NBER WORKING PAPER SERIES

MEASURING OPPORTUNITY IN U.S. HIGHER EDUCATION

Caroline M. Hoxby
Sarah Turner

Working Paper 25479
http://www.nber.org/papers/w25479

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2019

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. The authors acknowledge constructive comments from Sandy Baum, Sandra Black, John Bound, Damon Clark, Paul Courant, David Ellwood, Joshua Goodman, Michael McPherson, Richard Murnane, David Neumark, Jeffrey Smith, Christopher Taber, Martin West, and participants in several scholarly seminars and conferences. The authors also acknowledge helpful feedback from a number of government staff and a number of persons involved in academic governance. The data in this paper were previously generated as descriptive statistics used in the writing of Hoxby and Avery (2013) and Hoxby (2015a). We therefore gratefully acknowledge those who helped us with those projects including The College Board, ACT, and (under contracts TIR-NO-12-P-00378 and TIR-NO-15-P-00059) Barry W. Johnson, Michael Weber, and Brian Raub of the Statistics of Income Division, Internal Revenue Service.

Measuring Opportunity in U.S. Higher Education
Caroline M. Hoxby and Sarah Turner
NBER Working Paper No. 25479
January 2019
JEL No. H0,H75,I20,I22,I23,I24,I32

ABSTRACT

In identifying whether universities provide opportunities for low-income students, there is a measurement challenge: different institutions face students with different incomes and preparation. We show how a hypothetical university's "relevant pool"–the students from whom it could plausibly draw–affects popular measures: the Pell share, Bottom Quintile share, and Intergenerational Mobility. Using a proof by contradiction, we demonstrate that universities ranked highly on the popular measures can actually serve disproportionately few low-income students. We also show the reverse: universities slated for penalties on the popular measures can actually serve disproportionately many low-income students. Furthermore, the Intergenerational Mobility measure penalizes universities that face relatively equal income distributions, which are probably good for low-income students, and rewards universities that face very unequal income distributions. In short, by confounding differences in university effort with differences in circumstances, the popular measures could distort university decision making and produce unintended consequences. We demonstrate that, with well-thought-out data analysis, it is possible to create benchmarks that actually measure what they are intended to measure. In particular, we present a measure that overcomes the deficiencies of the popular measures and is informative about all, not just low-income, students.

Caroline M. Hoxby
Department of Economics
Stanford University
Landau Building, 579 Serra Mall
Stanford, CA 94305
and NBER
choxby@stanford.edu

Sarah Turner
University of Virginia
Department of Economics
Monroe Hall, Room 237
248 McCormick Rd
Charlottesville, VA 22903
and NBER
sturner@virginia.edu

MEASURING OPPORTUNITY IN U.S. HIGHER EDUCATION

Caroline M. Hoxby and Sarah Turner[†]

Abstract

In identifying whether universities provide opportunities for low-income students, there is a measurement challenge: different institutions face students with different incomes and preparation. We show how a hypothetical university's "relevant pool"-the students from whom it could plausibly draw-affects popular measures: the Pell share, Bottom Quintile share, and Intergenerational Mobility. Using a proof by contradiction, we demonstrate that universities ranked highly on the popular measures can actually serve disproportionately few low-income students. We also show the reverse: universities slated for penalties on the popular measures can actually serve disproportionately many low-income students. Furthermore, the Intergenerational Mobility measure penalizes universities that face relatively equal income distributions, which are probably good for low-income students, and rewards universities that face very unequal income distributions. In short, by confounding differences in university effort with differences in circumstances, the popular measures could distort university decision making and produce unintended consequences. We demonstrate that, with well-thought-out data analysis, it is possible to create benchmarks that actually measure what they are intended to measure. In particular, we present a measure that overcomes the deficiencies of the popular measures and is informative about all, not just low-income, students.

[†] The authors' affiliations are, respectively, Stanford University and the National Bureau of Economic Research and University of Virginia and the National Bureau of Economic Research. The authors acknowledge constructive comments from Sandy Baum, Sandra Black, John Bound, Damon Clark, Paul Courant, David Ellwood, Joshua Goodman, Michael McPherson, Richard Murnane, David Neumark, Jeffrey Smith, Christopher Taber, Martin West, and participants in several scholarly seminars and conferences. The authors also acknowledge helpful feedback from a number of government staff and a number of persons involved in academic governance. The data in this paper were previously generated as descriptive statistics used in the writing of Hoxby and Avery (2013) and Hoxby (2015a). We therefore gratefully acknowledge those who helped us with those projects including The College Board, ACT, and (under contracts TIR-NO-12-P-00378 and TIR-NO-15-P-00059) Barry W. Johnson, Michael Weber, and Brian Raub of the Statistics of Income Division, Internal Revenue Service.

1 Introduction

Higher education may be one of the most important channels through which people can attain improved life outcomes based on their merit rather than family background. If qualified students from lower-income families are underrepresented in higher education-owing to credit constraints, information barriers, or other obstacles-there is potentially a failure not just in equity but economic efficiency.

The question of which institutions lag (or lead) in providing opportunities for low-income students has become a front-line issue in national policy discussions. Legislative initiatives such as the Senate's ASPIRE Act propose to rank institutions on their representation of low-income students and to impose financial penalties on institutions below a certain ranking.[1] And, the most recent version of the U.S. News and World Report Rankings include a measure of "social mobility." Other news outlets such as The New York Times and Washington Monthly have prominently published rankings of colleges based on representation of low-income students, while taking editorial positions excoriating individual institutions and demanding policies based on such measures.[2]

Unfortunately, these initiatives ignore a thorny measurement

[1] In December 2017, Senators Chris Coons (D-Del.) and Johnny Isakson (R-Ga.) introduced the Access, Success, and Persistence In Reshaping Education Act (ASPIRE Act, S.2201 - 115th Congress).The U.S. Department of Education ("ED") has been working on a plan to reward and penalize colleges based on their Pell or Bottom Quintile shares. Several states already give larger appropriations to colleges that are ranked higher on one of these indices. ED (2018) continues to pursue a "College Ratings Framework." The National Conference of State Legislatures (2018) identifies 10 states that appropriate funds on the basis of the Pell Share. Private funders such as the Bloomberg Philanthropies' American Talent Initiative give funds to institutions based on their Pell Shares.

[2] Further, an editorial in The New York Times (2014) demanded that the federal government "set minimum performance standards for all colleges receiving federal aid: at least 17 percent enrollment of poor and working-class students …." The New York Times used the measures to single out individual institutions in articles such as "Virginia, Betraying Jefferson" (Leonhardt, 2017b).

challenge, one that can turn good intentions into penalties for institutions that are actually succeeding in providing opportunities for low-income students and trigger rewards for institutions that are far less successful than they might appear. What makes measurement challenging is that different institutions face students whose family income and preparation differ.

