Measuring Program-Level Outcomes in Higher Education

An Assessment of Program-Level Aggregation for Federal Earnings Data

Kristin Blagg

January 2026

As part of the One Big Beautiful Bill Act (OBBBA), Congress set forth a new rule, the "do no harm" (DNH) standard, that would revoke federal student loan eligibility for programs whose federally aided graduates do not meet a minimum earnings threshold four years after completion. A key concern for this standard is how to define a program, as many programs are small and would require multiple cohorts of graduates, or grouping with other programs, to reach the requirement that at least 30 observations be included in the earnings data.

In this analysis, I characterize which types of programs are most likely to be left out of the DNH standard, even after aggregation of graduates over time and across similar areas of study. My analysis finds the following:

- Even after five years of cohort aggregation, more than half of all programs of study (serving more than 10 percent of all graduates) do not have enough eligible graduates to meet the minimum cohort size for the DNH standard. Programs that award doctoral degrees and graduate certificates are less likely than other programs to meet the threshold. Adding more years of cohort data provides only small increases to the share of programs included and could make comparisons with larger programs, and with national data, more difficult.
- Aggregation of six-digit Classification of Instructional Programs (CIP) codes to broader categories (e.g., using alignment with career skills as determined by the US Department of

Education and the US Department of Labor) does improve the share of programs (about 60 percent) and graduates (about 93 percent) eligible for inclusion but cannot account for 100 percent of programs or graduates.

- Programs located in rural areas are least likely to be included in the DNH standard, even after aggregation by cohort year and with similar programs.
- Policymakers seeking to account for *all* federally aided programs, rather than the vast majority, may need to consider additional aggregation beyond what is specified in legislation for a small number of programs (e.g., across different program lengths or at the institution level).

Federal Program-Level Accountability

The DNH standard laid out in the OBBBA compares the median earnings of federally aided graduates from a given program, four years after graduation, with the median earnings for comparable working young adults who did not pursue a higher level of education. For example, for programs providing undergraduate degrees, the median earnings must be at or above the median earnings of young adults (ages 25 to 34) with only a high school diploma in the institution's state (or nationally if the institution draws more than 50 percent of enrollment from outside the state). Programs that do not pass this standard in two out of three years lose the ability to provide their students access to federal student loans. All federally aided programs of study, except undergraduate nondegree or certificate programs, are subject to this standard.

The OBBBA text requires that a minimum of 30 program graduates be identified in the earnings data to generate a median earnings estimate. Because few programs graduate more than 30 students each year, the legislation specifies that cohorts may be combined over time (e.g., three cohorts may be grouped together for the earnings measure). If grouping multiple cohorts from the same program still yields an insufficient sample size, programs of similar length may be grouped together, though the legislation does not suggest how this grouping should happen.

Methodology and Data

To understand what levels of aggregation may be necessary to account for all, or nearly all, programs and students, I analyze data from the Integrated Postsecondary Education Data System's (IPEDS) program-level completions dataset. These data provide information on the number of students who graduated from each program at the six-digit CIP code level. I merge these data with the most recent IPEDS directory file to ensure that I am reporting data for currently operating institutions and to capture institution characteristics. Students with two majors are included in the count of graduates from both programs. I then roll up the data from the unit ID (campus or college) level to the Office of Postsecondary Education ID level (which can represent a wider network of campuses or colleges) to link to the College Scorecard at the four-digit CIP code level.

Estimating the Number of Title IV Graduates with Measured Income

Because the DNH earnings measure is only for graduates who received federal financial aid and who had reported income in their fourth year, I compute the share of program graduates who met these criteria using the academic year 2014–15 and 2015–16 pooled graduating cohort in the College Scorecard, with earnings measured in the 2019 and 2020 calendar years, respectively. About 29 percent of programs have completion cohorts from 2014–15 and 2015–16 that can be linked to reported four-year earnings data. Bachelor's, master's, and first professional programs have relatively higher shares of programs with earnings data (42 percent for bachelor's programs and 32 percent for master's and first professional programs).

