The Education-Innovation Gap*

Barbara Biasi† Song Ma‡

March 29, 2021

Abstract

Higher education can play a crucial role for innovation and growth by by providing people with the frontier knowledge that is necessary to innovate. This paper proposes a novel approach to quantify the provision of this type of knowledge across and within schools: the education-innovation gap, a measure of the distance between the content of a university course and frontier knowledge. We calculate the gap for 1.75 million courses taught at US institutions by mapping the text of their syllabi to the text of 20 million old and new academic publications. We show evidence of substantial variation in the gap, both across and within schools. In particular, the gap is smaller in more selective schools and those serving fewer disadvantaged or minority students. The gap of a given course significantly declines when a new instructor takes over the course, and more research-active instructors teach courses with smaller gaps. Importantly, these differences bear a strong relationship with students' graduation rates and incomes after graduation, especially in less selective schools and those serving more disadvantaged students.

JEL Classification: I23, I24, I26, J24, O33

Keywords: Education, Innovation, Text Analysis, Inequality

First Draft: September, 2019. We thank David Deming, David Robinson, Kevin Stange and seminar and conference participants at AEA Annual Meetings, Duke, Erasmus, Maastricht, NBER (Entrepreneurship), Junior Entrepreneurial Finance and Innovation Workshop, and Yale (Applied Economics, Finance) for helpful comments. Xugan Chen provided excellent research assistance. We thank Yale Tobin Center for Economic Policy, Yale Center for Research Computing, Yale University Library, and Yale International Center for Finance for research support. All errors are our own.

†Yale School of Management and NBER, barbara.biasi@yale.edu, +1 (203) 432-7868;

‡Yale School of Management, song.ma@yale.edu, +1 (203) 436-4687.

1 Introduction

Endogenous growth theory has modeled human capital as a key component of the production of knowledge and non-rival new ideas, which fuel growth (Romer, 1990). Education systems, and particularly higher education, play a crucial role in equipping individuals with the skills necessary to produce new ideas and innovate (Biasi, Deming, and Moser, 2020). As the burden of knowledge required to innovate increases over time (Jones, 2009) and new ideas become harder to find (Bloom et al., 2020), the importance of these programs is likely to grow.

Just like there are differences in the economic returns to different types of education programs (Hoxby, 1998; Altonji et al., 2012; Chetty et al., 2019, among others), there might be differences in the extent to which programs are effective at fostering innovation.[1] For example, recent work has found that the majority of US inventors come from a very small set of elite schools (Bell et al., 2019). Whether these cross-school differences purely reflect student selection or whether they bear a relationship with the type of knowledge and skills students acquire, however, remains a "black box."

This paper tries to open this black box by examining the content of higher education instruction. Using a novel approach, we quantify the extent to which higher education systems are able to equip students with the valuable knowledge that leads to innovation and growth. To do so we develop a new measure: the *education-innovation gap*, which captures the distance between the content of college and university courses (described by course syllabi) and frontier knowledge, captured by academic articles published in top journals in recent times. The intuition behind this measure is the following: A course has a smaller gap if its content is more similar to newer articles than it is to older ones. For example, if the word *Julia* (the name of a popular programming language) is more frequently used in newer Computer Science articles than the word *Visual Basic*, an introductory programming course that teaches *Julia* in 2018 would be closer to the knowledge frontier than one that teaches *Visual Basic*, and it should have a smaller gap.

We construct this measure using a "text as data" approach applied to a very large set of course syllabi and academic papers. Specifically, we compare the text of 1.75 million college and university syllabi of about 570,000 courses in 62 different fields, taught at nearly 800 US institutions between 1998 and 2018, with the text of the abstract of 20 million academic publications published in top journals since each journal's foundation. To do this, we first project the text of each document (a

[1] Differences exist across fields as well: STEM programs, for example, have been shown to increase invention (Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019).

syllabus or an article) on an existing dictionary, such as the list of all words listed on *Wikipedia*. This allows us to represent each document as a binary vector, whose elements correspond to dictionary words. To better capture the relevance of each term for the content of a document, similarly to Kelly et al. (2018) we weigh each word by its relative frequency in the document relative to the pool of all older documents. We then compute the text similarity between each syllabus and each article, measured as the cosine similarity between each document's weighted word vector.

Using these measures of text similarity, we construct the education-innovation gap of a syllabus as the ratio between the average cosine similarity with articles published 13 to 15 years prior to the syllabus and the similarity with articles published one to three years prior. Intuitively, the gap is larger for syllabi that are more similar to older, rather than newer, knowledge. Importantly, by virtue of being constructed as a *ratio* of cosine similarities, the gap is not affected by idiosyncratic attributes of each syllabus, such as length, structure, or writing style.

Several empirical patterns confirm the ability of the education-innovation gap to capture a course's distance from the knowledge frontier. First, the gap is strongly correlated with the average "age" of articles and books listed in the syllabus as required or recommended readings. Second, graduate-level courses have the smallest gap on average; advanced undergraduate courses have the second smallest gap, and basic courses – more likely to teach the fundaments of the discipline rather than the latest research – have the largest gap. Third, gradually replacing "older" knowledge words with "newer" ones, as we do in a simulation exercise, progressively reduces the gap.

On average, the education-innovation gap is equal to 0.95, which indicates that courses tend to be more similar to newer research than to older one. However, substantial variation exists among syllabi. To move a syllabus from the 25th percentile (0.87) to the 75th percentile (0.98) of the gap distribution, approximately 22 percent of its content would have to be replaced with newer knowledge. Differences across fields explain 13 percent of the total variation, differences across schools explain an additional 3 percent, differences across courses within each school explain 57 percent, and differences among instructors explain an additional 13 percent of the variation.

This decomposition exercise suggests the existence of significant differences in the extent to which students at different schools are exposed to up-to-date knowledge. In particular, our data indicate that more selective schools (such as Ivy-Plus, Chetty et al., 2019) have a significantly smaller gap compared to less selective or non-selective schools: To make the average syllabus in non-selective schools comparable to the average for Ivy-Plus schools, 4 percent of its content would have to be replaced with newer knowledge.

More and less selective schools serve different populations of students, in terms of ability but also parental background (Chetty et al., 2019). If frontier knowledge is better received by students with higher ability, a smaller gap in more selective schools might reflect a school's attempt to provide students with better tailored educational content. We find, however, cross-school disparities in the gap even within selectivity tiers. Even controlling for selectivity, schools that enroll more students with a higher parental income have a significantly smaller gap. Similarly, the gap is larger in schools that enroll more students who belong to a racial or ethnic minority. These results reveal significant disparities in access to up-to-date content across students with different backgrounds.

The decomposition exercise also reveals a crucial role for instructors in determining the content of the courses they teach.[2] To more appropriately quantify this role, we show that the education-innovation gap of a course declines significantly when the instructor changes, which in turn suggests that people who take over a course previously taught by someone else substantially update its content. We also show that the gap is negatively correlated with measures of research productivity, such as the number of articles published in journals each year and the number of citations received by these publications. This finding might be due to more research-active instructors being better updated about the frontier of research, and more likely to cover this type of content in their courses. Instructors are also more likely to teach frontier knowledge when they teach a course that is closer to their research expertise. These results indicate that the assignment of instructors to courses can be a powerful tool to expose students to up-to-date content.

Having documented differences across schools and courses in the education-innovation gap, a natural question is: Do these differences matter for student outcomes? To answer this question, the ideal experiment would randomly allocate courses with different gaps across and within schools and compare the outcomes for students exposed to larger and smaller gaps. In the absence of this random variation we characterize the empirical relationship between the gap, measured as a school-level residual conditioning on field, course level, and year effects, and student outcomes such as graduation rates, incomes after graduation, and intergenerational mobility. To account for endogenous differences across schools we control for selectivity, parental income, and the share of enrolled students who are minority. We find that the gap is negatively related to graduation rates and incomes ten years after graduation; the relationship with intergenerational mobility is instead indistinguishable from zero.

Importantly, these patterns are almost entirely driven by less selective schools and those schools

[2]This finding is analogous to the finding of a large role for public-school teachers on the growth in achievement (Rockoff, 2004; Chetty et al., 2014).

serving students from lower-income families. In fact, the correlation between the gap and graduation rates and income is indistinguishable from zero in Ivy-Plus and elite schools and those serving more wealthy students, while it is negative and large for non-elective schools and those with a less wealthy student body.

Taken together, our findings indicate that the novelty of the academic content of university courses varies significantly across and within schools, relates to attributes of the instructors, and is correlated with students' outcomes. In the last part of the paper we test whether these patterns hold when, rather than with academic content, we measure novelty with the intensity soft skills, defined as non-cognitive attributes that shape the way people interact with others (Deming, 2017; Deming and Kahn, 2018). We measure the intensity of soft skills of a course using the assignments portion of the syllabus, calculating the share of terms which refer to teamwork, group assignments, and presentations. We find that the intensity of soft skills is negatively related to the gap. Furthermore, we show that also the prevalence of soft skills increases with school selectivity and is larger for schools where the median parental income is in the top portion of the distribution.

This paper contributes to several literature strands. First, it is among the first use the text of syllabi as information to characterize the content of higher-education instruction, relating it to the frontier of knowledge. Similarly to Kelly et al. (2018), who use cosine similarities between the text of patent documents to measure patent quality, and Gentzkow and Shapiro (2010), who use the language of newspaper articles to measure media slant, we use text analysis techniques to characterize the content of each course and to link it to frontier technologies. Our approach is similar to Angrist and Pischke (2017), who use hand-coded syllabi information to study the evolution of undergraduate econometrics classes.

Second, we contribute to the literature that has characterized heterogeneity in the production of human capital by proposing a novel approach to measure the content of higher education. This allows us to relate this content to the characteristics of schools, instructors, and students, as well as to students' outcomes. Earlier works have highlighted the role of educational attainment (Hanushek and Woessmann, 2012), majors and curricula (Altonji et al., 2012), college selectivity (Hoxby, 1998; Dale and Krueger, 2011), social learning ad interactions (Lucas Jr, 2015; Lucas Jr and Moll, 2014; Akcigit et al., 2018) and skills (Deming and Kahn, 2018) for labor market outcomes, innovation, and economic growth. Our analysis focuses instead on the specific concepts and topics covered in higher education courses, and aims at measuring the extent to which these are up-do-date with respect to the frontier of knowledge.

Third, this paper relates to the literature on the "production" of innovation. Earlier works (Nelson and Phelps, 1966; Benhabib and Spiegel, 2005) have highlighted an important role of human capital for technology diffusion. More recently, Jones (2009) has shown how innovators are increasingly required to update their knowledge and skills to keep up with a fast-paced world; failure to do so delays the correlation between researcher/inventor life-cycle productivity (Jones, 2010; Jones and Weinberg, 2011). Technical and scientific education has been associated with more innovation and growth (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019).[3] Our paper contributes to this body of evidence by taking a more "micro" approach to quantify the extent to which higher education is able to provide students with the frontier knowledge that is necessary to innovate.

Lastly, our findings contribute to a growing body of evidence on the "democratization" (or lack thereof) of access to valuable knowledge. Bell et al. (2019) have shown that US inventors (measured as patentees) come from a small set of top US schools, which admit very few low-income students. We confirm that these schools provide the most up-to-date educational content, which in turn suggests that access to this type of knowledge is not equally distributed across the population. This finding is particularly relevant in light of the fact that up-to-date knowledge bears the strongest correlation with outcomes in schools outside of the elite.