Suppose that the University of Maine and University of Connecticut enroll all students from their respective states who fit their academic standards, as defined by receiving a score on a college admissions test that is within the range of most students currently enrolled. Based on the different populations of their states, the University of Maine would draw 22 percent of its students from families with incomes below $40,000, while the University of Connecticut would draw only 10 percent of its students from this income range. The University of Maine would be judged much more favorably by popular measures of "opportunity" and rewarded by proposed accountability systems, while the University of Connecticut would be penalized. However, those rewards and penalties could not result from their differential success in enrolling low-income students since, in the example, all relevant students enroll at each university, regardless of their incomes. The universities would be rewarded based on their circumstances, not their behavior or effort.

More generally, popular measures of "opportunity" confound differences in universities' effort with differences in their circumstances. Specifically, while the measures mean to measure a university's effort to enroll well-qualified low-income students, what they actually measure can largely reflect differences in the pools of students from whom the universities could plausibly draw. The popular measures include a university's share of

students who receive federal Pell grants (the "Pell Share")[3], the share whose family income is in the bottom 20 percent of the national family income distribution (the "Bottom Quintile" measure), and the Intergenerational Mobility ("IGM") measure based on the percentage of enrolled students whose families are in the bottom 20 percent but who as adults end up in the top 20 percent of the national income distribution. That is, the IGM measure multiples the Bottom Quintile measure by an estimate of the probability that a Bottom Quintile student at the school ends up in the top quintile of the national income distribution.

Does a university that does well on these popular measures necessarily have more successful policies for recruiting low-income students than a university that does poorly on these measures? As we show in this analysis, the answer is no.

To be clear, we are not criticizing the intentions behind the efforts to measure the success of institutions in providing opportunities for low-income students. Rather, we are attempting to give higher education leaders the understanding and tools needed to conduct self-evaluation that is likely to further those good intentions.

A. What this Analysis Attempts to Do

Our analysis has two main aims. First, we provide a proof by contradiction. That is, we demonstrate that some universities slated for rewards based on the popular measures actually serve relatively few low-income students from their pool. The reverse is also true: some universities that are slated for penalties based on the popular measures actually serve disproportionately many low-income students from their pool. Thus, measurement matters greatly in this context: judging institutions using the popular measures is likely to produce unintended outcomes

[3] Tebbs and Turner (2005) discuss other reasons-such as adult and international students-why the Pell Share is problematic.

because they often give the wrong answer.

Second, in order to demonstrate that metrics that measure what they intend to measure are available, we construct a new measure of a university's success in providing opportunities to low-income students. Specifically, we show how to construct a university's "relevant pool"-the pool of students from which it could plausibly draw based on its academic mission and geographic location. We illustrate how to compare a university's students to its relevant pool, and we demonstrate that such comparisons are highly informative—not just to show how the university serves low-income students but how it serves all students.

The Pell, Bottom Quintile, and IGM measures could be regarded as reasonable proxies for universities' effort in recruiting low-income students if, when tested, they proved to be closely aligned with measures based on universities' relevant pools. However, we show that they are not, in fact, closely aligned and are, therefore, measuring something (a university's circumstances) that is different from what they are intended to measure (a university's effort). Even worse—because "top performers" and "bottom performers" receive most of the attention—is if the popular measures identify top performance as bottom performance and vice versa. Unfortunately, as we show, such top-to-bottom inversions do occur, and they affect some very salient universities.

In addition to our two main aims, we discuss the IGM measure in some detail because it appears to us that it may be substantially misunderstood. We show that it shares the issues that affect the Bottom Quintile measure but, moreover, that it has additional issues that lead it to punish universities that face relevant pools with high levels of income equality. It rewards universities that face relevant pools with very high levels of income inequality —for instance, universities located in California. Since the IGM measure penalizes income equality

rises and rewards income inequality, it may fail to embody its intended uses.

This paper concludes with a broad discussion of the process by which universities might evaluate themselves on the degree to which they are providing the opportunities that fulfill their educational missions. We also discuss why universities might benefit from such a process, which involves both self-examination (as to mission) and data-based metrics that measure what they intend to measure (such as the degree to which the university is enrolling low-income students from its relevant pool).

B. What this Analysis Does Not Attempt to Do

Having said what we attempt, it is worth saying what we do not attempt. Although we suggest methods by which schools could judge whether they are accomplishing their educational missions, we do not seek to define those missions. These differ in terms of the backgrounds and preparation of the students served. For instance, Berea College states that its mission is: "To provide an educational opportunity for students of all races, primarily from Appalachia, who have great promise and limited economic resources." This statement defines Berea's relevant pool (all races, primarily Appalachian, of great promise) and its income representation goal (disproportionate emphasis on low-income students). The methods we propose would allow Berea to judge itself against its own mission, but we do not propose to impose a mission on Berea.

Precisely because we do not want to impose missions on universities, we use examples drawn from states' most selective or "flagship" public universities for our proof by contradiction and our illustration of a sound way to measure opportunity. We use them because their key undergraduate mission and constraints are a matter of public record—largely to educate well-prepared students from their own state. Thus, we know approximately how they would define their relevant pools, and we can construct

those pools with a fair degree of confidence. However, the measurement issues we confront apply just as much to non-flagship institutions that are more or less selective and public, nonprofit, or for-profit. Even universities like Harvard and Stanford, which claim to recruit students nationally, in fact have relevant pools that differ substantially owing to strong geographical skews.

Although flagships' missions and constraints are quite public and the pools we construct for them are grounded in empirical evidence about their behavior, we emphasize that our assumptions are meant only to facilitate illustration. They do not preclude a university specifying alternative parameters.

Moreover, we do not attempt—in this analysis—to answer fundamental questions such as (i) why students' preparation varies with family background; (ii) why different institutions have curricula and resources designed to serve students with different levels of preparation; and (iii) why students often prefer more proximate institutions even when not constrained to attend them. These questions are of absorbing interest to us and other economists of higher education, but we stick to a simpler question in this paper: Given the curricula offered by various institutions (which implicitly constrain the students for whom their offerings generate a high return), given the legal and market conditions under which institutions operate (which affect how attractive they are to out-of-state or otherwise distant students), and given the correlation between income and preparation, how can we measure an institution's enrollment of students from across the income distribution?[4]

[4] We and others investigate such questions in prior and continuing research. Among the important explanations are: (i) market forces that tend to align educational investments with students' capacity to benefit from them; (ii) market forces that induce institutions to offer skills demanded by local employers; (iii) forces that induce institutions to minister to students with preparation typical of local high schools; (iv) states' structuring their public postsecondary sectors in accordance with their populations sizes and with economies of scale and scope.