With these data, I compute the share of completers who would likely appear in later earnings data. For example, if 40 percent of master's program completers appear in the 2014–15 and 2015–16 earnings data at a given institution for CIP code 44.05 (public policy analysis), I assume the two underlying six-digit CIP programs (44.0501, public policy analysis, general, and 44.0502, education policy analysis) would each have the same 40 percent share of completers appearing in the data. I then compute the number of individuals we might find in each year of program graduate data, based on this rate. For example, if the institution graduates 100 public policy analysis master's candidates each year, I calculate that 40 would be found in the earnings data. If the institution graduates 10 education policy analysis candidates, 4 would be found in the earnings data.

Because many programs are too small (or have too few federally aided students) for inclusion in the College Scorecard data, I impute the share of graduates identified in the earnings data for these programs. I impute this share using the relationship between the share of graduates in the earnings data and the program level, the four-digit CIP program identifier, and the institution sector. With this imputed share, I can then estimate how many graduates may be identified as Title IV recipients in the earnings data.

This approach has several caveats. First, this imputation relies on data from programs that are, by definition, large enough to support a College Scorecard earnings estimate. If, for example, smaller programs are less likely to have federally aided students, the share of graduates that could be captured in earnings data could be lower than what I estimate. In a counterexample, if graduates from smaller programs are more likely to be identified as having earnings four years after completion, the share in the earnings data could be higher than I estimate. Second, this approach assumes that the share of federally aided students (and students with earnings) will stay relatively consistent over time. Finally, to ensure confidentiality, the Internal Revenue Service perturbs the data on the number of graduates in the earnings cohort using a differentially private algorithm, which adds further uncertainty to this estimation approach.

Implementing a "Do No Harm" Aggregation

Because of the lack of specificity in the legislation on how program cohorts should be aggregated, I make several decisions around defining program length, number of cohort years, and aggregation across CIP code. These decisions are guided by both data availability and by potential policy concerns.

Assessing Degree Type and Program Length

The DNH provision in the legislation suggests that after programs are rolled up by cohort year, they may be combined with other programs of the same length. Program length—the typical number of years or credits required for a given credential—is generally not available in a national dataset such as IPEDS. For this analysis, I aggregate programs within credential level, which may serve as a reasonable proxy for program length (e.g., as most associate's and master's degree programs are two years, and most bachelor's degree programs are four years).[1] It is unclear whether allowing aggregation by program length could permit the combination of different degree types, though it appears possible. For example, some institutions offer an accelerated two-year law degree program, which could theoretically be combined with two-year master's programs.

Selection of Cohort Year Aggregation

To assess how many additional students are captured by additional cohort years, I present data for one-, three-, and five-year program cohorts. I selected these benchmarks for three reasons. First, these benchmarks align with the way American Community Survey (ACS) data are aggregated into estimates using surveys from one, three, and five years of annual survey data. Because the legislation regarding graduate degree earnings requires producing state-level estimates of the median earnings of working adults ages 25 to 34 with a bachelor's degree in the same field of study, it is likely that the five-year ACS data will be used in some cases to estimate earnings by field. Matching cohort data years to comparable ACS survey years could be important; for example, a five-year (2020–24) median income estimate for leisure and hospitality workers includes 2020 and 2021, when employment in this sector dropped sharply because of the COVID-19 pandemic. The earnings data from this five-year estimate could be substantially lower than a one-year estimate for 2024.

Second, the aggregation of data beyond five cohort years means that the results are representative of individuals who could have first enrolled in the program as much as 10 to 15 years earlier. For example, a person who completes a bachelor's degree in six years and has their four-year earnings measured in 2020, as part of a 2020–24 five-year measurement cohort, would have started their program in 2010. It seems infeasible to hold programs accountable for decisions made many years ago.