2 Data

Our empirical analysis combines different types of data. These include the text of course syllabi, the abstracts of academic publications, information on US higher education institutions, and labor market outcomes for the students at these institutions. More detail on the construction of our final data set can be found in the Online Data Appendix.

2.1 College and University Course Syllabi

The text of a large sample of college and university syllabi was collected by the American Assembly's Open Syllabus Project (OSP).[4] The raw data cover more than seven million English-language

[3] The literature on the effects of education on innovation encompasses studies of the effects of the land grant college system (Kantor and Whalley, 2019; Andrews, 2017) and, more generally, of the establishment of research universities (Valero and Van Reenen, 2019) on patenting and economic activity.

[4] OSP collects data from a variety of sources, including publicly accessible university websites and archives, as well as personal websites of faculty members that list teaching materials. Voluntary faculty and student contributions make up a small portion of the collection. The main purpose of the Project is to support educational research and novel teaching and learning applications.

syllabi of courses taught in over 80 countries; the oldest syllabi date as far back as the 1960s.

Most syllabi share a standard structure. Basic details of the course (such as title, code, and the name of the instructor) are followed by a description of the content and material and a list of references and recommended readings for each class session. In addition, syllabi often contain information on the course's requirements (such as assignments and exams) and general policies regarding grading, absences, lateness, and misconduct. From the text of each syllabus, we extract four pieces of information: (i) basic course details, (ii) the course's content, (iii) the list of required and recommended readings for each class session, and (iv) information on assignments and exams.

Basic course details These include the name of the institution, the course's title and code, the name and email of the instructor, as well as the academic year in which the course is taught (e.g., Fall 2020).[5] Using names and codes we construct the course level, distinguishing between introductory, advanced, or graduate-level courses.[6] We also use information on the course's field as provided by OSP. Specifically, OSP assigns each syllabus to one of 62 detailed fields, e.g., English Literature, History, Computer Science, Economics, and Mathematics (see Online Data Appendix).[7] We further aggregate these fields into four broader areas: STEM, Humanities, Social Sciences, and Business.[8]

Course content To extract this information, essential for our analysis, we identify the portion of a syllabus that contains the course description by searching for section titles such as "Summary," "Description," and "Content."[9] This portion describes the basic structure of the course, the key concepts that are covered, and (in many cases) a timeline of the content and materials for each lecture. The average syllabus contains a course description of 1307 words; the median is 626.

List of references These include bibliographic information on the required and recommended readings for each class session. We identify this list by searching for section titles such as "References", "Book", "Guidebook", and "Textbook." We also collect other in-text citations such as "Biasi

[5] Information on course codes allows us to track how a given course changes over time. For example, we are able to study how the content of a course evolves when its instructor changes.

[6] For example, we distinguish between introductory courses (e.g., Intro to Microeconomics), advanced courses (e.g., Intermediate or Advanced Microeconomics), and graduate-level courses (e.g., PhD Microeconomics).

[7] The taxonomy of OSP draws extensively from the 2010 Classification of Instructional Programs of the Integrated Postsecondary Education Data System: https://nces.ed.gov/ipeds/cipcode/default.aspx?y=55.

[8] This aggregation is outlined in Appendix Table AI.

[9] The full list of section titles used to identify the course description contains: "Syllabi", "Syllabus", "Title", "Description", "Method", "Instruction", "Content", "Characteristics", "Overview", "Tutorial", "Introduction", "Abstract", "Methodologies", "Summary", "Conclusion", "Appendix", "Guide", "Document", "Module", "Apporach", "Lab", "Background", "Requirement", "Applicability", "Objective", "Archivement", "Outcome", "Motivation", "Purpose", "Statement", "Skill", "Competency", "Performance", "Goal", "Outline", "Schedule", "Timeline", "Calendar", "Guideline", "Material", "Resource", and "Recommend".

and Ma (2020)." We successfully identify this portion for 62 percent of all syllabi. For each reference item, we also collect information on full title, abstract, journal, textbook edition, and keywords using information from Elsevier's SCOPUS database (see Section 2.2 for additional details).

Assessed skills To gather information on the set of skills required by and assessed in the course, we use information on exams and assignments policies. We identify and extract the related portion of each syllabus by searching for section titles such as "Exam," "Assignment," "Homework," "Evaluation," and "Group."[10] Using the text of these sections, we distinguish between hard skills (assessed through exams, homework, assignments, and problem sets) and soft skills (assessed through presentations, group projects, and teamwork). We successfully identify this information in 95.7 percent of all syllabi.

Sample restrictions and description Panel A of Table 1 describes the characteristics of the syllabi sample. To maximize consistency across time, we focus our attention on syllabi taught between 1998 and 2018 in four-year US institutions with at least one hundred syllabi in our sample.[11] We exclude 35,917 syllabi (1.9 percent) with less than 20 words or more than 10,000 words (the top and bottom 1 percent of the length distribution). Our final sample contains about 1.75 million syllabi from 570,247 courses at 792 institutions. A syllabus contains an average of 2,218 words, with a median of 1,779. The content description, which we use to construct the education-innovation gap, contains 1,011 words on average.

2.2 Academic Publications

To compare the content of each course to frontier research, we compiled a data set of all peer-reviewed articles that appeared in the top academic journals of each field since the journal's foundation, using data from Elsevier's SCOPUS dataset.[12] We define as top journals those ranked among the top 10 by Impact Factor (IF) in each field at least once since 1975. Our final list of publications includes 20 million articles in the same fields as the syllabi, corresponding to approximately 100,000 articles per year.[13] We capture the knowledge content of each article with its title, abstract, and

[10]The full list of section titles used to identify the skills is as follows: "Exam", "Quiz", "Test", "Examination", "Final", "Examing", "Midterm", "Team", "Group", "Practice", "Exercise", "Assignment", "Homework", "Evaluation", "Presentation", "Project", "Plan", "Task", "Program", "Proposal", "Research", "Paper", "Essay", "Report", "Drafting", "Survey".

[11]We remove 129,429 syllabi from one online-only university, the University of Maryland Global Campus.

[12]We access the SCOPUS data through the official API in April-August 2019.

[13]SCOPUS classifies articles into 191 fields. To map each of these to the 62 syllabi fields, we calculate the cosine similarity (see Section 3) between each syllabus and each article. We then map each syllabi field with the SCOPUS field

keywords.

2.3 Course Instructors

We use information from Microsoft Academic (MA) to measure the research output of all people listed as instructors in the syllabi. For each researcher, MA lists publications, working papers, other manuscripts, and patents, together with the counts of citations to each of these documents. We link instructor records from the text of the syllabi to MA records using names and institutions; we are able to successfully match 38.93 percent of all instructors. Using data from MA, we measure each instructor's research quantity and quality.

Panel B of Table 1 summarizes our measures of instructor research productivity: the number of publications and the number of citations received by these articles, either per year (on average) or in the previous five years.[14] On average, instructors publish 1.4 articles per year and 4.5 articles in the most recent five years, and they receive 104 citations per year. The distribution of both citations and publication counts are highly skewed: The median instructor in our sample does not publish any article and, as a result, receives no citations.

2.4 Information on US Higher Education Institutions

The last component of our dataset includes information on all US universities and colleges where syllabi courses are taught. Our primary source of data is the the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES).[15] For each college or university, IPEDS lists a set of institutional characteristics (such as name and address, control, and affiliation); the types of degrees and programs offered; tuition and fees; characteristics of the student population, such as the distribution of SAT and ACT scores of all admitted students, enrollment figures for different demographic groups, completion rates, and graduation rates.We link each syllabus to the corresponding IPEDS record using a fuzzy matching algorithm based on the name of the institution. We are able to successfully link 89 percent of all syllabi.

We complement IPEDS data with information on schools and students from two additional sources. The first one is the dataset assembled and used by Chetty et al. (2019), which includes

with the highest average similarity.

[14]Using citations and publications in the previous five years helps address issues related to the life cycle of publications and citations, with older instructors having a higher number of citations and publications per year even if their productivity declines with time.

[15]IPEDS includes responses to surveys from all postsecondary institutions since 1993. Completing these surveys is mandatory for all institutions that participate in, or are applicants for participation in, any federal financial assistance programs.

school-level characteristics such as selectivity (defined using Barron's selectivity scale) and the incomes of students and parents, along with measures of of intergenerational mobility such as the share of students with income in the top quintile and parental income in the bottom quartile (from the universe of US tax records).

Our second second source of data is the College Scorecard Database of the US Department of Education, an online tool designed to help consumers compare costs and returns from attending various colleges and universities in the US. This database includes graduation rates of students by school and cohort and the incomes of graduates ten years after the start of the program. We use these variables, available for the academic years 1996-97 to 2017-18, to measure student outcomes for each school.

Panel C of Table 1 displays summary statistics of school-level variables for the colleges and universities of our syllabi sample. The median parental income is equal to $97,544 on average. Across all schools, the average share of students with parental income in the top 1 percent is 3 percent, with a standard deviation of 0.041. The share of minority students is equal to 0.15, with a standard deviation of 0.15. Graduation rates average 60.8 percent in 2018, whereas students' incomes ten years after school entry, for the 2003–04 and 2004–05 cohorts, are equal to $47,348. Students' intergenerational mobility, defined as the probability that students from the bottom quintile of parental income reach the top income quintile during adulthood, is equal to 0.29 on average.

2.5 Data Coverage and Sample Selection

Our sample of syllabi corresponds to a subset of all courses taught in US colleges and universities. The number of syllabi, syllabi per instructor, and syllabi per institution in the sample increases over time, indicating that sample coverage improves across the years (Appendix Figure AII). To better interpret our empirical results, it is useful to compare our sample to the population of all syllabi and to explore possible patterns of selection into the sample, overall and across time. To this purpose, we compiled the full list of courses offered between 2010 and 2019 in a subsample of 161 US institutions, representative of all institutions included in IPEDS, from hand-collected course catalogues in the archives of each school.[16] This allows us to compare the syllabi sample to the population for these schools and years.

[16]We begin our collection from the year 2010 because most universities started listing their catalogues online around this time. For an example of a course catalogue, please see https://registrar.yale.edu/course-catalogs. In the Appendix Table AII we provide a list of institutions for which we collected the catalogs, and we show that these institutions are representative of all IPEDS institutions (Appendix Table AIII).

The fraction of catalogue courses included in our sample is stable over time at 5 percent (Appendix Figure AIII). This suggest that, at least among the schools with catalogue information, the increase in the number of syllabi over time is driven by an increase in the number of courses that are offered, rather than an increase in sample coverage.

Next, we test whether selection into the sample is related to observables features of schools and fields. Our data reject this hypothesis. Appendix Figure AI shows that the field composition of our sample is similar to that of course catalogues, with STEM courses representing 25-35 percent of the sample, Humanities representing 30-40 percent, and the Social Sciences representing 25 percent of all syllabi in a year.

Second, Appendix Figure AIV shows that our sample does not disproportionately cover from certain geographic areas of the US. The number of institutions (panel (a)) and of syllabi in the sample (panel (b)) are similarly distributed across states.

Lastly, Table 2 illustrates more broadly that the share of courses in each school that are included in the syllabi sample is unrelated to a set of institutional characteristics, such as selectivity, finances, and enrollment. Panel (a) shows means and standard errors of the share of covered syllabi across selectivity tiers. In 2013, this share ranged from 0.03 percent for non-selective schools to 4.42 percent for highly selective and selective public schools (right columns); these shares are, however, statistically indistinguishable across tiers. The same is true for the 2010-2013 *change* in the share of covered syllabi (left columns). Panel (b) shows instead the correlation between the share of syllabi included in the sample and a set of financial attributes (such as expenditure on instruction, endowment per capita, sticker price, and average salary of all faculty), enrollment, the share of students in different demographic categories (Black, Hispanic, alien), and the share of students graduating in Arts and Humanities, STEM, and the Social Sciences. These correlations are all statistically indistinguishable from zero.