2. How a University is Affected by the Income-Achievement Distribution of its Pool

Consider a hypothetical university that enrolls a student body that is fully representative of its relevant pool. How is an indicator such as the Bottom Quintile measure affected by the divergence between the relevant pool's and nation's income-achievement distribution?

To illustrate the issues, we assume that the relevant pool's and nation's distributions of income and preparation are bivariate normal and differ only by mean achievement, mean income, the variance of achievement, the variance of income, and the covariance of achievement and income. Think of these means and variances as representing differences across geographically-defined pools. The illustration varies these characteristics over ranges inspired by their ranges among U.S. states.

To represent most U.S. schools, we vary the university's curriculum to be designed for:

i. "open enrollment"—that is, designed to serve students whose preparation is above the 25th percentile, approximately those who attain a high school diploma on-time;

ii. students who achieve above the median on a national basis (for instance, flagship universities associated with certain—usually small population—states);

iii. students who achieve in the top quartile on a national basis;

iv. students who achieve in the top decile on a national basis (for instance, the most selective several flagship universities).

Although we attempt to select reasonable ranges to characterize the distributions of income and achievement, keep in mind that this is a hypothetical exercise designed to show what happens as we change each mean and variance, keeping the

others the same. Below, we provide evidence based on real universities' pools.

Table 1 varies the earnings and achievement distributions (rows) and preparation required at different levels of college selectivity (columns), showing striking implications for the Bottom Quintile measure. (The results would be similar if we showed the Pell Share.) The first row shows the case in which the university draws from all the nation. Even for an open enrollment school in this case, the percentage of the university's pool that falls into the Bottom Quintile is only 14.7 percent. It is 14.7 percent, as opposed to 20 percent, because drop-outs are concentrated in the bottom income percentiles.[5] Moreover, the Bottom Quintile share falls as we increase the preparation expected of potential students, dropping to 11.1 percent at the 50th percentile and then to 5.1 percent at institutions with a curriculum designed for top decile students.

Changing the mean income of the area from which a university draws its relevant pool, shown in the next two rows, illustrates how circumstances affect a university's Bottom Quintile measure.[6] Facing a high income area, even an open enrollment university has a Bottom Quintile share of only 7.0

[5] Empirically, students who lack college-readiness are concentrated at the bottom of the income distribution. For instance, in the American Community Survey ("ACS" 2015), 12 percent of 18 year olds in the bottom income decile are high school dropouts. This percentage declines monotonically as we move to higher deciles. Chetty et al (2017) do not condition on any measure of preparation (not even age-for-grade) and therefore set the threshold for their Bottom Quintile measure well below the 20th percentile among students reasonably likely to obtain a high school diploma on-time. Chetty et al's (2017) threshold is $25,000 for the 1980 birth cohort whereas it is about $31,000 if we use the nationally representative ACS and merely eliminate drop-outs and the institutionalized (most of whom are in juvenile detention and are not ready to enroll in college). In fact, among students who are not drop-outs or institutionalized, $25,000 is approximately the 16th percentile of income, not the 20th percentile. The foregoing statistics are for the 1980 birth cohort when they are age 17, based on the ACS 3-year file 2006-2008.

[6] Such changes correspond to a university serving an area with average income that differs from the nation. Our high (low) mean income assumption is 70 (30) compared to 50 for the nation. If we set the nation's mean income to 50, a state with mean income of about 70 is Connecticut. A state with mean income of about 30 is New Mexico.

percent. It is a mere 1.9 percent if the university serves students with preparation in the top decile. In contrast, a university facing a low income area has a pool with a Bottom Quintile share of 26.3 percent if it is open enrollment and 11.0 percent if it serves students with top decile preparation.

The next two rows change the variance of the income distribution, corresponding to areas in which incomes are more and less equal than they are nationally.[7] In a more equal area, fewer families are in the national bottom quintile so there are fewer bottom quintile students who could attend any university—leading to a low Bottom Quintile share. Conversely, in an unequal area, the income distribution has fat tails, with the consequence that there are more students in families below the national 20th percentile—leading to a high Bottom Quintile share. The less equal the area, the higher is the Bottom Quintile share and vice versa.

Low and high variance cases generate very different Bottom Quintile shares. For instance, a top-decile-serving university facing a relatively equal income distribution would find that only 2.6 percent of its relevant pool was in the bottom quintile. If it were facing an unequal income distribution, 7.2 percent of its pool would be in the bottom quintile. It is important to grasp that the low and high variance cases are so different precisely because the quantile being considered is so low. The lower is the quantile, the more the tails of the distribution affect measurement. By focusing on the 20th percentile (as opposed to—say—the 40th), the Bottom Quintile measure exacerbates problems due to the relevant pool's having income equality that differs from that of the nation.

[7] We change the coefficient of variation from the nation's average of 1 (which is approximately correct) to a low of 0.75 and a high of 1.25. Alaska resembles our low case and New York resembles our high case.

The next row of the table increases the correlation between income and achievement from 0.4 to 0.6.[8] This has the effect of reducing the Bottom Quintile share greatly—especially at schools with curricula designed for those with top decile preparation. Those schools face a Bottom Quintile share that falls from an already low 5.0 percent to a minimal 1.2 percent as the achievement-income correlation rises.

The final row of the table reduces mean preparation in a local area.[9] This causes the Bottom Quintile share to fall for all universities. In a low achieving area, an open enrollment school would have only 12.6 percent of its students in the Bottom Quintile (relative to 14.7 percent in the national case) while a top-decile-serving university would have only 3.8 percent in the Bottom Quintile (relative to 5.1 percent in the national case).

The highest Bottom Quintile share in the table is 26.3 percent; the lowest is 1.2 percent. These are very large differences that cannot be attributed to differential university "effort" (since that is assumed away in the exercise). These statistical matters seem to be greatly under-appreciated by analysts who attribute the differences in schools' ranking on the popular measures to institutional effort. They confound behavior with circumstances.

---

[8] There is no ideal range of achievement-income correlations for consideration in our hypothetical example. The empirical correlation varies with scaling (whether achievement is measured in SAT points, for instance) and selection (how we deal with drop-outs and other students who do not take college assessments, for instance). One useful benchmark, though, is the correlation between a student's college assessment percentile (among test-takers) and his or her family's income percentile (among families who a child aged 17 who is approximately on-grade). This percentile-percentile correlation is about 0.37 for the nation-the reason that we first use a 0.4 correlation. U.S. states' correlations generally fall within 0.2 of the nation's.