Finally, the addition of more years of data makes it more difficult for a program to shift the median measure of earnings for their graduates. Imagine a program that enacted substantial improvements in their graduates' career prospects, starting with the 2021–22 graduation cohort. If the program is big enough to require only one year of data, the median 2025 calendar year earnings (measured for the

2020–21 cohort) are completely replaced in the next year by the 2026 calendar year earnings (measured for the "new and improved" 2021–22 cohort graduates). If the program has two years of cohort data, only about half the earnings data would be replaced (assuming the same number of eligible program graduates in each class). At the five-year cohort aggregation mark, a new cohort of data would likely replace only 20 percent of the graduates making up the earnings metric.

Designing a Program-Level Aggregation Scheme

The earnings framework allows for program-level aggregation of graduating cohorts across program length but does not describe how that aggregation should happen across different program categories. In this analysis, I explore two versions of how to aggregate across different six-digit CIP categories. First, I use the CIP SOC crosswalk, developed by the Bureau of Labor Statistics and the National Center for Education Statistics.[2] This crosswalk aligns postsecondary programs of study (at the six-digit CIP code level) with Standard Occupational Classification (SOC) codes, based on the assessment of what skills or knowledge are needed for a given career.

To aggregate CIP codes together using the crosswalk, I look at CIP codes that are linked to the same occupations. I first link CIP codes that have the same occupation codes listed, meaning that the two programs have been determined to generate skills and knowledge that match with the exact same set of occupations. After determining exact matches, I look for any CIP program that has at least one occupation that has been linked to another CIP program. For example, the occupation "architectural and civil drafters" is linked to both "architectural and building sciences/technology" (04.0902) and "3-D modeling and design technology/technician" (15.1307). These CIP codes do not lead to identical career sets (e.g., the crosswalk indicates that graduates from the former CIP code can serve as "architecture teachers, postsecondary" but graduates from the latter cannot) but demonstrate some commonalities, even across different two-digit CIP classifications.

As an alternative to the CIP SOC method, I roll up programs more directly to the four-digit and two-digit CIP reporting level, within degree type. Previous research suggests that not many programs would benefit from rollup to the four-digit CIP code level; 17 percent of four-digit programs of study in 2018–19 had more than one six-digit program of study contained within it (Blagg et al. 2021). As there are fewer than 50 two-digit CIP codes, aggregating to this level will likely increase the number of programs included but potentially at a cost of similarity in program content or in skills and knowledge developed.

Results

Broadly, my results show that any aggregation scheme, aside from aggregation at the institution level, will likely exclude some programs. This could be because of the newness of a program (i.e., there are few completer cohorts to aggregate) or the lack of other programs within a given set of similar CIP codes. Programs could also be excluded because most of their graduates do not use federal financial aid or because many of their graduates opt for additional education or not to participate in the labor market at the four-year mark.

But broad aggregation schemes—rolling up both years of cohorts and combining similar programs—tend to include more than 80 percent of program *graduates*. This is because programs that cannot be included in the DNH framework (as I model it), while numerous, tend to represent very small numbers of graduates. As a result, the typical program graduate is far more likely to graduate from a program that is eligible for DNH evaluation than from a program that is ineligible. I do find, however, that programs offered in rural areas, and those offered at public institutions offering programs of two years or less, are more likely to have a larger share of programs (and graduates) ineligible because of cohort size, even after aggregation.

Because the DNH framework excludes undergraduate programs that do not lead to a degree (presumably because they are covered by the gainful employment rule), I provide estimates for these programs in the first two figures for context, but I do not include these programs in my estimates by sector or rurality.

In my results, I present data using the CIP SOC framework for program aggregation after five years of cohorts are combined. Results for four-digit and two-digit CIP code aggregation schemes generally fall above and below the results for the CIP SOC framework and are presented at the end of this section. I present results using my Title IV imputation. Results using all graduates instead of the imputation are similar in trends across categories but tend to be more inclusive of programs because the cohort is not limited to Title IV–supported graduates with earnings.