These findings indicate that our syllabi sample does not appear to be selected on the basis of observable characteristics of schools and fields. While these findings are reassuring, we are not able to test for selection driven by unobservables. Our results should therefore be interpreted with this caveat in mind.

3 Measuring the Education-Innovation Gap

To construct the education-innovation gap, we combine information on the content of each course, captured by its syllabus, with information on frontier knowledge, captured by academic publica-

tions. We now describe the various steps for the construction of this measure, provide the intuition behind it, and perform validation checks.

Step 1: Measuring Similarities in Text

To construct the gap, we begin by computing textual similarities between each syllabus and each academic publication. To this purpose, we represent each document d (a syllabus or an article) in the form of a vector $\tilde{V}_d$ of length $N_W = |W|$, where W is the set of unique words in a given language dictionary (we define dictionaries in the next paragraph). Each element w of $\tilde{V}_d$ equals one if document d contains word $w \in W$. To measure the textual proximity of two documents d and k we use the cosine similarity between the corresponding vectors $\tilde{V}_d$ and $\tilde{V}_k$:

$$\rho_{dk} = \frac{\tilde{V}_d}{\|\tilde{V}_d\|} \cdot \frac{\tilde{V}_k}{\|\tilde{V}_k\|}$$

In words, ρ_{dk} measures the proximity of d and k in the space of words W. To better capture the distance between the knowledge content of each document (rather than simply the list of words), we make a series of adjustments to this simple measure, which we describe below.

Accounting for term frequency and relevance Since our goal is to measure the knowledge content of each document, we assign more weight to terms that best capture this type of content relative to terms that are used frequently in the language (and, as such, might appear often in the document) but do not necessarily capture content. To this purpose, we use the "term-frequency-inverse-document-frequency (TFIDF)" transformation of word counts, a standard approach in the text analysis literature (Kelly et al., 2018). This approach consists in comparing the frequency of each term in the English language and in the body of all documents of a given type (e.g., syllabi or articles), assigning more weight to terms that appear more frequently in a given document than they do across all documents. For example, "genome editing" is used rarely in the English language, but often in some Biology syllabi syllabi; "assignment" is instead common across all syllabi. Because of this, "genome editing" is more informative of the content of a given syllabus and should therefore receive more weight than "assignment".

We construct the weight of a term w in document d as:

$$TFIDF_{wd} = TF_{wd} \times IDF_w$$

where $TF_{wd} \equiv \frac{c_{wd}}{\sum_k c_{kd}}$ is the frequency of word w in document d, c_{wd} counts the number of times term w appears in d, and

$$IDF_w \equiv \log\left(\frac{|D|}{\sum_d \mathbb{1}(w \in \tilde{V}_d)}\right)$$

is the inverse document frequency of term w in the set D of all documents of the same type as d. Intuitively, the weight will be higher the more frequently w is used in document d (high TF_{wd}), and the less frequently it is used across all documents (low IDF_d). In words, words that are more distinctive of the knowledge content of a given document will receive more weight.

To maximize our ability to capture the knowledge content of each document, in our analysis we focus exclusively on words related to knowledge concepts and skills, excluding words such as pronouns or adverbs. We do this by appropriately choosing our "dictionaries," lists of all relevant words (or sets of words) that are included in the document vectors. Our primary dictionary is the list of all unique terms ever used as keywords in academic publications from the beginning of our publication sample until 2019. As an alternative, we have also used the list of all terms that have an English Wikipedia webpage as of 2019; our results are robust to this choice.

Accounting for changes in term relevance over time The weighting approach described so far calculates the frequency of each term by pooling together documents published in different years. This is not ideal for our analysis, because the resulting measures of similarity between syllabi and publications would ignore the temporal ordering of these documents. Instead, we are interested in the novelty of the content of a syllabus d relative to research published in the years prior to d, without taking into account the content of future research. To see this consider, for example, course CS299 at Stanford University, taught by Andrew Ng in the early 2000 and one of the first entirely focused on *Machine Learning*. Pooling together documents from different years would result in a very low $TFIDF_{wd}$ for the term "machine learning" in the course's syllabus: Since the term has been used very widely in the last years, its frequency across all documents would be very high and its IDF very low. Not accounting for changes in the frequency of this term over time would then lead us to misleadingly underestimate the course's path-breaking content.

To overcome this issue, we modify the traditional $TFIDF$ approach and construct a retrospective or "point-in-time" version of IDF, meant to capture the inverse frequency of a word among all articles published *up to a given date*. We call this measure "backward-IDF," or $BIDF$, and define it

as

$$BIDF_{wt} \equiv \log\left(\frac{\sum_d \mathbb{1}(t(d) < t)}{\sum_d \mathbb{1}(t(d) < t) \times \mathbb{1}(w \in \tilde{V}_d)}\right)$$

where $t(d)$ is the publication year of document d. Unlike IDF, $BIDF$ varies over time to capture changes in the frequency of a term among documents of a given type. This allows us to give the term its temporally appropriate weight. Using the $BIDF$ we can now calculate a "backward" version of $TFIDF$, substituting $BIDF$ to IDF:

$$TFBIDF_{wd} = TF_{wd} \times BIDF_{wt(d)}$$

Building the weighted cosine similarity Having calculated weights $TFBIDF_{wd}$ for each term w and document d, we can obtain a weighted version of our initial vector $\tilde{V}_d$, denoted as V_d, multiplying each term $w \in \tilde{V}_d$ by $TFBIDF_{wd}$. We can then re-define the cosine similarity between two documents d and k, accounting for term relevance, as

$$\rho_{dk} = \frac{V_d}{\|V_d\|} \cdot \frac{V_k}{\|V_k\|}.$$

Since $TFBIDF_{wd}$ is non-negative, ρ_{dk} lies in the interval $[0, 1]$. If d and k are two documents of the same type that use the exact same set of terms with the same frequency, $\rho_{dk} = 1$; if instead they have no terms in common, $\rho_{dk} = 0$.

3.1 Calculating the Education-Innovation Gap

To construct the education-innovation gap, we proceed in 3 steps.

Step 1: We calculate ρ_{dk} between each syllabus d and article k.

Step 2: For each syllabus d, we define the average similarity of a syllabus with all the articles published in a given three-year time period τ:

$$S_d^\tau = \sum_{k \in \Omega_\tau(d)} \rho_{dk}$$

where ρ_{dk} is the cosine similarity between syllabus d and a article k (defined in equation (3)) and $\Omega_\tau(d)$ is the set of all articles published in the three-year time interval $[t(d) - \tau - 2, t(d) - \tau]$.[17]

Step 3: We construct the education-innovation gap as the ratio between the average similarity of a

[17] For our main analysis we use three-years intervals; our results are robust to the use of one-year or two-years intervals.

syllabus with older technologies (published in τ) and the similarity with more recent ones ($\tau' < \tau$):

$$Gap_d \equiv \left(\frac{S_d^{\tau}}{S_d^{\tau'}} \right) \tag{1}$$

It follows that a syllabus published in t has a lower education-innovation gap if its text is more similar to more recent research than older research. In our analysis, we set $\tau = 13$ and $\tau' = 1$, and we scale the measure by a factor of 100 for readability.

It is worth emphasizing the advantage of a ratio measure over a simple measure of similarity (S_d^1). In particular, the latter could be sensitive to idiosyncratic differences in the "style" of language across syllabi in different fields, or even within the same field. A ratio of similarity measures *for the same syllabus* is instead free of any time-invariant, syllabus-specific attributes.

3.2 Validation and Interpretation of Magnitude

To gauge the extent to which the education-innovation gap captures the "novelty" of a course's content, we perform a series of checks. First, we show that the relationship between the gap and the average age of its reference list (defined as the difference between the year of each syllabus and the publication year of each reference) is strong and almost linear, with a correlation of 0.99 (Figure 1).

Second, we show that more advanced and graduate courses have a lower gap compared with basic undergraduate courses. The latter have a gap of 97.7; more advanced undergraduate courses have a gap of 95.4, and graduate courses have a gap of 94.9 (Appendix Figure AV). This suggests that more advanced courses cover content that is closer to frontier research.

Lastly, we simulate the extent to which a unit change in the education-innovation gap translates into actual differences in the content of a course. We do so with a simple simulation exercise: We replace "old" words with "new" words in a randomly selected subsample of 100,000 syllabi. "Old" and "new" are words in the bottom and top 5 percent of the distribution of the frequency of a term among all publications in the same field in the year prior to the one of the syllabus. We then calculate the gap for each syllabus as we gradually replace more words.

This exercise shows that the gap monotonically decreases if we replace old words with new ones (Figure 2). In particular, a unit reduction in the gap requires replacing 14 old words with newer ones – an almost 10 percent change to the average syllabus, which contains 331 terms.

3.3 Decomposing The Education-Innovation Gap

As a stepping stone for our empirical analysis, we now describe how the education-innovation gap varies across fields, institutions, courses, and instructors, decomposing its variation among these factors.

Figure 3 (solid line) shows the distribution of the gap across all syllabi taught between 1998 and 2018. The average course has a gap of 95.2, with a standard deviation of 5.9, a 25th percentile of 86.6, and a 75th percentile of 98.2. To better quantify the extent of this variation, we make use of the relationship illustrated in Figure 2: In order to move a syllabus from the 75th to the 25th percentile one would have to replace approximately 22 percent of its content (or 35 words).

Figure 3 also shows how the dispersion in the gap decreases as we progressively control for institution, field, course, and instructor fixed effects; this is helpful to understand the contribution of these factors to the overall variation in the gap.[18] Controlling for institution reduces the standard deviation to 5.0; controlling for field reduces it to 4.7; controlling for courses reduces it to a much smaller 2.6, and controlling for instructors brings it to 1.9.

To more rigorously quantify the part of the variation in the gap explained by each of these factors, in Table 3 we estimate OLS regressions of the gap on various sets of fixed effects (column 1), and we report how much the R^2 of a baseline regression (including only year fixed effects) decreases as we add controls for each different factor. This exercise reveals that differences among institutions explain 3 percent of the variation in the gap; differences among fields explain an additional 13 percent, differences among courses explain an additional 57 percent, and differences among instructors explain 13 percent.[19]

The results from our decomposition exercise indicate a substantial amount of variation in the education-innovation gap of syllabi taught across fields, institutions, and by different instructors. In the next sections, we focus more in depth on two of these factors: institutions and instructors. Specifically, we study how the gap varies across different types of schools serving different populations of students, and we explore how it relates to the research productivity and focus of the person who teaches the course.

[18] We obtained the within-field, within-institution, and within-instructor distributions using the residuals from a regression of the gap on the corresponding field, institution, course, and instructors fixed effects. We then added the mean gap to each set of residuals.

[19] For example, the R^2 of a regression with year fixed effects equals 0.23, that of a regression with instructor and year fixed effects equals 0.26, and that of a regression of field, instructor, and year fixed effects equals 0.36. As a result, instructor fixed effects explain (0.26-0.23)/(1-0.23)= 3 percent of the extra variation once year fixed effects are accounted for, and field fixed effects explain an additional (0.36-0.26)/(1-0.23)=13 percent.