[9] If we normalize the nation's mean achievement to 50 (as in the example), then low-achieving states like Mississippi are in the 30s and high-achieving states like Minnesota are in the 60s.

3. The Income Distribution of Academically Prepared Students Varies across States

In this section, we begin our proof by contradiction by demonstrating that differences in real universities' relevant pools generate substantial issues for the popular measures. As examples, we use the main campuses of the flagship universities of Connecticut (Storrs), Maine (Orono), Illinois (Urbana-Champaign), Montana (Missoula), New Mexico (Albuquerque), and Wisconsin (Madison). We chose these universities because their relevant pools are distinct in ways that affect measurement. Since the main contributions of this analysis are the proof by contradiction and demonstration that sound measures are possible, we needed to choose interesting universities, not average ones. Our aim is certainly not to rank all universities—indeed, we deliberately refrain from doing so because it is a university's responsibility (and not the prerogative of outside economists) to define its mission and, thereby, its relevant pool.

This analysis employs statistics from de-identified tax data and the population of college test takers that were constructed for use in Hoxby and Avery (2013) and Hoxby (2015a). The statistics are for the high school class of 2008.[10]

To construct each university's relevant pool, we include all students from the state whose scores on a college assessment put them in their flagship's "core" preparation range. Universities report these core ranges—the 25th and 75th percentiles of the scores of their students—to the U.S. Department of Education and college guides. While selective universities consider multiple indicators of preparation and most practice holistic admissions, these core ranges efficiently summarize academic standards and are comparable across geographic areas as, for instance, letter

[10] This facilitates comparison with Chetty et al (2017) who use approximately the same cohort but lack data on academic preparation.

grades are not. (Our use of test scores should not be taken as indicating that we endorse their exclusive or formulaic use in admissions. Nor do we suggest that a university should define its relevant pool only in terms of test scores or geography.[11] This is an illustration, not a policy prescription.)

Figure 1 shows the relevant pool's income distribution for the flagship universities. The 20th, 40th, 60th, and 80th percentiles of each income distribution are marked to facilitate comparisons. For instance, compare the University of Connecticut and University of Maine distributions. Maine's 20th percentile is much lower than Connecticut's 20th percentile. In fact, Connecticut's 20th percentile is approximately the same as Maine's 40th percentile. Connecticut's 40th percentile is midway between Maine's 60th and 80th percentiles. The Illinois-Montana comparison of relevant pools generates similar insights. Illinois' 20th percentile is higher than Montana's 40th percentile, and Illinois' 40th percentile is between Montana's 60th and 80th percentiles. Clearly, if one sets any low-income threshold based on a national distribution, as the Pell and Bottom Quintile measures do, a larger share of Maine's or Montana's relevant pool will fall below it. These comparisons illustrate how universities could be penalized for facing higher income distributions (Connecticut, Illinois) or rewarded for facing lower ones (Maine, Montana, New Mexico).

The University of Wisconsin's relevant pool is interesting because the state of Wisconsin has a relatively equal income distribution. (Notice that although Wisconsin's 40th, 60th, and 80th percentiles are well below those of Connecticut and Illinois, Wisconsin's 20th percentile is about the same as theirs.)

[11] Any observable student characteristic could be used be used to construct a pool of prospective applicants. For instance, a university could use students' high school grades, performance on their state's mandatory examinations, reported postsecondary goals, etc. If a university concludes that tests are biased against certain groups, it could construct its pool using test score ranges that differ by group.

Wisconsin's income equality translates into relatively few students with very low incomes by national standards. Thus, Wisconsin's relatively equal income distribution—which is probably good for disadvantaged students—generates penalties for the university when it is evaluated on Bottom Quintile or Pell measures. Ironically, the university would be less penalized if Wisconsin had more unequal incomes—as does California, say.

4. Relevant-Pool Based Measures of a University's Success in Enrolling Students from all Income Backgrounds

By incorporating information on each university's relevant pool, we can address the measurement challenge and create a metric that measures what it intends to measure—namely, universities' effort rather than their circumstances. In this section, we illustrate this measure using the flagship universities we selected as examples.

Figure 2 illustrates how the universities' in-state enrolled students' income distributions fit into the income distributions of their relevant pools. Specifically, we compute what percentage of each university's in-state students fall into each of the relevant pool's 5-percentile-wide bins. If the university is enrolling students of all incomes equally, each bin will contain 5 percent of students. We divide the bin's percentage by 5 so that the number 1 is a useful marker on the "measuring stick." For instance, if the height of the 20th to 25th percentile bin is 1, then the university's representation of enrolled students from the 20th to 25th percentiles is exactly the same as their representation in the relevant pool. If the height is 1.5, the university's representation of enrolled students is 50 percent greater than their representation in the relevant pool. If the height is 0.5, its representation is 50 percent lower.

Although 1 is useful marker, it is just a marker—not a mission we impose on schools. For instance, Berea College and

many other universities—public and private—that have a mission to serve disadvantaged students especially might want to see numbers above 1 for low-income students. A flagship university might be unconcerned if its numbers were less than 1 for high-income students—especially if the school were aware that it offered opportunities to high-income students but that some chose to attend private universities with comparable curricula at their own expense (saving taxpayers' money, thereby, for potential reallocation to needier students).

At each of the universities of Connecticut, Illinois, and Wisconsin, the height of the bars is consistently above 1 for enrolled students from low-income backgrounds—up through at least the 40th percentile of the relevant pool's income distribution. In other words, these universities recruit low-income students sufficiently effectively that such students' representation is disproportionately large. In contrast, the height of the bars is consistently below 1 for enrolled students from low-income backgrounds at the universities of Maine, Montana, and New Mexico-indicating that low-income students' representation is disproportionately small.

At the universities of Connecticut, Illinois, and Wisconsin, the height of the bars for middle-income students is about 1, indicating that their representation is similar to their representation in the relevant pool. At the universities of Maine, Montana, and New Mexico, the height of the bars for middle-income students is consistently above 1, indicating that middle income students' representation is disproportionately large. Recall that, for the same schools, low-income students' representation was disproportionately small.

At all six universities, the height of the bars tends to be below 1 for high-income students. Although we cannot be sure, this is probably not due to the flagships' failing to provide upper-income students with opportunities but, rather, those students choosing to attend private universities at their own expense.

These examples show the key advantages of our method:

1. Unlike threshold-based metrics like the Pell, Bottom Quintile, and IGM measures, our method shows how each university is enrolling students across the whole of its relevant pool's income distribution. This comprehensiveness allows observers to take in the entire picture or focus on whatever part of the income distribution interests them.
2. Our method provides a measuring stick but does not impose a mission on a university. A university can choose its own targets across the income distribution—which may include enrolling low-income students disproportionately
3. Our method does not encourage perverse behavior such as neglecting students just above an arbitrary income threshold. This is unlike the popular measures that make such students—who may need substantial financial aid and encouragement—fail to count towards a university's ranking. Moreover, when—as in proposed federal legislation—all universities face rewards and penalties based on the same threshold, there is increased likelihood of an "arms race" to enroll threshold-eligible students (e.g. Pell students), exacerbating any tendency to focus aid on them at the expense of other modest-income students.