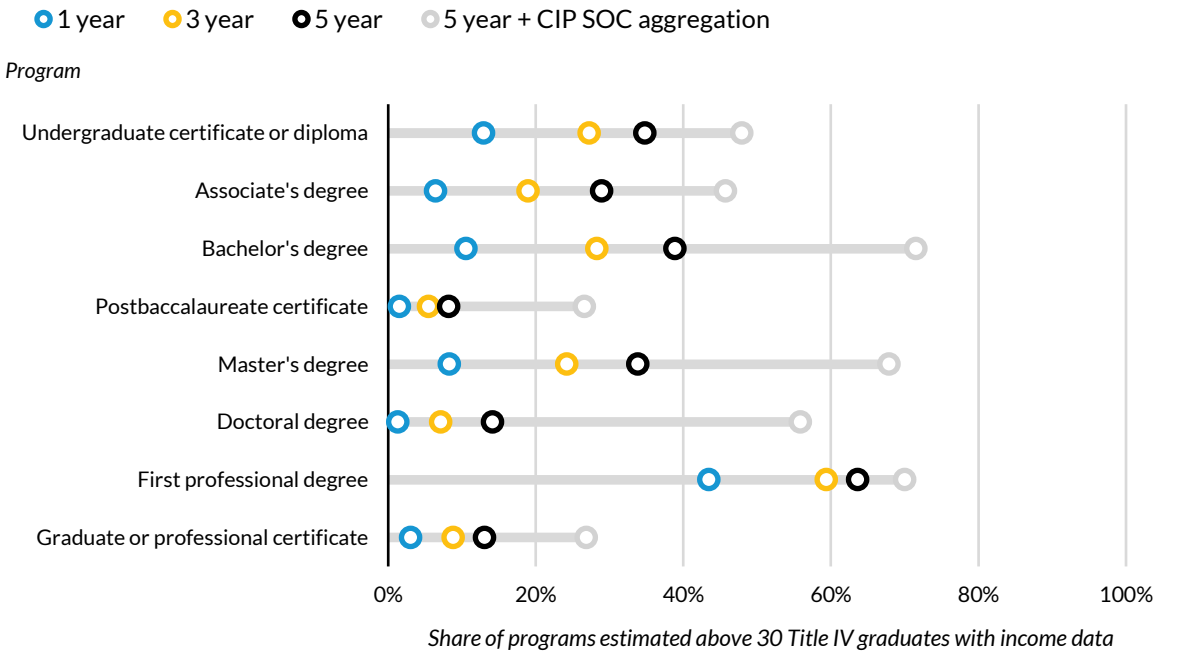
By Degree Type

Very few six-digit CIP code programs meet the 30-person criteria with one year of data, and even with five years of cohort data, only first professional programs have more than 50 percent of programs meet the 30-person benchmark for inclusion (figure 1). Using the CIP SOC framework for further aggregation, I can identify substantially more programs that could meet the eligibility criteria, especially for bachelor's degree, master's, and doctoral degree programs.