4 The Education-Innovation Gap Across Schools

To investigate how the education-innovation gap varies among different types of institutions, we focus on three school characteristics: selectivity, students' parental income, and the racial and ethnic makeup of the student population. This exercise is helpful to to assess whether access to up-to-date educational content differs among students with different backgrounds, enrolled at different types of schools.

4.1 Selectivity

We begin by studying how the gap differs across more and less selective schools. We divide schools into five "tiers" according to their level of control (public vs private) and the degree of selectivity in admissions; our classifications follows Chetty et al. (2019) and is based on Barron's 2009 selectivity ranking. Our tiers are as follows. "Ivy Plus" include Ivy League universities and the University of Chicago, Stanford, MIT, and Duke. "Elite" schools are all the other schools classified as Tier 1 in Barron's ranking. "Highly selective and selective public" and "Highly selective and selective private" correspond to schools in Barron's Tiers 2 to 5. Lastly, "Non-selective" schools include those in Barron's Tier 9 and all four-year institutions not included in Barron's classification.[20]

To compare the gap across different school tiers, we use the following equation:

$$\text{Gap}_i = \mathbf{S}_i'\boldsymbol{\beta} + \phi_{f(i)l(i)t(i)} + \varepsilon_i,$$

where Gap_i measures the education-innovation gap of syllabus i, taught in school $s(i)$ in year $t(i)$. The vector $\mathbf{S}_i'$ contains indicators for selectivity tiers. Field-by-level-by-year fixed effects ϕ_{flt} control for systematic, time-variant differences in the gap that are common to all syllabi in the same field and course level. We cluster standard errors at the institution level.

Point estimates of the coefficients vector $\boldsymbol{\beta}$ in equation (2), shown in panel (a) of Figure 4, represent conditional mean gaps for schools in each tier. These estimates indicate that the gap is significantly lower for more selective schools, and it progressively increases as selectivity declines. Ivy Plus schools have the smallest gap, at 94.7, followed by Elite schools at 95.1. The gap increases up to to 95.6 for non-selective schools.

In order to quantify the economic magnitude of the difference in gaps between more and less selective schools, we make use of the simulation results illustrated in Figure 2. The simulation

[20] For comparability, we exclude two-year institutions.

indicates that in order to close the difference in the gap between Ivy-Plus and non-selective schools, one would have to replace approximately 4 percent of the knowledge content in syllabi of non-selective schools (13 terms).

In Table 4 (panel (a)) we re-estimate equation (2) for syllabi in different fields and course levels, using non-selective schools as the reference tier. Column 1, estimated on the full sample, confirms the findings of Figure 4: Ivy-Plus, Elite, Selective public, and Selective private schools have a 0.89, 0.53, 0.34, and 0.29 lower gap compared with Non-selective schools.

Columns 2-5 show estimates by macro-fields. The differences in the gaps across selectivity tiers are most pronounced for Business, STEM, and Social Science; they are instead indistinguishable from zero for Humanities. For example, syllabi of Business and Social Sciences courses have a 1.3 lower gap in Ivy-Plus schools compared with non-selective schools (or 18 newer knowledge words, significant at 1 percent). Similarly, syllabi of STEM courses have a 1.06 lower gap in Ivy-Plus schools.

Columns 6-8 show estimates by the level of the course (undergraduate basic, advanced, and graduate). Differences in the gap across selectivity tiers are more pronounced for undergraduate courses, especially the basic ones. For example, basic undergraduate courses taught in Ivy Plus schools have a 0.98 lower gap (or 14 additional new knowledge words) compared with those taught in non-selective schools (significant at 1 percent). This difference is equal to 0.81 for advanced undergraduate courses (8 knowledge terms, significant at 1 percent), and only 0.55 (indistinguishable from zero) for graduate courses.

4.2 Parental Income

What can explain the difference in gaps across more and less selective schools? If students are allocated to schools based on their ability and lower-gap courses are better suited for higher-ability students, this finding could be driven by schools adjusting the content of each course to their students' ability. However, students are not allocated to schools uniquely on ability; for example, Ivy-Plus and Elite schools are disproportionately more likely to enroll students from wealthier backgrounds (Chetty et al., 2019). A consequence of this is that access to up-to-date content might be unequally distributed across more and less advantaged students, even conditional on their ability.

To more directly test for his hypothesis, we now investigate how the gap differs across schools serving students from different socio-economic backgrounds, overall and conditional on the school's selectivity. We measure students' backgrounds with two use two school-level measures: Median

parental income and the share of parents with incomes in the top percentile of the national distribution, constructed using tax returns for the years 1996 to 2004 (Chetty et al., 2019).

Median Parental Income The data indicate that the gap is unequally distributed across schools serving more and less wealthy students. The education-innovation gap is negatively related to the median parental income of students at each school: a $10,000 higher median income is associated with a 0.15 lower gap (Figure 5, panel (a)).

To explore non-linearities in this relationship, we also -reestimate a specification similar to equation (2), where the vector $\mathbf{S}_i'$ contains indicators for schools with parental income in the bottom 25 percent, 25-50 percent, 50-75 percent, 75-99 percent, and top 1 percent of distribution of across all schools. Estimates of this specification, shown in the darker series in panel (b) of Figure 4, confirm a negative relationship between the gap and median parental income. Schools with parental income in the bottom 25 percent have a gap equal to 95.6, schools in the middle of the distribution (25 to 99 percentile) have a gap between 95.2 and 95.3, and schools with median parental income in the top percentile of the distribution have a significantly smaller gap, equal to 94.5.

Of course, these patterns might be driven by differences in ability among more and less wealthy students. To control for these differences, we obtain these estimates further controlling for a school's selectivity tier. These estimates, shown in the lighter series in panel (b) of Figure 4, indicate that the negative relationship between the gap and median parental income at each school is present even within selectivity tiers. Schools in the same tier with parental income in the bottom 25 percent have a gap equal to 95.2, schools in the middle of the distribution (25 to 99 percentile) have a gap between 94.9 and 95.0, and schools with median parental income in the top percentile of the distribution have a significantly smaller gap, equal to 94.5. These estimates imply that, in order to close the difference in the gap between schools with median parental income in the bottom quartile and those with income in the top one percent, one would have to replace approximately 6 percent of the total knowledge content of the average syllabus, or 20 knowledge terms.

In panel (b) of Table 4 we re-estimate these specifications for different subgroups of syllabi, using schools in the bottom quarter of the median parental income as the reference group. The difference in the gap between schools with parental income in the top 1 percent and those in the bottom quartile is significant across all macro-fields, but most pronounced for Business (1.6, column 2) and least pronounced for Humanities (0.77, column 3). The difference is present across all levels but largest for graduate courses (1.2, column 8).

Share of Parents in the Top Income Percentile We repeat our analysis using the share of parents with incomes in the top percentile in each school as a measure for students' background. Panel (b) of Figure 5 shows the relationship between each school's share of students with parental income in the top percentile and the education-innovation gap. The two variables are negatively correlated, with a slope coefficient of -7.9 (significant at 1 percent). This correlation implies that a ten-percent increase in the share of students with parental income in the top percentile is associated with a 0.8 lower gap, equivalent to a 6 percent difference in the syllabus syllabus content, or 20 newer knowledge terms.

As before, we further investigate this relationship by dividing schools into bins depending on the share of students with parental income in the top percentile. These estimates confirm that the gap is smallest for schools enrolling more students with parental incomes in the top percentile. In particular, the gap is equal to 94.7 for schools where more than 15 percent of students are in the top percentile, whereas it is much larger at 95.6 for schools where less than 0.1 percent of students have parental incomes at the very top of the distribution (Figure 4, panel (c), darker series). Estimates are robust to controls for the selectivity of each school (lighter series). These results also imply that, in order to close the gap between schools with almost no students and those with 15 percent or more students with parental incomes in the top percentile, one would have to replace approximately 13 knowledge terms to the average syllabus, or 4 percent of its content.

This pattern is confirmed by the estimates in Table 4 panel (c) column 1, where the reference group are schools with less than 0.1 percent of students with parental income in the top percentile. Differences appear most pronounced for Business courses (column 2) and basic undergraduate courses (column 6).

4.3 Students' Race and Ethnicity

Lastly, we investigate whether schools enrolling more Black or Hispanic students (which we refer to as "minority") offer courses with significantly different gaps. The relationship between the share of minority students in each school and the average education-innovation gap is equal to 2.8 and significant at 1 percent (Figure 5, panel (c). This indicates that a ten-percent increase in the share of minority students in each school is associated with a 28 percent difference in the content of the average syllabus, or 45 older knowledge terms.

To more transparently explore how access to university courses with smaller gaps varies across students of different races and ethnicities, we divide schools in five bins depending on their share

of minority students. We then estimate a specification similar to equation (2), where the vector $\mathbf{S}_i'$ contains indicators for each bin as independent variables. This exercise confirms that schools with more than 40 percent of minority students students have a larger gap, equal to 95.3. By comparison, schools with a share of minority students lower than 5 percent have a gap of 95.0 (Figure 4, panel (d), darker series). These estimates imply that, in order to close the difference in the gap between schools with more than 40 percent and those with less than 5 percent of students who are minority, one would have to replace 6 knowledge words in the average syllabus, or 3.7 percent. These estimates are robust to controls for school selectivity (Figure 4, panel (d), lighter series).

These patterns are confirmed by the estimates in Table 4 panel (d), where we use schools with more than 40 percent minority students as the reference group. Compared with the reference group, schools with less than 5 percent of students who are minority have a 0.29 larger gap. Differently from before, these differences are larger for courses in STEM (-0.29, column 4), Humanities (-0.27, column 3), and Social Sciences (-0.26, column 5), and insignificant for Business (column 2). They are also larger for basic and graduate courses (-0.36 and -0.28, columns 6 and 8).

5 The Role of Instructors

The results presented so far indicate a significant amount of variation in the education-innovation gap across schools serving different populations of students. Nevertheless, the decomposition exercise in Table 3 and Figure 3 suggests that differences are also present within schools, across courses taught by different people. We now explore this variation, focusing on the role of instructors in explaining it.

5.1 The Education-Innovation Gap When The Instructor Changes

To better quantify the importance of instructors in shaping the education-innovation gap of the courses they teach, we follow the literature on the effects of teachers on student achievement (Rivkin et al., 2005; Chetty et al., 2014) and characterize how the gap of a course varies when there is a change of instructor. To so do, we estimate an event study of the gap in a 10-years window around the time of the instructor change:

$$\text{Gap}_i = \sum_{k=-4}^{5} \delta_k \mathbb{1}(t(i) - T_{c(i)} = k) + \phi_{c(i)} + \phi_{s(i)f(i)} + \phi_{t(i)} + \varepsilon_i, \tag{2}$$

where i, c, s, f, and t denote a syllabus, course, school, field, and year respectively, and the variable T_c represents the first year in our sample in which the instructor of course c changed.[21] We restrict our attention to courses taught by only one person in each year and we set $t(i) - T_c = 0$ for all courses without an instructor change, which in practice serve as a comparison group. We cluster our standard errors at the course level. In this equation, the parameters δ_k capture the differences between the gap k years after an instructor change relative to the year preceding the change.

OLS estimates of δ_k, shown in Figure 6, indicate that a change in a course's instructor is associated with a progressive decline in the education-innovation gap. Estimates are indistinguishable from zero and on a flat trend in the years leading to an instructor change; five years after the change, the gap is 0.3 lower. This decline is equivalent to replacing 2 percent of the content of a syllabus, or 6 knowledge words.

