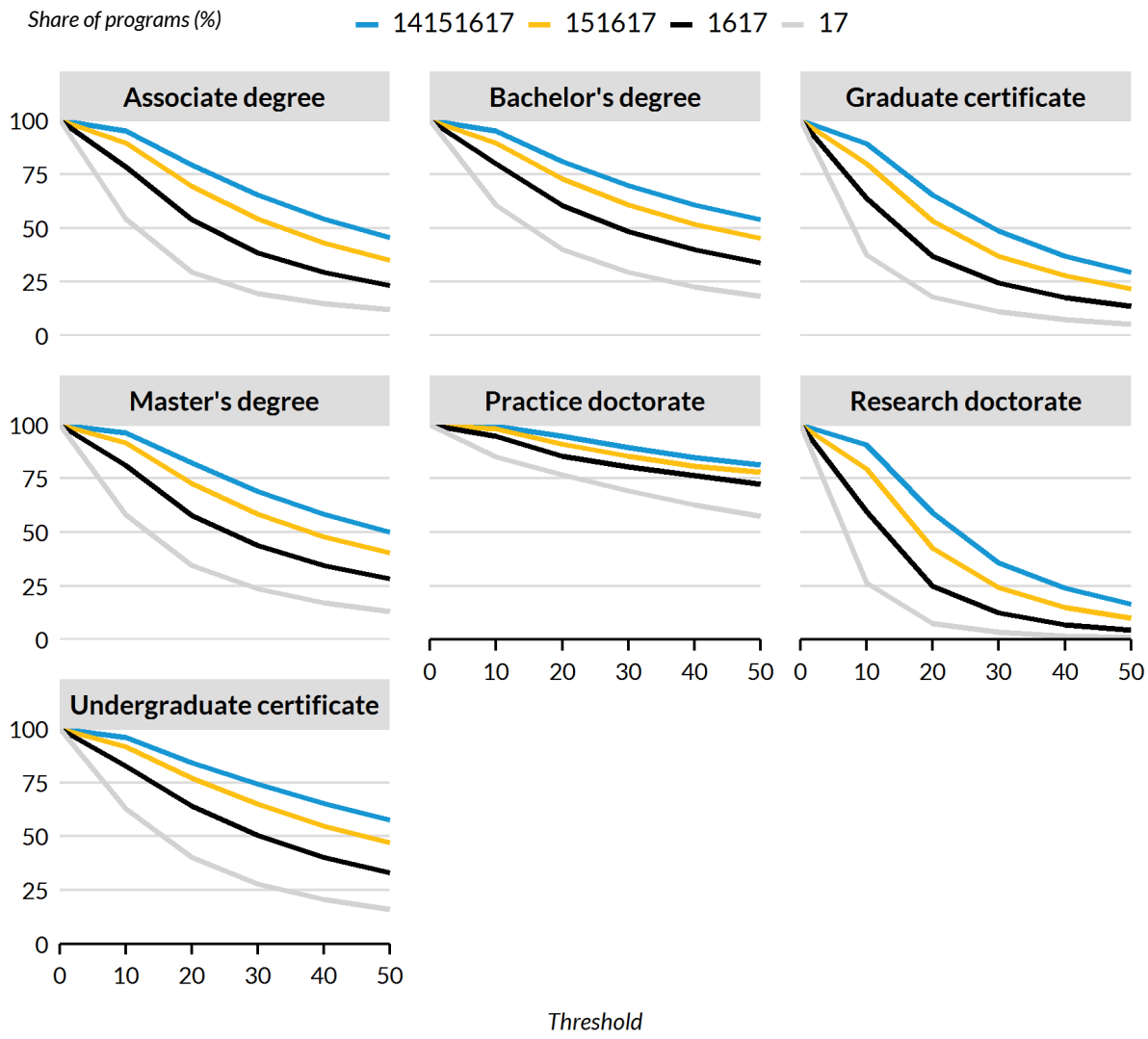
Source: Authors' calculations using College Scorecard data.

Notes: CIP = Classification of Instructional Programs; OPEIDs = Office of Postsecondary Education Identification codes. This counts only programs with data on debt or earnings levels and the number of students in that cohort. The programs are counted as being offered at the OPEID level, even though some OPEIDs include offerings in the same CIP across multiple unit IDs. The sample size refers to the number of two-digit CIP codes across all OPEIDs.