FIGURE 1

Share of Programs Eligible for Do No Harm Evaluation

By aggregation scheme and program level



URBAN INSTITUTE

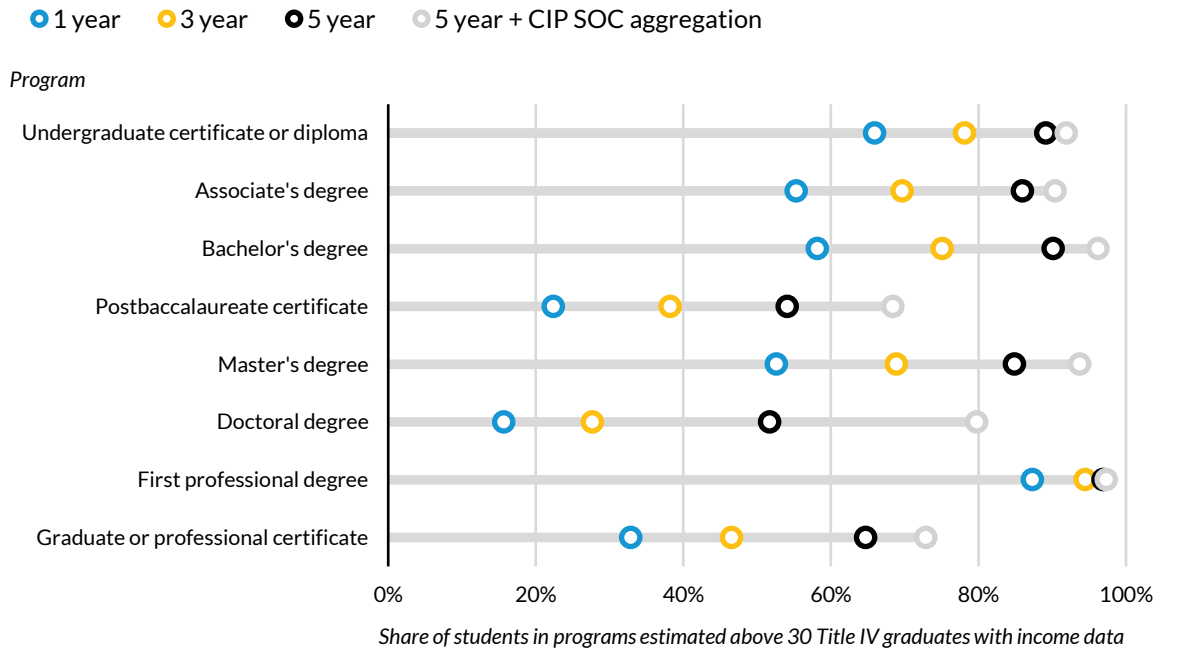
Source: Urban Institute analysis of Integrated Postsecondary Education Data System and College Scorecard data.
Note: CIP = Classification of Instructional Programs; SOC = Standard Occupational Classification.

But when I look at the share of 2023–24 graduates of these programs that would be represented under these aggregation schemes, the picture changes markedly (figure 2). Nearly all degrees and credentials—except postbaccalaureate and graduate or professional certificates—have at least 80 percent of program graduates eligible for the DNH earnings framework when aggregating five years of cohorts and across the CIP SOC crosswalk. Of note, across different credentials, different types of aggregation (cohort year and CIP SOC) do different work. For example, nearly 90 percent of those graduating from first professional degree programs could likely be judged under this framework using only one year of data. But adding five years of data makes a difference for capturing the earnings of students graduating from doctoral degree programs, nor does aggregation across six-digit CIP codes.

FIGURE 2

Share of Graduates from Programs Eligible for Do No Harm Evaluation

By aggregation scheme and program level



Source: Urban Institute analysis of Integrated Postsecondary Education Data System and College Scorecard data.
Note: CIP = Classification of Instructional Programs; SOC = Standard Occupational Classification.