In Table 5 (panel (a)) we re-estimate equation (2) for different subsamples of syllabi, pooling together years preceding and following an instructor change. After such a change, the gap declines for all fields and course levels. The decline is largest for Business courses (column 2) and smallest for STEM (column 4), and it is comparable across course levels (columns 6-8).

These results confirms that instructors play a crucial role in shaping the content of the courses they teach. They also suggest that instructors who take over an existing course from someone else significantly update the content of the syllabus, bringing it closer to the knowledge frontier.

5.2 The Education-Innovation Gap and Instructors' Research Productivity

Having confirmed the importance of the instructor in determining the education-innovation gap of the courses they teach, we now study how characteristics of the instructors relate to the gap. We focus on research productivity, measured using individual counts of citations and publications either in the five years preceding the syllabus or per year, since the instructor became research-active. We obtain these measures using data from Microsoft Academic.

In our data, the median instructor does not publish any article nor receive any citations. The top echelons of the distribution of citations and publications vary across macro-fields; for citations per year, the 75th percentile of the distribution ranges from a minimum of 0 for Humanities to a maximum of 11.4 for STEM, whereas for publications per year it ranges from 0 for Humanities to 1 for STEM.

Panels (a) and (b) of Figure 7 show a binned scatterplot of the gap and either citations per

[21] Our results are robust to using the median or last year of the instructor change.

year ((panel (a)) or publications per year (panel (b)), controlling for field effects.[22] The relationship between the gap and instructors' productivity is significantly negative for both measures of productivity.

This negative relationship is confirmed in Table 5 (panels (b) and (c), column 1), which shows estimates of the slope of the relationship between the gap and instructors' research productivity (expressed as the inverse hyperbolic sine of counts of publications and citations, to account for the skewed distribution of these variables). In these specifications we control for course and school-by-year fixed effects, to account for unobserved determinants of the gap that are specific to a course in a given school and year. Estimates on the full sample of syllabi indicate that a ten-percent increase in the instructor's number of publications per year is associated with a .04 lower gap (panel (b), column 1, significant at one percent).[23] Similarly, a ten-percent increase in the instructor's number of citations per year is associated with a 0.008 lower gap (panel (c), column 1, significant at 5 percent). Consistent with the previous finding of a larger decline in the gap following instructor changes in STEM, these relationships are stronger for STEM courses (column 4) and for courses at the basic and graduate level (columns 6 and 8).

The findings presented above indicate that instructors who produce more and better cited research teach courses with a lower gap. This could be due to the fact that more research-active instructors are better updated about the frontier of research, and more likely to cover this type of content in their courses. It is also possible that research-active instructors chose to teach their own work. If the relationship between research productivity and the gap is driven by instructors being more informed about the research frontier, we should expect this relationship to be stronger, the closer the course is to the instructor's own research interests. We test this possibility next.

The role of "fit" between the instructor and the course To measure the closeness between each instructor's research expertise and the content of the course they teach, or "fit," we calculate the cosine similarity between the text of instructor's publications in the previous five years and the description of the course contained in the syllabus, restricting attention to instructors with at least one publication.

Panel (c) of Figure 7 shows a binned scatterplot of the gap against our measure of fit: The relationship between these two variables is negative and significant. This is confirmed by the estimates

[22] In this figure, the horizontal axis corresponds to centiles of each productivity measures; the vertical axis shows the average gap in each centile.

[23] Bellemare and Wichman (2020) show that in a specification such as $y = \beta x + \varepsilon$, where x is expressed as the inverse hyperbolic sine of a variable $\tilde{x}$, the elasticity of y to $\tilde{x}$ is β/y. In our case, the average y is 95.

in column 1 of Table 5, panel (d), which represent the slope of this relationship controlling for course and school-by-year fixed effects. The association between instructor fit and the education-innovation gap appears largest for courses in Humanities and Social Sciences and for basic undergraduate courses.

Taken together, our findings outline an important role for instructors in shaping the content of the course they teach. Research-active instructors are particularly likely to cover frontier knowledge in their courses, especially if the course is in an area close to their research expertise. This suggests that a well-thought assignment of instructors to courses can be a valuable tool to ensure students are exposed to up-to-date content, which–as we show next–is associated with better educational and economic outcomes.

6 The Education-Innovation Gap and Students' Outcomes

The results presented until here show significant differences in access to up-to-date knowledge across schools serving different types of students and across courses within the same school. We now explore whether these differences are related to a host of students' outcomes. We focus on three outcomes: graduation rates, income, and intergenerational mobility. Graduation rates and income figures are from IPEDS and the College Scorecard, respectively, and they cover years from 1997 to 2018. To measure intergenerational mobility we borrow a measure from Chetty et al. (2019), defined as the probability that students with parental incomes in the bottom quintile of the distribution who graduated between 2002 and 2004 reach the top quintile during adulthood. We first study whether each of these outcomes is related to the education-innovation gap. Then, we test whether this relationship is different for different groups of schools and students.

Since our outcomes are measured at the school level, this exercise requires the construction of a school-level gap. We follow the school value-added literature (see Deming, 2014, for example) and estimate the school component of the gap using the following model:

$$\text{Gap}_i = \theta_{s(i)} + \phi_{f(i)l(i)t(i)} + \varepsilon_i. \tag{3}$$

In this equation, the quantity θ_s captures the average education-innovation gap of school s, accounting flexible time trends that are specific to the level l and the field f of the course. Appendix Figure AVI shows the distribution of θ_s; the standard deviation is 0.85.

6.1 Graduation Rates

We begin by studying how the gap relates to graduation rates. Column 1 of Table 6 shows estimates of the slope coefficient of the gap on graduation rates, measured at the school level for the years 1998-99 to 2016-17, controlling for year fixed effects. An estimate of -0.06, significant at 1 percent, indicates that a unit decline in the gap (equivalent to 25 extra words, or a 15 percent change in the content of the median syllabus) is associated with a 6 percentage points increase in graduation rates. Compared with an average graduation rate of 56 percent, this implies a 11 percent increase.

A negative relationship between the gap and graduation rates might be driven by idiosyncratic differences between schools and the students they enroll, related to both the gap and graduation rates. In an attempt to control for these differences, in column 2 of Table 6 we test whether the relationship holds when we control for a range of school-level observables: selectivity tier, median parental income, the share of parents in the top 20 percent of the income distribution, and the share of students who are minority. Even controlling for these observables, the relationship between the gap and graduation rates remains significant. A coefficient of -0.0134 indicates that a unit decline in the gap is associated with a 1.3 percentage points increase in graduation rates, or 2.3 percent.

Estimates by Selectivity Tier Estimates for the coefficient *Gap* in Table 6 represent averages across all schools. The relationship between the gap and graduation rates, however, could be different depending on the type of schools and the population of students they serve.

We test this hypothesis in Table 7. In column 1, we allow for the slope coefficient of the gap to vary by selectivity tier, controlling for the same covariates as before (selectivity tier, median parental income, share of parents in top income quintile, and share of students who are minority). Our estimates indicate that the relationship between the gap and graduation rates is indistinguishable from zero for more selective schools and negative for less selective schools. In particular, estimates of the coefficient *Gap * Ivy Plus/Elite* are positive, although insignificant. On the other hand, estimates for *Gap * (Highly selective) private* are equal to -0.013, indicating that a unit decline in the gap is associated with a 1.3 percentage points increase in graduation rates, or 2.3 percent. Estimates of *Gap * (Highly selective) public* are even large at -0.016, and estimates of *Gap * Non-selective* are largest at -0.022.

Estimates by Parental Income One possible explanation for the stronger relationship between the gap and graduation rates for less selective institutions is that more up-to-date content is particularly

valuable for students of lower ability. However, less selective institutions are also more likely to enroll economically worse-off students. Estimates by school selectivity could therefore also capture a heterogeneous relationship by students' economic background, rather than ability. We test for this possibility in column 2 of Table 7, where we allow the slope coefficient of the gap to vary by quartiles of median parental income at each institution and, importantly, we control for selectivity. While imprecise, our estimates confirm the second hypothesis and suggest that, even within tiers, the relationship is stronger for schools with median parental income outside of the top quartile (with estimates equal to -0.012, -0.014, and -0.016 for the first, second, and third quartile respectively, and an estimate of -0.0033 for the fourth and highest quartile).

6.2 Students' Incomes

Graduation rates are a strictly academic measure of student success; however, they are also likely to affect students' long-run economic trajectories. To directly examine the relationship between the education-innovation gap and students' economic success after they leave college, we study whether the gap is related to the income of students who graduate from each school ten years after graduation measured in the years 1998-99 to 2016-17. An estimate of -0.06, significant at 1 percent, indicates that a unit decline in the gap is associated with a 11 percent increase in incomes ten years after graduation (Table 6, column 3). This relationship becomes weaker, but remains significant, once we account for differences across schools in selectivity, median parental income, the share of parents in the top income quintile, and the share of students who are minority (column 4).

In columns 3 and 4 of Table 7 we allow the relationship between the education-innovation gap and student incomes to vary by school selectivity and median parental income. As with graduation rates, we find that the relationship is stronger for less-selective schools and essentially flat for Ivy-Plus and Elite and selective private schools. A unit decline in the gap is associated with a 1.9 percentage points (3.4 percent) higher graduation rate in selective public schools and a 2.3 percentage points (4 percent) higher rate in non-selective schools (column 3). Within selectivity tiers, the relationship is stronger in schools with median parental quintile between the 25th and 50th percentile (Q2, column 4).

6.3 Gap and Intergenerational Income Mobility

The last outcome we examine is intergenerational mobility. Our data indicate that a unit decline in the gap is associated with a 3.1 percentage points increase in this probability (Table 6, column 3).

This relationship, however, becomes essentially zero when we control for a school's observables.

6.4 Summary

Our analyses of student outcomes indicate that a lower education-innovation gap at the school level is associated with improved academic and economic outcomes of the students at each school, including graduation rates and incomes after graduation, and especially for less selective schools and those that serve a lower proportion of wealthy students. The lack of experimental variation in the gap across schools prevents us from pinning down a causal relationship with certainty. Nevertheless, our results are robust to controls for the ability and economic background of a school's student and, at a minimum, suggestive of a positive effect of up-to-date course content on students' outcomes.

7 Soft Skills

Our analysis of the education-innovation gap focuses on the novelty of a syllabus with respect to its academic content. We now take a broader perspective and explore another dimension of novelty, not necessarily captured by purely academic content: soft skills, defined as non-cognitive abilities that define how a person interacts with their colleagues and peers, and identified by recent literature as increasingly demanded in the labor market (Deming, 2017).

To assess the soft-skills intensity of a syllabus, we focus on the course's evaluation scheme. Specifically, we consider a course to be more soft-skills intensive if the assignments portion of the syllabus has a higher share of words such as "group", "team", "presentation", "essay", "proposal", "report", "drafting", and "survey". In the average syllabus, 33 percent of the words in the assignment portion of the syllabus refers to soft skills (Table 1, panel (a)).

A binned scatterplot of soft-skills intensity and the education-innovation gap of a syllabus reveals that the two measures are negatively correlated, with a correlation of -0.16 (Figure 8, panel (a)). Using the same empirical framework used above, we find that the cross-school differences in the skill intensity of the courses display the same patterns found above for the education-innovation gap. The prevalence of soft skills increases with school selectivity (panel (b)) and it is larger for schools where the median parental income is in the top portion of the distribution.