FIGURE A.3

Share of Programs, by Rule, for Each Credential Level

Six-digit CIP codes



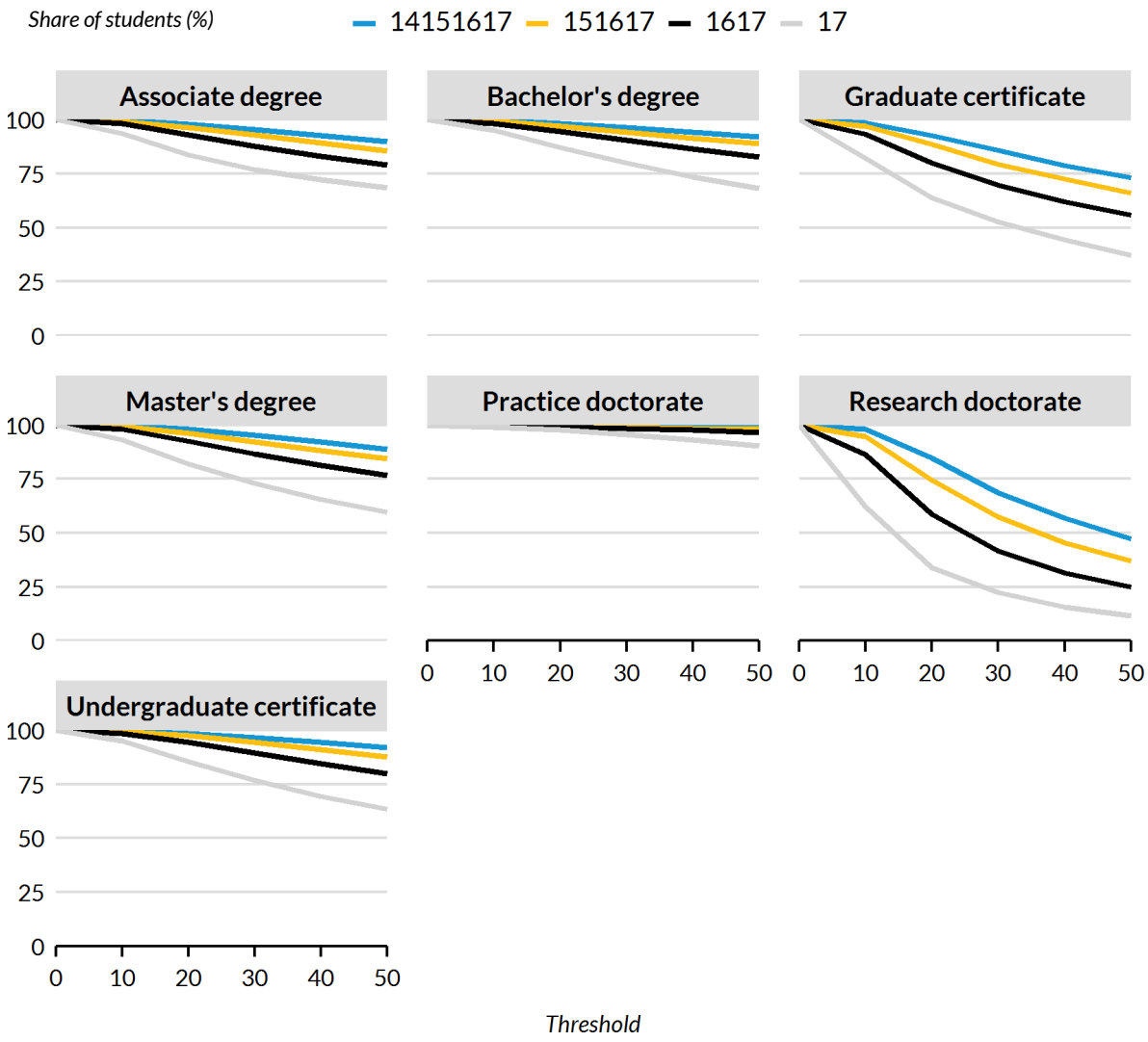
URBAN INSTITUTE

Source: Integrated Postsecondary Education Data System completions data.
Notes: CIP = Classification of Instructional Programs.