Two comments are in order. First, because there is year-to-year variation in a university's applicants and relevant pool, a university might compromise its academic standards if it tries to achieve certain income representation targets each year, exactly. A university might want to employ moving averages or confidence intervals.[12]

Second, a university might wish to assess the extent to which it has exhausted the pool of relevant students

[12] Kane and Staiger (2002) note that schools overinterpreted year-to-year movements on accountability measures.

or—alternatively—"left some on the table." It could do this by constructing a simple variant of Figure 2 that shows its "utilization rate": its number (not percentage) of enrolled students in each bin divided by the relevant pool's number of students in that bin. To assess pool exhaustion, the university would then need to consider the size of its class relative to its market. This is best explained with examples.

Suppose that the University of Wyoming were assessing whether it had exhausted its pool. Since it is the only baccalaureate-granting public university in a state that has only one (tiny) private baccalaureate-granting institution, it might look for utilization rates fairly close to one as indicating exhaustion. (A rate of one would be overexhaustion because some Wyoming students attend out-of-state.) In contrast, a utilization rate that would indicate exhaustion for the University of California-Berkeley or University of California-Los Angeles would be well below 0.5. These two flagships share the same relevant pool and, moreover, have a pool that overlaps with those of numerous other public and private institutions in California and the West. And this is before accounting for California students' tendency to attend out-of-state.

5. Proof by Contradiction QED: Popular Measures Generate Rankings in Which Schools that Disproportionately Enroll Low-Income Students from their Pools are "Bottom Performers" while Schools that Fail to Do So are "Top Performers"

Acknowledging that threshold-based measures and rankings based on them are fundamentally flawed, it is nevertheless important to our proof by contradiction to compare rankings based on universities' relevant pools to rankings based on the Pell or Bottom Quintile measures to illustrate the magnitude of mismeasurement. We ranked all 50 flagship universities on the shares of their enrolled students whose family incomes fall below

the 20th and 40th percentiles of the relevant pool distribution. We also ranked the universities using the Pell and Bottom Quintile measures. The rankings are such that 1 is the "best" at enrolling low-income students according to the measure being used, and 50 is the "worst." See Table 2.

The University of Illinois is ranked 2nd best on both of the relevant-pool-based measures. However, it is ranked 36th on the Pell measure and 26th on the Bottom Quintile measure. Similarly, the universities of Connecticut and Wisconsin are among the several best on the relevant-pool-based measures. However, they are in the bottom fifteen schools on the Pell and Bottom Quintile measures.

Despite the fact that low-income students are well-represented in relation to the relevant pools at these three universities, policies based on the popular measures would punish these universities in various ways, because the mean income of their relevant pools is high, because their income distribution is relatively equal, or because of both.

The University of Montana is ranked 47th and 40th on—respectively—the first and second relevant-pool-based measures. However, it is ranked 3rd on the Pell measure and 7th on the Bottom Quintile measure. Similarly, the universities of Maine and New Mexico rank among the bottom fifteen schools on the relevant-pool-based measures but rank in the top five on the Pell and Bottom Quintile measures. Therefore, despite their own states' low-income students being underrepresented at these universities, policies based on the popular measures would reward the universities of Montana, Maine, and New Mexico because they face relevant pools with low incomes, relatively unequal income distributions, or both.

For our proof by contradiction, we selected six universities to demonstrate that measurement matters. Of course, there are universities that rank somewhat similarly regardless of whether we use the relevant-pool-based, Pell, or Bottom Quintile

measures. This is because some states happen to have income-achievement distributions in their relevant pools that are fairly similar to the national distribution. Such states' similar rankings, across the measures, are not a reason to endorse the Pell or Bottom Quintile measures. Observing that the measure does not affect these states' rankings much is akin to observing that it would not matter how we measured height if everyone were of average height.

6. The Intergenerational Mobility Measure

The increasingly popular intergenerational mobility (IGM) measure has received ample attention, including favorable coverage in The New York Times, and has been presented as a measure of the effect of universities on the economic success of low-income students. The Intergenerational Mobility or IGM measure is calculating by multiplying a university's Bottom Quintile measure by an estimate of the probability that the university's students from the national bottom income quintile end up, as adults, in the national top quintile ("Bottom-Top Mobility").[13]

The IGM measure has two problems. First, the IGM measure is dominated not by Bottom-Top Mobility—as one might think, given the "intergenerational" in its name. Rather, two-thirds of the variation in the IGM measure is generated by variation in the Bottom Quintile measure. Thus, the IGM measure is something of a "Bottom Quintile `Plus'", rather than a benchmark that measures something different (or mainly different). As a result, the IGM measure shares all the issues that affect the Bottom Quintile measure.[14] Second and importantly, the IGM measure

[13] As noted at greater length in footnote 5, the Bottom Quintile threshold used in the IGM measure in Chetty et al (2017) is not at the 20th percentile among students reasonably likely to obtain a high school diploma on-time. Rather, it is approximately the 16th percentile among students who are not drop-outs and not institutionalized.
[14] The correlation between the Bottom Quintile and the IGM measure is 0.65 whereas the correlation between bottom-top mobility and the IGM measure is only 0.29.

has peculiar problems that flow through Bottom-Top Mobility.

The University of Wisconsin exemplifies the problems. We have seen that because the state of Wisconsin has an unusually equal income distribution, it has relatively few prospective students who fall into the national bottom quintile. Thus, despite the university's success in enrolling low-income students from its pool, its ranking on the Bottom Quintile measure is poor. But, a second implication of the state's unusually equal income distribution is that it has comparatively few adults in the national top income quintile. Thus, University of Wisconsin students are disproportionately unlikely to end up in the national top quintile if they stay in Wisconsin, regardless of the income with which they grew up. Thus, the IGM measure penalizes Wisconsin's income equality twice: once through the Bottom Quintile measure and again through Bottom-Top Mobility. Thus, it should be no surprise that the University of Wisconsin ranks 22nd out of the 25 highly selective public colleges for which this measure is reported via the New York Times. Ironically, the University of Wisconsin would be less likely to suffer IGM-based penalties if the state of Wisconsin had less equal incomes.