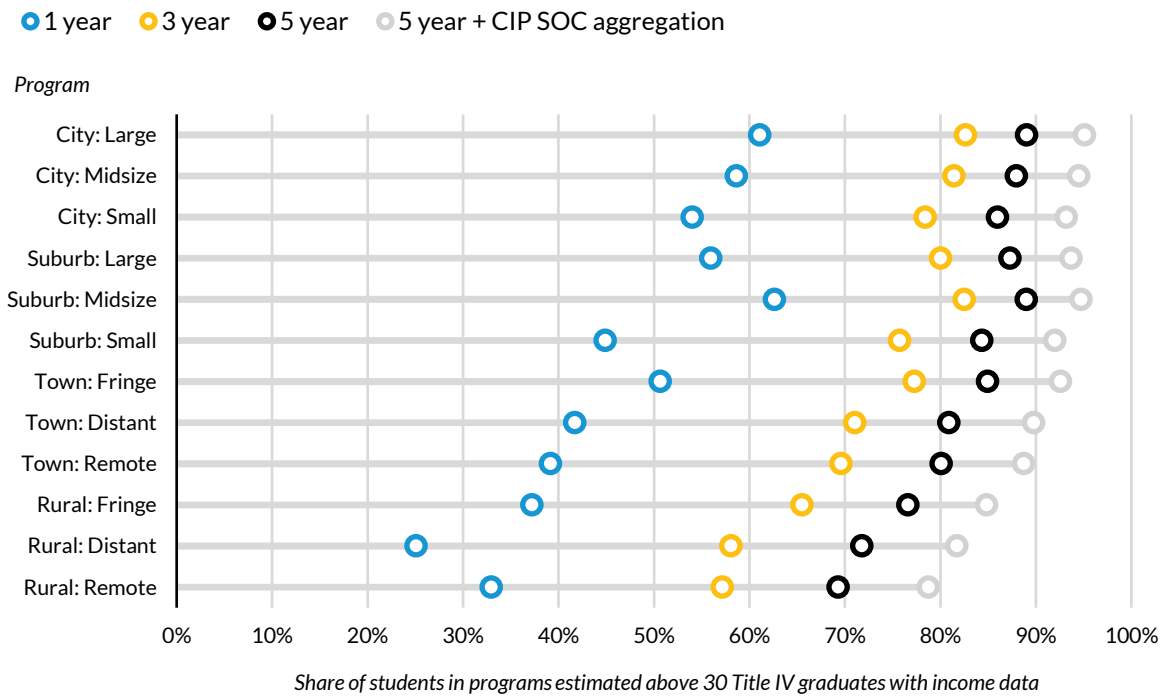
By Locale

A particular concern for program-level accountability is geography. When a program is no longer available in a given region (because of either the loss of federal loan eligibility or the stigma of less-than-stellar earnings data), students in rural areas may not have another nearby option to pursue a desired career path (Blagg 2022). In my analysis, I find that the availability of programs that are eligible for the DNH framework under my aggregation scheme varies by urbanicity (figure 3). Programs located in cities and suburbs tend to have about 95 percent of their graduates represented in the fully aggregated data, while rural programs have less than 90 percent of their graduates represented.

FIGURE 3

Share of Graduates from Programs Eligible for Do No Harm Evaluation

By aggregation scheme and locality



URBAN INSTITUTE

Source: Urban Institute analysis of Integrated Postsecondary Education Data System and College Scorecard data.
Notes: CIP = Classification of Instructional Programs; SOC = Standard Occupational Classification. Undergraduate certificate or diploma data are excluded.