8 Conclusion

This paper examines the production of human capital by investigating the knowledge content of higher education. Our approach centers around a new measure, the "education-innovation gap," defined as the textual distance between syllabi of courses taught in colleges and universities and the frontier knowledge published in academic journals. We measure this gap with a novel measure based on textual analysis techniques, using information on the text of 3 million university syllabi taught in nearly three decades and 20 million academic publications.

This new approach allows to document a set of new findings about the offering of frontier knowledge across US higher education institutions. First, a significant amount of variation in frontier knowledge exists across university courses, both across and within institutions, the largest part of which is explained by instructors. Second, more selective schools, schools serving students from wealthier backgrounds, and schools serving a smaller proportion of minority students offer courses with a smaller gap. Third, instructors play a large role in shaping the content they teach, and more research-active instructors are more likely to teach courses with a lower gap. Fourth, the gap is correlated with students' outcomes such as graduation rates, income ten years after graduation, and intergenerational mobility, and the correlation is particularly pronounced for schools serving more disadvantaged students. Taken together, our results suggest that that the education-innovation gap can be an important indicator to study how human capital is produced in higher education.

References

Akcigit, Ufuk, Santiago Caicedo, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi, 2018, Dancing with the stars: Innovation through interactions, Technical report, National Bureau of Economic Research.

Altonji, Joseph G, Erica Blom, and Costas Meghir, 2012, Heterogeneity in human capital investments: High school curriculum, college major, and careers, *Annu. Rev. Econ.* 4, 185–223.

Andrews, Michael, 2017, The role of universities in local invention: evidence from the establishment of us colleges, *Job Market Paper* .

Angrist, Joshua D, and Jörn-Steffen Pischke, 2017, Undergraduate econometrics instruction: through our classes, darkly, *Journal of Economic Perspectives* 31, 125–44.

Baumol, William J, 2005, Education for innovation: Entrepreneurial breakthroughs versus corporate incremental improvements, *Innovation policy and the economy* 5, 33–56.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, 2019, Who becomes an inventor in america? the importance of exposure to innovation, *The Quarterly Journal of Economics* 134, 647–713.

Bellemare, Marc F, and Casey J Wichman, 2020, Elasticities and the inverse hyperbolic sine transformation, *Oxford Bulletin of Economics and Statistics* 82, 50–61.

Benhabib, Jess, and Mark M Spiegel, 2005, Human capital and technology diffusion, *Handbook of economic growth* 1, 935–966.

Bianchi, Nicola, and Michela Giorcelli, 2019, Scientific education and innovation: from technical diplomas to university stem degrees, *Journal of the European Economic Association* .

Biasi, Barbara, David J Deming, and Petra Moser, 2020, Education and innovation, in *The Role of Innovation and Entrepreneurship in Economic Growth* (University of Chicago Press).

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, 2020, Are ideas getting harder to find?, *American Economic Review* 110, 1104–44.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, 2014, Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates, *American Economic Review* 104, 2593–2632.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan, 2019, Income segregation and intergenerational mobility across colleges in the united states, *NBER Working Paper* .

Dale, Stacy, and Alan B Krueger, 2011, Estimating the return to college selectivity over the career using administrative earnings data, *NBER Working Paper* .

Deming, David, and Lisa B Kahn, 2018, Skill requirements across firms and labor markets: Evidence from job postings for professionals, *Journal of Labor Economics* 36, S337–S369.

Deming, David J, 2014, Using school choice lotteries to test measures of school effectiveness, *American Economic Review* 104, 406–11.

Deming, David J, 2017, The growing importance of social skills in the labor market, *Quarterly Journal of Economics* 132, 1593–1640.

Gentzkow, Matthew, and Jesse M Shapiro, 2010, What drives media slant? evidence from us daily newspapers, *Econometrica* 78, 35–71.

Hanushek, Eric A, and Ludger Woessmann, 2012, Do better schools lead to more growth? cognitive skills, economic outcomes, and causation, *Journal of economic growth* 17, 267–321.

Hoxby, Caroline M, 1998, The return to attending a more selective college: 1960 to the present, *Unpublished manuscript, Department of Economics, Harvard University, Cambridge, MA* .

Jones, Benjamin F, 2009, The burden of knowledge and the death of the renaissance man: is innovation getting harder?, *Review of Economic Studies* 76, 283–317.

Jones, Benjamin F, 2010, Age and great invention, *The Review of Economics and Statistics* 92, 1–14.

Jones, Benjamin F, and Bruce A Weinberg, 2011, Age dynamics in scientific creativity, *Proceedings of the National Academy of Sciences* 108, 18910–18914.

Kantor, Shawn, and Alexander Whalley, 2019, Research proximity and productivity: long-term evidence from agriculture, *Journal of Political Economy* 127, 819–854.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, 2018, Measuring technological innovation over the long run, *NBER Working Paper* .

Lucas Jr, Robert E, 2015, Human capital and growth, *American Economic Review* 105, 85–88.

Lucas Jr, Robert E, and Benjamin Moll, 2014, Knowledge growth and the allocation of time, *Journal of Political Economy* 122, 1–51.

Nelson, Richard R, and Edmund S Phelps, 1966, Investment in humans, technological diffusion, and economic growth, *American Economic Review* 56, 69–75.

Rivkin, Steven G, Eric A Hanushek, and John F Kain, 2005, Teachers, schools, and academic achievement, *Econometrica* 73, 417–458.

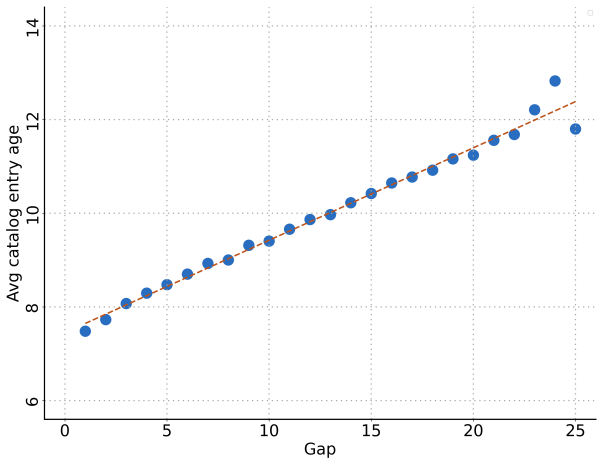
Rockoff, Jonah E, 2004, The impact of individual teachers on student achievement: Evidence from panel data, *American Economic Review* 94, 247–252.

Romer, Paul M, 1990, Endogenous technological change, *Journal of Political Economy* 98, S71–S102.

Toivanen, Otto, and Lotta Väänänen, 2016, Education and invention, *Review of Economics and Statistics* 98, 382–396.

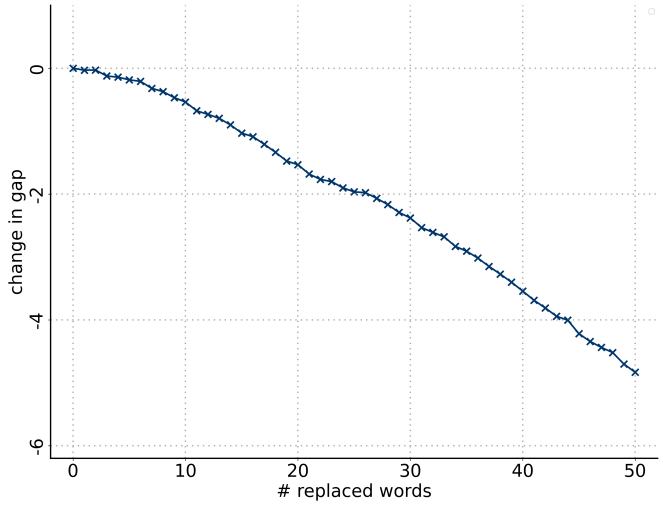
Valero, Anna, and John Van Reenen, 2019, The economic impact of universities: Evidence from across the globe, *Economics of Education Review* 68, 53–67.

Figure 1: Validating the Education-Innovation Gap Measure With Syllabi References



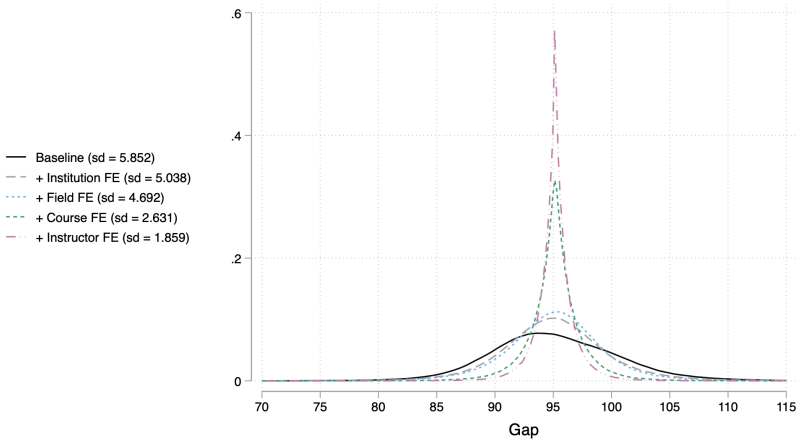
Note: This figure shows the correlation between the gap and the reference age of each syllabus. The reference age is defined as the average difference between the year of the syllabus and the year of each reference listed in the syllabus as a required or recommended reading. We divide syllabi in 25 equally-sized bins ranked by gap; the vertical axis correspond to the average reference age of each bin.

Figure 2: Simulation Experiment: Replacing Syllabus Content With Newer Words



Note: This figure links the differences in education-innovation gap with the associated number of "knowledge words" that must be replaced with newer words in each syllabus. We obtain this relationship by a) randomly choosing 100,000 syllabi from the sample, b) replacing a varying number of "old" knowledge words with "new" knowledge words, where "old" and "new" are defined with respect to the popularity of these terms among all publications in the same field and in the year prior to that of the syllabus, and c) measuring the change in the gap.

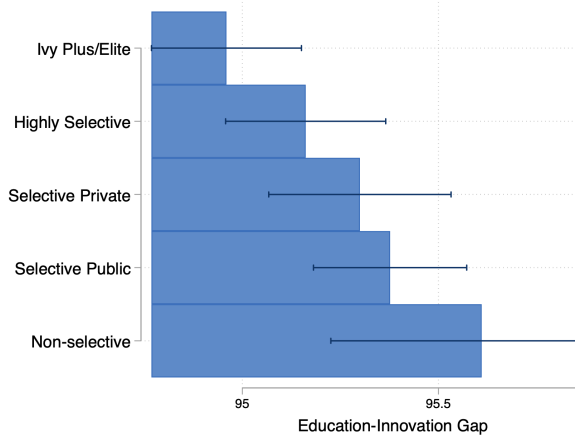
Figure 3: Education-Innovation Gap: Variation



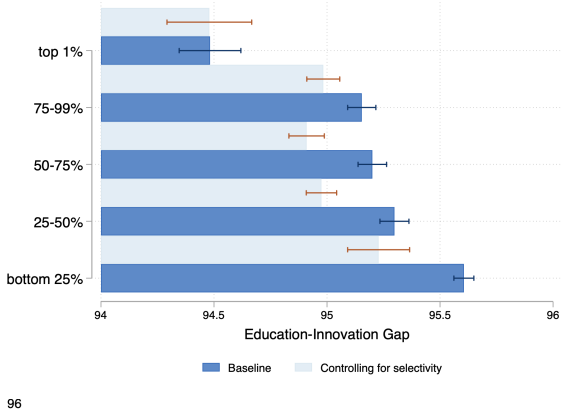
Notes: The figure shows the distribution of the gap. The solid line shows the raw data; the other series show the residuals of regressions as we progressively control for additional sets of fixed effects.