For universities that face unusually unequal distributions, the situation is reversed. For instance, the California flagships face a relevant pool with large percentages of students at both the very top and very bottom of the income distribution. This is because California is a state that, by almost any measure, has one of the highest levels of income inequality.[15] Its income distribution exhibits strikingly fat tails. Thus, the IGM (especially) and other popular measures "reward" the California flagships for their state's income inequality. Unless the intended use of the IGM measure is to provide incentives for increased

[15] Regardless of which commonly used measure of income inequality is employed-the Gini Coefficient, the Atkinson Index, the Theil Index, Relative Mean Deviation-California is always among the top five most unequal states. See Mark W. Frank (2019).

income inequality, its construction is problematic.

7. Discussion: Pursuing Educational Missions and Providing Opportunities Regardless of Background

So far, we have emphasized the good intentions behind initiatives to measure the degree to which universities are providing opportunities for low-income students. However, as we stated in the opening paragraph, well-being and economic growth tend to be maximized when students—regardless of family income—are provided with opportunities that allow them to make optimal investments in their own education. For many colleges and universities, providing need-based financial aid to enable well-qualified students to attend regardless of family circumstances is a matter of enlightened self-interest, not simply a response to external pressures. For instance, Hoxby (2009) and Hoxby (2015b) argue on the basis of, respectively, historical evidence and economic theory that the universities that have been most successful in making admissions decisions to "craft a class" of academically outstanding students from a diversity of backgrounds without regard to ability-to-pay-tuition have gained in resources and innovative capacity and are often regarded as the best in the world.[16] Thus, even if a university were to lack good intentions and to lack motives to pursue society's goals, its own self-interest might induce it to engage in rigorous self-examination, asking itself whether it was providing opportunity regardless of family income.

The autonomy, decentralization and variation in mission of colleges and universities have been widely hailed as strengths that distinguish the U.S. market; public and non-profit institutions span local, regional and national markets. While this

[16] See Epple et al. (2018) for additional economic theory on the reason why selective institutions have incentives to recruit high-achieving students regardless of family income.

extraordinary diversity of mission and market geography is often celebrated, it also severely limits the extent to which "one-size-fits-all" performance standards or accountability measures such as Pell Shares or Bottom Quintile measures can provide constructive incentives or useful information. As we have shown with examples based on public flagship universities, states differ markedly in the level and inequality of incomes resulting in differences in the popular measures that are often unrelated to a university's success (or lack thereof) in drawing students from across the income distribution of its relevant pool. Moreover, colleges and universities differ significantly in mission including the academic requirements of the curricula or the emphasis on serving particular student populations such as students from Appalachia in the case of Berea or students with particular interests in science (such as California Institute of Technology and Massachusetts Institute of Technology). Given constraints, which include overall enrollment capacity and funding available for student aid, efforts to use the Pell Share or the Bottom Quintile measures as a cudgel to change behavior through either regulation or external pressure are likely to distort universities' behavior, ultimately reducing the capacity of the higher education sector to serve as an engine of opportunity for the long run. Indeed, given geographic differences combined with differences in mission among colleges and universities, optimizing institutions (and those operating in the public interest) would be expected to differ markedly on the popular measures.

Although low-income students are less likely to be high achievers than high-income students are, nevertheless more of low-income high achievers exist than are enrolled at selective universities when those universities are considered as a group (Hoxby and Avery, 2013). This now well-known aggregate result is too often misinterpreted to imply that all selective universities can achieve the same absolute representation of low-income

students. How successful a university is at enrolling low-income students or the potential for increasing enrollment of low-income students cannot consistently be revealed by the popular measures—as we have demonstrated.

Furthermore, the main takeaway from Hoxby and Avery (2013)—which many readers seem to miss—is that the "missing" students cannot be identified and recruited without rigorous data analysis. For instance, admissions staff often reach out only to secondary schools with high concentrations of students who participate in the national free lunch program: such crude targeting, which is akin to using the Pell Share or Bottom Quintile measure, allows institutions to find only a small fraction of the low-income students in their relevant pool. Similarly, in Hoxby and Turner (2013), we found that the low-income high achievers who were induced to enroll in selective universities by an informational intervention were exactly those students whom universities tended to overlook with crude identification methods. We identified the students in question using sound data analysis, not shortcuts akin to the popular measures.

In short, well-intentioned commentators and leaders appear to be ignoring one of the most important conclusions of recent research—namely that true improvements on providing opportunity can be attained only with sound data analysis. Measurement does matter. The "quick and dirty" popular measures can generate "dirty" incentives and policies because they confound differences in universities' effort with differences in their circumstances.

As we have demonstrated, differences in the income-achievement distributions faced by universities can produce Pell/Bottom Quintile/IGM-based penalties and rewards that are not only unintended but even the inverse of what was intended. Recall the example of the University of Wisconsin which faces a relevant pool with an unusually equal income distribution. Facing strong policy incentives or public pressure

to improve on the popular measures could produce distortions such as (i) enrolling less prepared students who meet the Pell or Bottom Quintile threshold even when there are much better-prepared students just above the threshold, (ii) substituting out-of-state students who meet the threshold for in-state students, (iii) encouraging graduates to migrate to less equal states where their earnings are more likely to be in the top quintile. In other words, a university that pursues the popular measures may find that the easiest way to attain a better ranking is to deviate substantially from its educational mission.

A university that evaluated its success in enrolling students from low- and moderate-income families using measures that assessed outcomes relative to the relevant pool, as presented in our analysis, would not find a conflict between pursuing its educational mission and providing opportunities to students regardless of background. If universities use metrics that measure what is intended, they can further both equity and excellence simultaneously.

8. References

Chetty, Raj, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility," The Equal Opportunity Project Working Paper.

Dynarski, Susan, Katherine Michelmore, and Carmello Libassi. 2018. "Increasing Economic Diversity at a Highly Selective University: Results from a Large Field Experiment," University of Michigan manuscript.

Education Trust. 2014. Tough love: Bottom-line quality standards for colleges. Washington, DC: Author.

Epple, Dennis; Richard Romano, Sinan Sarpca, Holger Sieg and Melanie Zaber. 2018. "Market Power and Price Discrimination in the U.S. Market for Higher Education." RAND Journal of Economics (forthcoming).

Frank, Mark W. "U.S. State-Level Income Inequality Data." Internet site https://www.shsu.edu/eco_mwf/inequality.html, accessed January 2019.

Heckman, James J., Carolyn Heinrich and Jeffrey Smith. "The Performance Of Performance Standards," Journal of Human Resources, 2002, v37 (4,Fall), 778-811.

Hoxby, Caroline M. 2009. "The Changing Selectivity of American Colleges." Journal of Economic Perspectives, 23 (4): 95-118.