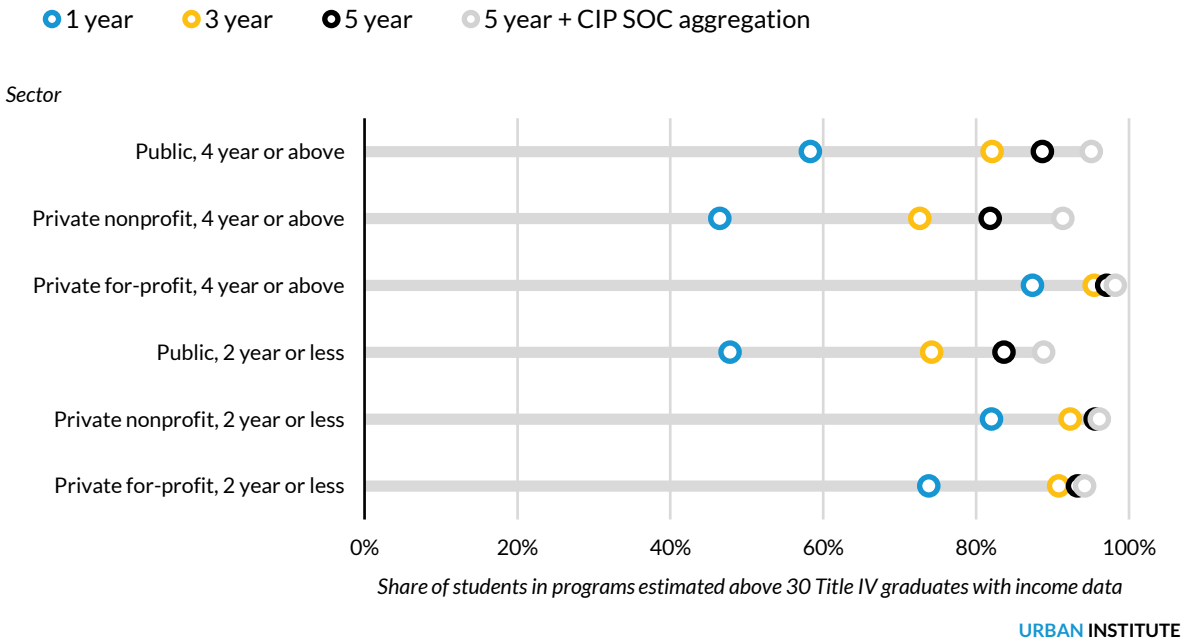
By Sector

There are some differences in DNH eligibility by sector. Notably, nearly all graduates of private for-profit four-year institutions are likely represented in the data as aggregated here (figure 4). Indeed, many students are enrolled in programs at private for-profit institutions that would likely need only one year of data to meet the DNH threshold for inclusion. Private nonprofit four-year institutions and public two-year institutions are less likely to have all their graduates represented in this aggregation scheme.

FIGURE 4

Share of Graduates from Programs Eligible for Do No Harm Evaluation

By aggregation scheme and program sector



URBAN INSTITUTE

Source: Urban Institute analysis of Integrated Postsecondary Education Data System and College Scorecard data.
Notes: CIP = Classification of Instructional Programs; SOC = Standard Occupational Classification. Undergraduate certificate or diploma data are excluded.

Alternate CIP Aggregation Strategies

For simplicity, I presented a single six-digit CIP aggregation scheme, which uses the CIP SOC crosswalk to combine programs with similar occupational profiles. I also modeled a more direct aggregation approach, where programs are joined with other programs at the same degree level that have the same four-digit and two-digit CIP codes (table 1). Broadly, the four-digit aggregation tends to capture fewer students in eligible programs than my original analysis, and the two-digit CIP aggregation tends to capture more. Even at five years of cohort aggregation and combining programs at the two-digit CIP code level, there is no point at which 100 percent of programs (or federally aided graduates) are completely accounted for.

TABLE 1

Share of Graduates from Programs Eligible for Do No Harm Evaluation

By CIP aggregation scheme

	5 year + CIP SOC aggregation	5 year + 4-digit CIP aggregation	5 year + 2-digit CIP aggregation
Level			
Undergraduate certificate or diploma	92%	92%	97%
Associate's degree	90%	89%	95%
Bachelor's degree	96%	93%	98%
Postbaccalaureate certificate	68%	61%	80%
Master's degree	94%	89%	96%
Doctoral degree	80%	58%	85%
First professional degree	97%	98%	99%
Graduate or professional certificate	73%	74%	87%
Locale			
City: Large	95%	92%	97%
City: Midsize	95%	91%	97%
City: Small	93%	89%	96%
Suburb: Large	94%	90%	96%
Suburb: Midsize	95%	91%	97%
Suburb: Small	92%	87%	95%
Town: Fringe	93%	88%	95%
Town: Distant	90%	85%	93%
Town: Remote	89%	84%	92%
Rural: Fringe	85%	81%	90%
Rural: Distant	82%	77%	87%
Rural: Remote	79%	74%	84%
Institution sector			
Public, 4 year or above	95%	91%	97%
Private nonprofit, 4 year or above	91%	86%	95%
Private for-profit, 4 year or above	98%	98%	99%
Public, 2 year or less	89%	87%	95%
Private nonprofit, 2 year or less	96%	96%	98%
Private for-profit, 2 year or less	94%	95%	97%

Source: Urban Institute analysis of Integrated Postsecondary Education Data System and College Scorecard data.
Notes: CIP = Classification of Instructional Programs; SOC = Standard Occupational Classification. Undergraduate certificate or diploma data are excluded by location and sector.