Figure 4: The Education-Innovation Gap By School Characteristics

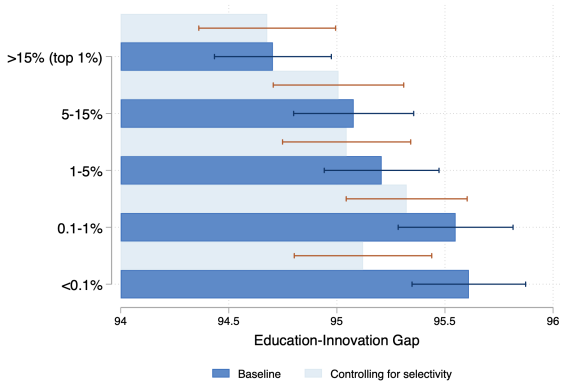
(a) By school tier



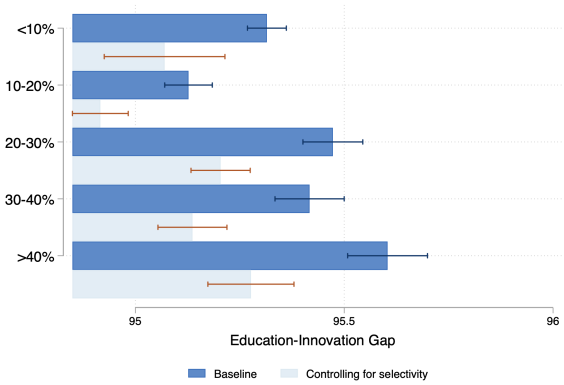
(b) By percentile of parental income



(c) By % of parents in top income percentile



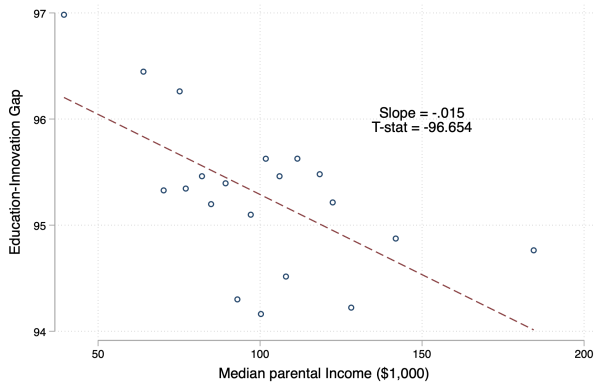
(d) By % of minority students (Black/Hispanic)



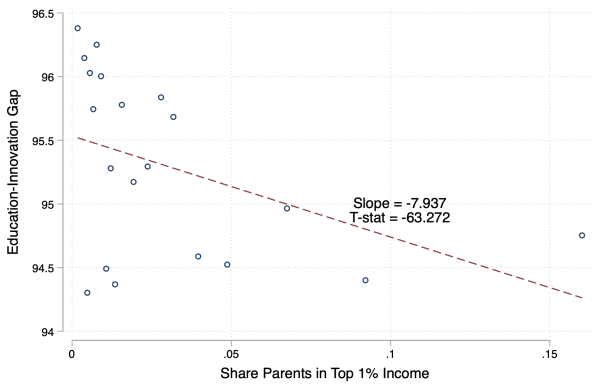
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by school tier (panel a), percentile of median parental income in the school (panel (b)), share students with parents in the top income percentile in the school (panel (c)), and share of students who are either Black or Hispanic (panel (d)). The gap is defined as the ratio between syllabi's cosine similarity with publications 13 to 15 years prior to the syllabus date and the cosine similarity with publications one to three years prior. Parental income percentiles for panel (b) are calculated using the distribution of median parental incomes across all schools. Percentiles for panel (c) are based on the national income distribution. Estimates are obtained pooling data for the years 1998 to 2018, and controlling for field and syllabus year fixed effects. In panels (b) to (d), the lighter series also control for selectivity tiers. Standard errors are clustered at the school level.

Figure 5: The Education-Innovation Gap And Student Demographics

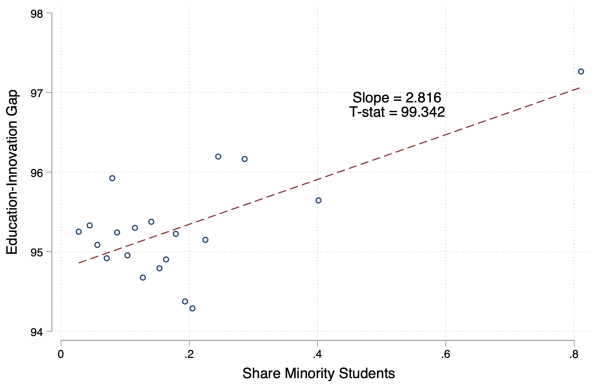
(a) Median parental income



(b) % of parents in top income percentile

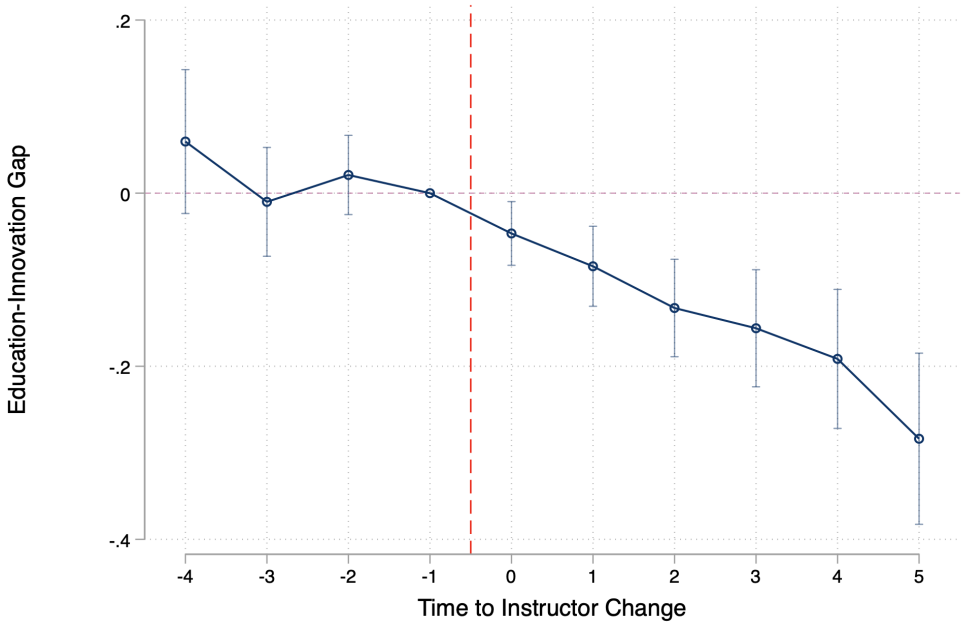


(c) % of minority students (Black/Hispanic)



Notes: Binned scatterplots of the education-innovation gap (vertical axis) and median parental income at each school (panel (a)), the share of parents with income in the top percentile (panel (b)), and the share of students who are Black or Hispanic ("minority", panel (c)).

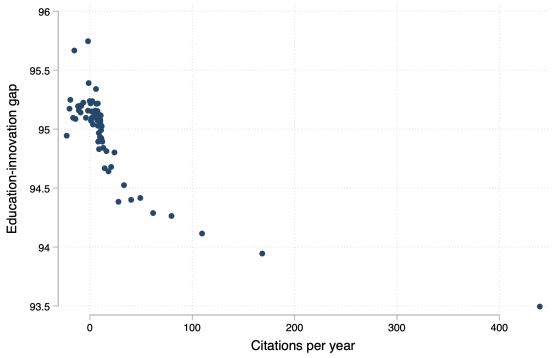
Figure 6: Education-Innovation Gap Around The Time of An Instructor Change



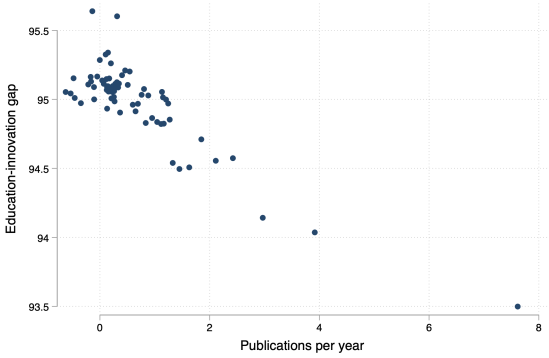
Notes: Estimates and standard deviations of the parameters δ_k, corresponding to an event study of the gap around an instructor change and specified in equation (2). Standard errors clustered at the course level.

Figure 7: Instructors' Research Productivity and The Education-Innovation Gap

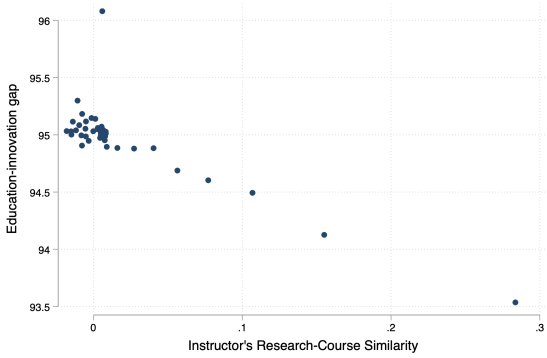
(a) Citations Per Year



(b) Publications Per Year



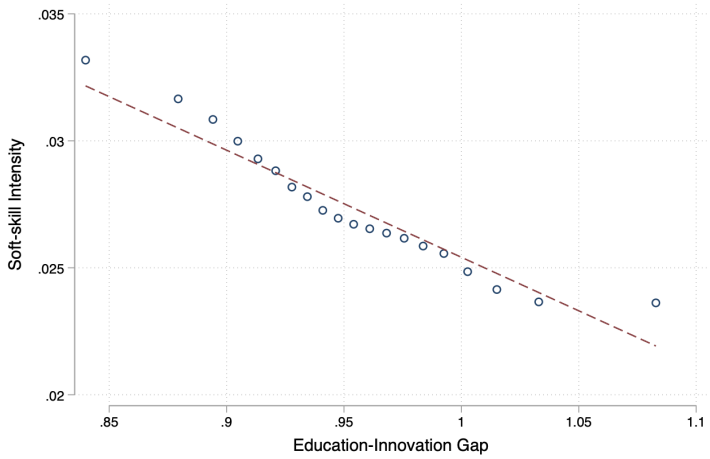
(c) Instructor's Fit with the Course



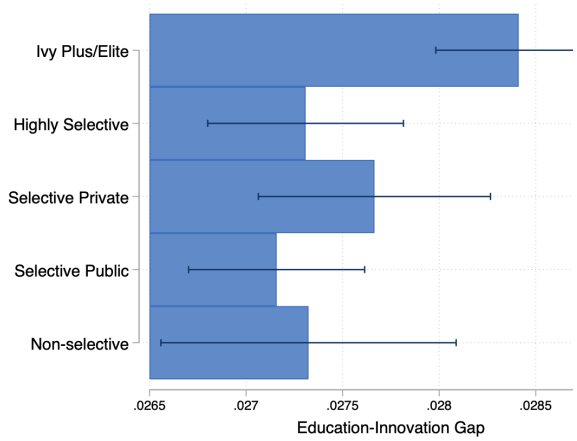
Notes: Binned scatterplot of the gap (vertical axis) and measures of research productivity and fit (horizontal axis): number of citations per year (panel (a)), number of article publications per year (panel (b)), and cosine similarity of each syllabus with an instructor's publications (panel (c)). Feld fixed effects are absorbed.