Hoxby, Caroline M. 2015a. "Computing the Value-Added of American Postsecondary Institutions," Internal Revenue Service Statistics of Income working paper.

Hoxby, Caroline M. 2015b. "Endowment Management Based on a Positive Model of the University," in How the Financial Crisis and Great Recession Affected Higher Education, Jeffrey R. Brown and Caroline M. Hoxby, editors. Chicago: University of Chicago Press.

Hoxby, Caroline M., and Christopher Avery. 2013. "The Missing "One-Offs": The Hidden Supply of High-Achieving, Low Income Students." Brookings Papers on Economic Activity, 2013 (1): 1-65.

Hoxby, Caroline M. and Sarah Turner. 2013. "Expanding College Opportunities for High-Achieving, Low Income Students" SIEPR Discussion Paper 12-014.

Kane, Thomas, J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." Journal of Economic Perspectives, 16(4): 91-114.

Leonhardt, David. 2017. "College Access Index Methodology." The New York Times. (May 26) https://www.nytimes.com/2017/05/26/opinion/2017-college-access-index-methodology.html

Leonhardt, David. 2017. "Virginia, Betraying Jefferson" The New York Times. (September 18).

National Conference of State Legislatures. 2018. "Performance Based Funding for Higher Education." Retrieved from: http://www.ncsl.org/research/education/performance-funding.aspx (April 25).

Sallee, James, Alexandra M. Resch, and Paul Courant. 2008. "On the optimal allocation of students and resources in a system of higher education." B.E. Journal of Economic Analysis and Policy, 8(1).

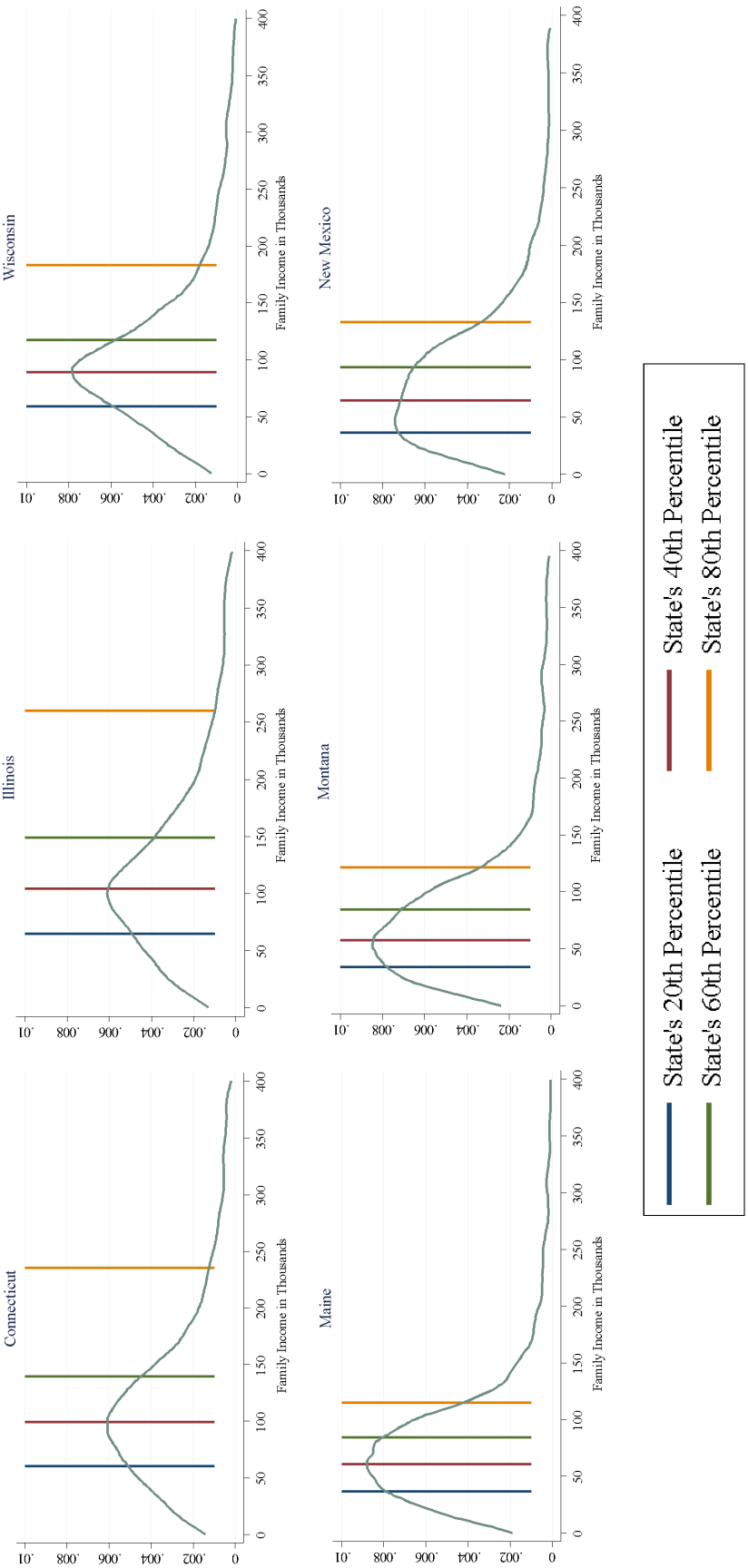
Tebbs, Jeffrey and Sarah Turner. 2005. "Low-Income Students: A Caution about Using Data on Pell Grant Recipients." Change: The Magazine of Higher Learning.

The New York Times, The Editorial Board. 2014. "Tying Federal Aid to College Ratings" (June 25) Page A22.

U.S. Department of Education. 2018. "For Public Feedback: A College Ratings Framework." Retrieved from https://www.ed.gov/collegeratings (April 25).