FIGURE A.4

Share of Students, by Rule, for Each Credential Level

Six-digit CIP codes



URBAN INSTITUTE

Source: Integrated Postsecondary Education Data System completions data.
Notes: CIP = Classification of Instructional Programs.

Notes

[1] The timing of an earnings measurement introduces further issues. Institution-level earnings (as collected by the College Scorecard) include all students not currently enrolled in school and are measured 6, 8, and 10 years after enrollment. Program-level earnings are currently measured one year after completion, with plans to expand the time horizon in future years. Neither is perfect. Pegging the measurement of earnings to a given number of years postcompletion accounts for students who take different amounts of time to complete, but it excludes students who did not complete the program. Conversely, measuring from point of entry penalizes programs that have disproportionate shares of students studying part time.

Different programs have different earnings trajectories (Webber 2016). It is commonly accepted that earnings "stabilize" 10 to 15 years after program entry (Chetty et al. 2017), but this makes accountability challenging because the program may have changed substantially since those students enrolled. This is a concern not merely for program-level earnings metrics but for earnings metrics more broadly.

There is also the question of how to account for students who pursue another credential after graduation. This interacts with the other measurement concerns. Arguably, students currently enrolled, whether at the same institution or elsewhere, should not be included in earnings metrics (as is the current approach). But if earnings are measured shortly after completion, and enrolled students are excluded, this might penalize two-year institutions who successfully prepare their students to transfer to four-year institutions, as their future highest-earning graduates are not included in the metric.

[2] Amu Gorel, "Harvard's A.R.T. Institute Freezes Enrollment after Federal Report Cites High Student Debt," *ARTery*, WBUR, January 17, 2017, https://www.wbur.org/artery/2017/01/17/art-institute-harvard.

[3] Doing so would require Congress to pass new legislation, such as the College Transparency Act, introduced in 2019.

References

Altonji, Joseph G., Erica Blom, and Costas Meghir. 2012. "Heterogeneity in Human Capital Investments: High School Curriculum, College Majors, and Careers." *Annual Review of Economics* 4:185–223.

Blagg, Kristin, and Macy Rainer. 2020. *Measuring Program-Level Completion Rates: A Demonstration of Metrics Using Virginia Higher Education Data*. Washington, DC: Urban Institute.

Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2 (1): 67–90.

Chetty, Raj, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. *Mobility Report Cards: The Role of Colleges in Intergenerational Mobility*. Working Paper 23618. Cambridge, MA: National Bureau of Economic Research.

Kelchen, Robert. 2018. *Higher Education Accountability*. Baltimore: Johns Hopkins University Press.

Kelchen, Robert, and Zhuoyao Liu. 2019. *Did Gainful Employment Regulations Result in College and Program Closures? An Empirical Analysis*. Working paper. South Orange, NJ: Seton Hall University.

Webber, Douglas A. 2016. "Are College Costs Worth It? How Ability, Major, and Debt Affect the Returns to Schooling." *Economics of Education Review* 53:296–310.

About the Authors

Kristin Blagg is a senior research associate in the Center on Education Data and Policy at the Urban Institute. Her research focuses on K–12 and postsecondary education. Blagg has conducted studies on student transportation and school choice, student loans, and the role of information in higher education. In addition to her work at Urban, she is pursuing a PhD in public policy and public administration at the George Washington University. Blagg holds a BA in government from Harvard University, an MSEd from Hunter College, and an MPP from Georgetown University.

Erica Blom is a senior research associate in the Center on Education Data and Policy, where she studies higher education policy. Blom received a bachelor's degree in mathematics and political science from Queen's University and a master's degree in economics from Western University. She also earned a doctoral degree in economics from Yale University, where her research focused on students' choices in college major.

Robert Kelchen is an associate professor of higher education and chair of the Department of Education Leadership, Management, and Policy at Seton Hall University. His research interests include higher education finance, accountability policies, and student financial aid. He received his PhD in educational policy studies at the University of Wisconsin–Madison.

Carina Chien is a research assistant in the Center on Education Data and Policy. She graduated from Cornell University with a bachelor's degree in economics and comparative literature.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

URBAN
INSTITUTE · ELEVATE · THE · DEBATE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org