Conclusion and Further Considerations

These results illustrate the difficulties of one aspect of implementation of the "do no harm" standard: the aggregation of program cohorts so that all (or nearly all) programs are assessed against an earnings threshold. My analysis suggests that an approach that includes nearly all programs would likely require at least five cohorts of data and substantial aggregations across CIP codes, perhaps even beyond two-digit CIP categories. But an approach that accounts for the results of more than 80 percent of graduates of these programs seems feasible with five years of data and some CIP aggregation. Programs in rural areas, and programs that provide doctoral degrees or graduate or postbaccalaureate certificates, are most likely to have fewer graduates represented.

In my analysis, I discovered a few additional questions that will need consideration:

- **Estimation of "Donor" Program Earning Thresholds:** In many cases in my data, a small program reaches the earnings threshold by being paired with a bigger program. For example, a program with 10 eligible graduates in the five-cohort sample might reach the threshold by being paired with a program with 100 graduates in five cohorts. It is unclear whether these programs will be assessed together (i.e., the median threshold will be from the 110-graduate sample for both programs) or whether the larger program will be assessed on its own (i.e., the small program uses the 110-graduate sample, but the large program uses only its own 100 graduates).
- **Use of Either a Consistent or Rolling Number of Cohorts:** Rulemakers will need to decide whether programs are measured with a consistent number of cohort years (e.g., all programs measured with three cohorts) or whether the number of cohorts will depend on whether the 30-eligible-graduate benchmark is reached (rolling up cohorts). Earlier regulation on program-level earnings—the gainful employment and financial value transparency rules—allowed for the use of either two cohorts or four cohorts (if two is insufficient to identify 30 students). But these earlier regulations did not allow for combinations across different programs. In this case, a consistent number of cohorts, at least within institution and program length, may be preferable to avoid combining a small program with five cohorts of data with a large program needing only one cohort of data.
- **Development of a "Participation" Appeal:** Some programs, particularly graduate programs where federal grant aid is less common, tend to have relatively small shares of students that use federal financial aid. Rulemakers may want to consider whether programs with very low rates of federal loan use (e.g., fewer than 5 percent of students, or fewer than 10 students total, borrow federal loans for the program) should be exempt. These programs are difficult to include in the DNH threshold because of the low borrowing rate, and they present relatively little risk to the taxpayer. This type of appeal is similar to what is available for the cohort default rate, where programs can appeal a high default rate using a participation rate index appeal, showing that the share of students borrowing federal loans is low.

The range of program offerings at institutions that offer federal financial aid is wide and diverse. Policymakers should work to find a DNH standard that can accommodate most programs while maintaining a consistent and logical approach for identifying earnings cohorts.

Notes

[1] For this analysis, I combine all undergraduate certificates and diploma programs of less than two years, though in practice, these programs are not part of the earnings accountability scheme. I also combine academic doctoral degrees with other (nonprofessional practice) doctoral degrees.

[2] "CIP SOC Crosswalk," US Department of Education, Institute of Education Sciences, National Center for Education Statistics, accessed January 5, 2026, https://nces.ed.gov/ipeds/cipcode/post3.aspx?y=56.

References

Blagg, Kristin. 2022. "The Limits and Potential of Program-Level Earnings in Higher Education Accountability." In *Student Outcomes and Earnings in Higher Education Policy*, edited by Jason D. Delisle. American Enterprise Institute.

Blagg, Kristin, Erica Blom, Robert Kelchen, and Carina Chien. 2021. *The Feasibility of Program-Level Accountability in Higher Education: Guidance for Policymakers*. Urban Institute.

About the Author

Kristin Blagg is a principal research associate in the Work, Education, and Labor Division 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. Blagg holds a BA in government from Harvard University, an MSEd from Hunter College, an MPP from Georgetown University, and a PhD in public policy and public administration from the George Washington University.

Acknowledgments

This brief was funded 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 author 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.

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization founded on one simple idea: To improve lives and strengthen communities, we need practices and policies that work. For more than 50 years, that has been our charge. By equipping changemakers with evidence and solutions, together we can create a future where every person and community has the opportunity and power to thrive.