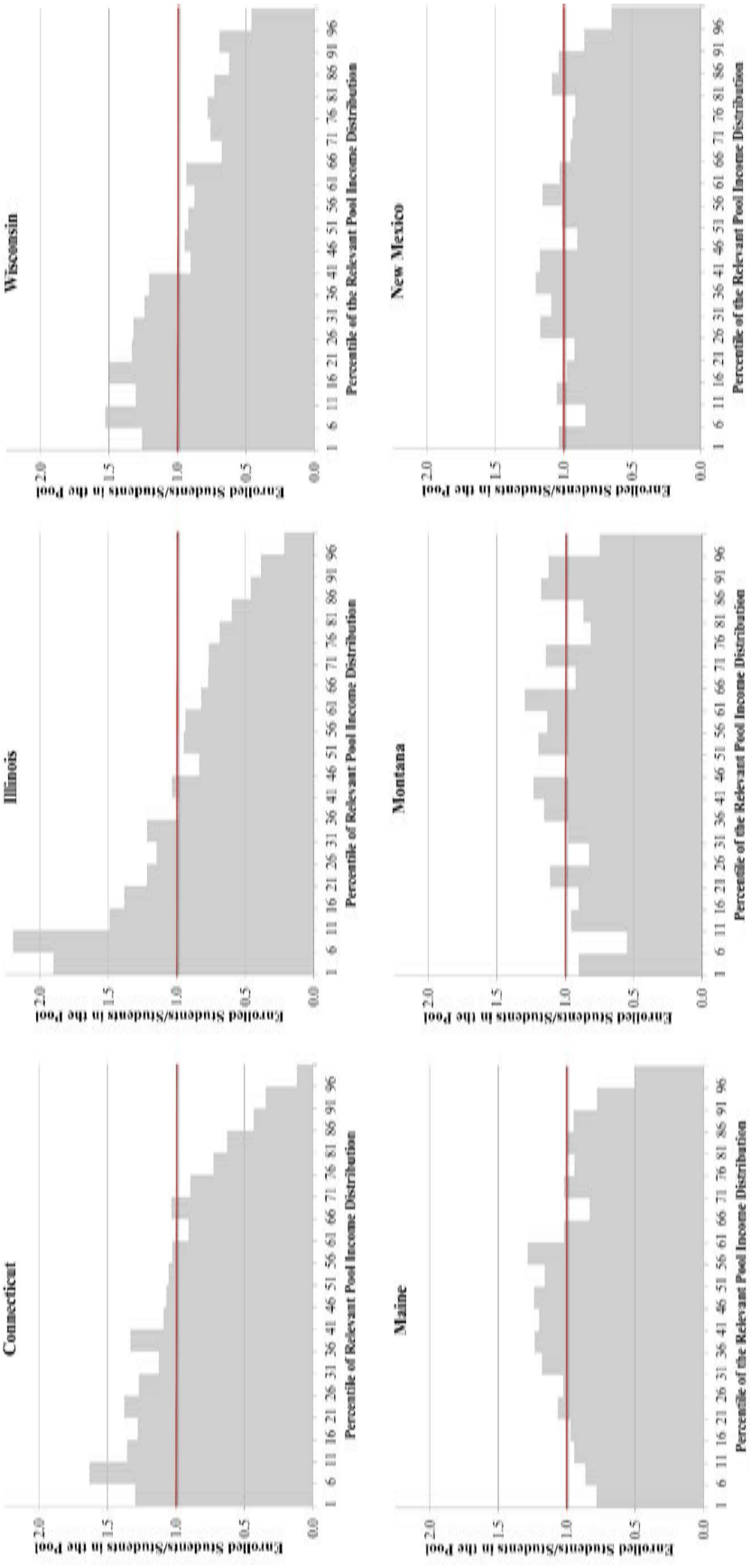
Washington Monthly. 2017. "The 2017 College Guide." Vol 49 No 9/10 (September/October) https://washingtonmonthly.com/2017college-guide?ranking=2017-rankings-national-universities

Figure 1
Income Distributions of Students who are in the Core Preparation Range of their State's Flagship University
Connecticut, Maine, Illinois, Montana, Wisconsin, New Mexico



Notes: Each figure shows the distribution of family income for students in a state whose test scores put them in the core preparation range of their state's flagship university. Family income in thousands of dollars is on each horizontal axis. Density is on each vertical axis. The vertical lines in each figure mark the 20th, 40th, 60th, and 80th percentiles of the income distribution.

Figure 2
How Enrolled Students' Income Distribution Fits into the Income Distribution in the Relevant Pool
Flagship Universities of Connecticut, Maine, Illinois, Montana, Wisconsin, New Mexico



Notes: Each figure shows the distribution of family income of the students at a state's flagship university fit into the distribution of family income of the students in the state who are in the university's core preparation range (the relevant pool). The bars are based on 5-percentile ranges of the relevant pool's income distribution. The horizontal line set at 1 is an equal representation marker (see text).

Table 1. Simulations of Low-Income Students' Representation in the Pool Relevant to a University When Its Pool's Income and College Preparation Differ from National Norm

	College Preparation	Family Income	Overall Percent Below the National 20th Percentile in Income	Percent with Income Below the National 20th Percentile and College Preparation At or Above the 25th percentile	50th percentile	75th percentile	90th percentile
Baseline							
mean	50.1	50.1	20.0%	14.7%	11.1%	7.6%	5.1%
standard deviation	50.0	49.9					
Increase Mean Income							
mean	50.1	70.1	10.7%	7.0%	4.9%	3.0%	1.9%
standard deviation	50.0	49.9					
Decrease Mean Income							
mean	50.1	30.1	32.9%	26.3%	21.1%	15.5%	11.0%
standard deviation	50.0	49.9					
Decrease Variance of Income							
mean	50.1	50.1	13.2%	8.9%	6.3%	4.0%	2.6%
standard deviation	50.0	37.4					
Increase Variance of Income							
mean	50.1	50.1	25.0%	19.1%	14.8%	10.5%	7.2%
standard deviation	50.0	62.4					
Increase Income-Achievement Correlation from 0.4 to 0.6							
mean	50.1	50.1	20.0%	11.3%	6.4%	2.8%	1.2%
standard deviation	50.0	50.0					
Decrease Mean College Preparation							
mean	30.1	50.1	20.0%	12.6%	8.9%	5.9%	3.8%
standard deviation	50.0	49.9					

Notes: Baseline estimates assume a bivariate normal distribution of income with the indicated means and standard deviations of income and college preparation. The baseline income-achievement correlation is 0.4. Estimates are from a simulation with 500,000 draws from the indicated distribution.

Table 2
Rankings Based on Relevant Pool Indicators Versus Rankings Based on Pell Share and Bottom Quintile Measure

	Rank Based on...			
	Percent Below the 20th Percentile in the Relevant Pool	Percent Below the 40th Percentile in the Relevant Pool	Pell Share	Bottom Quintile Measure
University of Illinois	2	2	36	26
University of Wisconsin	6	4	48	36
University of Connecticut	7	5	35	41
University of New Mexico	34	30	2	1
University of Maine	42	34	5	3
University of Montana	47	40	3	7

Notes: The table shows a comparison between rankings based on universities' relevant pools and rankings based on the Pell or Bottom Quintile measures. This is important to our proof by contradiction. We ranked all 50 flagship universities on the shares of their enrolled students whose family incomes fall below the 20th and 40th percentiles of the relevant pool distribution. We also ranked all 50 universities using the Pell and Bottom Quintile measures. The rankings are such that 1 is the "best" at enrolling low-income students according to the measure being used, and 50 is the "worst."