Figure 8: Soft Skills: Relationship With the Education-Innovation Gap and Variation Across Schools Training By School Characteristics

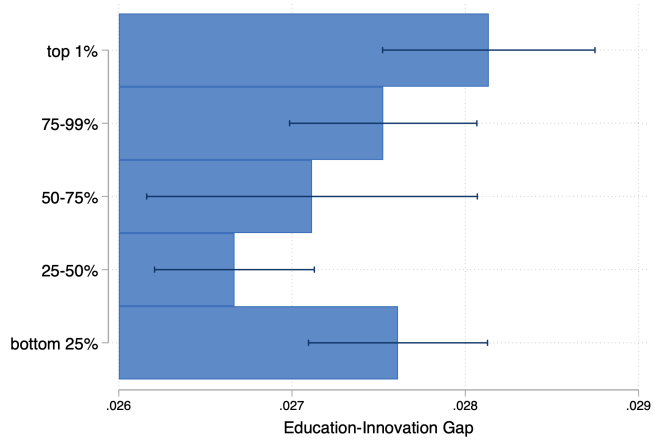
(a) Correlation w/Education-Innovation Gap



(b) By school tier



(c) By percentile of parental income



Notes: Panel (a) shows a binned scatterplot of the gap with a measure of soft skills, defined as the share of words that refer to soft skills in the assignment portion of each syllabus, and the education-innovation gap. Panels (b) and (c) show averages of the measure of soft skills by selectivity tier (panel (b)) and by percentile of the parental income distribution (panel (c)), controlling for field-by-year fixed effects. Standard errors in panels (b) and (c) are clustered at the school level.

Table 1: Summary Statistics: Courses, Instructors, and Schools

Panel (a): Syllabus (Course) Characteristics

	count	mean	std	25%	50%	75%
# Words	1,752,795	2,218.082	1,978.138	1,065	1,779	2,787
# Knowledge words	1,752,795	1,010.821	1,105.030	350	660	1,239
# Unique knowledge word	1,752,795	419.561	324.818	204	331	534
Soft skills	1,750,212	33.400	22.914	14.163	30.588	50
STEM	1,570,275	0.318	0.466	0	0	1
Business	1,570,275	0.114	0.318	0	0	0
Humanities	1,570,275	0.305	0.461	0	0	1
Social science	1,570,275	0.263	0.440	0	0	1
Basic	1,752,795	0.360	0.480	0	0	1
Advanced	1,752,795	0.289	0.453	0	0	1
Graduate	1,752,795	0.351	0.477	0	0	1

Panel (b): Instructor (Professor) Research Productivity

	count	mean	std	25%	50%	75%
# publications (all papers) per year	682,286	1.455	1.696	1	1	1.286
# publications (all papers), last 5 years	682,286	4.575	12.238	0	0	3
# citations (all papers) per year	682,286	24.146	82.553	0	1.333	15.286
# citations (all papers), last 5 years	682,286	103.961	621.203	0	0	17

Panel (c): Students' Characteristics and Outcomes at University Level

	count	mean	std	25%	50%	75%
Median parental income	778	97.544	30.945	77.500	93.200	109.575
Share parents in top income quintile	778	0.029	0.041	0.005	0.013	0.031
Share minority students	709	0.146	0.152	0.054	0.100	0.179
Graduation rates (2012–13 cohort)	785	0.608	0.191	0.467	0.604	0.758
Income (2003–04 and 2004–05 cohorts)	785	47,348.535	11,489.585	39,700	45,300	52,500
Intergenerational mobility	778	0.291	0.139	0.179	0.276	0.374
Admission rate	737	0.640	0.218	0.524	0.681	0.795
SAT score	732	1,180.232	136.670	1,093	1,157	1,265

Note: Summary statistics of main variables.

Table 2: Patterns of Sample Selection: Share of Syllabi Included in the Sample and Institution-Level Characteristics

Panel (a): Share and Δ Share, By School Tier

	Share in OSP		ΔShare in OSP, 2010-13	
	Mean	SE	Mean	SE
Ivy Plus	0.0082	(0.0023)	-0.0016	(0.0008)
Elite	0.0219	(0.0068)	0.0115	(0.0054)
Highly Selective Private	0.0016	(0.0002)	-0.0047	(0.0000)
Highly Selective Public	0.0066	(0.0031)	0.0068	(0.0000)
Selective Private	0.0268	(0.0206)	0.0047	(0.0034)
Selective Public	0.0337	(0.0111)	0.0149	(0.0076)
Non-selective Private	0.0001	(0.0000)	0.0000	(0.0000)
Non-selective Public	0.0013	(0.0004)	0.0008	(0.0000)

Panel (b): Share and Δ Share, Correlation w/ School Characteristics

	Share in OSP		ΔShare in OSP, 2010-13	
	Corr.	SE	Corr.	SE
ln Expenditure on instruction (2013)	-0.0099	(0.0068)	-0.0035	(0.0021)
ln Endowment per capita (2000)	0.0050	(0.0078)	-0.0030	(0.0048)
ln Sticker price (2013)	-0.0051	(0.0097)	-0.0047	(0.0038)
ln Avg faculty salary (2013)	0.0194	(0.0281)	0.0087	(0.0080)
ln Enrollment (2013)	0.0084	(0.0079)	0.0038	(0.0024)
Share Black students (2000)	-0.0201	(0.0334)	-0.0254	(0.0177)
Share Hispanic students (2000)	0.0390	(0.0387)	-0.0252	(0.0359)
Share alien students (2000)	0.2092	(0.2289)	-0.0654	(0.0507)
Share grad in Arts & Humanities (2000)	0.0002	(0.0005)	-0.0000	(0.0001)
Share grad in STEM (2000)	-0.0003	(0.0006)	-0.0001	(0.0001)
Share grad in Social Sciences (2000)	-0.0002	(0.0006)	0.0000	(0.0001)

Note: The top panel shows OLS coefficients ("means") and syllabus-clustered standard errors ("SE") of a regression of each dependent variable on indicators for school tiers. The bottom panel shows OLS coefficients ("means") and syllabus-clustered standard errors ("SE") of separate regressions of each dependent variable with each independent variable. The dependent variables are the school-level share of syllabi contained in the OSP sample in 2013 (columns 1-2) and the change in this share between 2010 and 2013 columns (3-4).

Table 3: Decomposing the Gap: Contribution of Institutions, Years, Fields, Courses, and Instructors

Specification	R2	Additional share of explained variation
Year FE	0.23	.
+ Institution FE	0.26	0.03
+ Field FE	0.36	0.13
+ Course FE	0.79	0.57
+ Instructor FE	0.89	0.13

Note: Column 1 shows the R-squared of a set of OLS regressions of the gap as functions of the corresponding set of fixed effects. Column 2 shows the fixed effects of each regression, divided by one minus the R-squared of the previous regression. Each observation corresponds to a course, instructor, and year.

Table 4: The Education-Innovation Gap Across School Tier and Student Backgrounds

Panel (a) Across School Tiers	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
Ivy Plus	-0.885***	-1.375***	-0.237	-1.061***	-1.315***	-0.940***	-0.559***	-0.403
	(0.193)	(0.448)	(0.388)	(0.367)	(0.231)	(0.207)	(0.183)	(0.289)
Elite	-0.533**	-0.942**	0.078	-0.581	-0.923***	-0.647***	-0.388**	0.022
	(0.218)	(0.469)	(0.372)	(0.459)	(0.223)	(0.180)	(0.161)	(0.370)
(Highly) Selective Private	-0.337*	-0.375	0.215	-0.704**	-0.680***	-0.476**	-0.228	0.119
	(0.179)	(0.434)	(0.375)	(0.313)	(0.254)	(0.191)	(0.162)	(0.280)
(Highly) Selective Public	-0.291*	-0.376	0.340	-0.848***	-0.494**	-0.357**	-0.152	0.082
	(0.170)	(0.419)	(0.368)	(0.283)	(0.213)	(0.174)	(0.145)	(0.265)
Mean of Gap for Benchmark Group	95.595	93.651	95.191	97.626	94.598	96.757	95.334	94.195
N	1,752,602	179,167	479,389	498,480	413,060	630,315	506,839	615,392
# Inst-by-field	30,333	1,811	9,000	6,418	5,445	22,844	20,915	20,763

Panel (b) Across Median Parental Income	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
top 1%	-1.059***	-1.606***	-0.766***	-0.954***	-1.017***	-1.010***	-0.642***	-1.212***
	(0.123)	(0.357)	(0.195)	(0.280)	(0.146)	(0.159)	(0.164)	(0.180)
75-99%	-0.540***	-0.525*	-0.341***	-0.558***	-0.476***	-0.441***	-0.270***	-0.706***
	(0.087)	(0.269)	(0.132)	(0.169)	(0.117)	(0.112)	(0.098)	(0.152)
50-75%	-0.378***	-0.221	-0.205*	-0.378**	-0.390***	-0.220**	-0.172*	-0.670***
	(0.090)	(0.265)	(0.123)	(0.171)	(0.150)	(0.112)	(0.101)	(0.162)
25-50%	-0.121	-0.235	-0.023	0.097	-0.325**	0.176	-0.251**	-0.419***
	(0.108)	(0.294)	(0.203)	(0.202)	(0.165)	(0.139)	(0.127)	(0.160)
Mean of Gap for Benchmark Group	95.713	93.610	95.704	97.282	94.464	96.658	95.394	94.903
N	1,718,204	174,138	471,891	489,035	406,527	618,531	500,084	599,532
# Inst-by-field	29,838	1,784	8,863	6,304	5,355	22,468	20,611	20,401
Field x Level x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

(continued)

Table 4. Continued

Panel (c) Across % of Top Income Parents	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
>15%	-0.911***	-1.746***	-0.849***	-1.067***	0.107	-0.983***	-0.813***	-0.552
	(0.268)	(0.347)	(0.199)	(0.359)	(0.766)	(0.170)	(0.242)	(0.591)
5-15%	-0.536**	-1.502***	-0.672***	-0.410	0.482	-0.720***	-0.328	-0.115
	(0.271)	(0.303)	(0.177)	(0.383)	(0.763)	(0.159)	(0.231)	(0.601)
1-5%	-0.403	-1.041***	-0.512***	-0.615**	0.680	-0.507***	-0.322	-0.059
	(0.261)	(0.284)	(0.170)	(0.306)	(0.764)	(0.143)	(0.225)	(0.590)
0.1-1%	-0.060	-0.744***	-0.245	-0.101	0.905	-0.113	-0.173	0.237
	(0.261)	(0.287)	(0.170)	(0.308)	(0.768)	(0.140)	(0.227)	(0.590)
Mean of Gap for Benchmark Group	95.597	94.215	95.890	97.295	93.345	96.766	95.439	94.253
N	1,718,204	174,138	471,891	489,035	406,527	618,531	500,084	599,532
# Inst-by-field	29,838	1,784	8,863	6,304	5,355	22,468	20,611	20,401

Panel (d) Across Ratios of Minority Students	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
<5%	-0.290***	0.135	-0.271**	-0.289*	-0.260**	-0.362***	-0.133	-0.283*
	(0.087)	(0.260)	(0.113)	(0.168)	(0.123)	(0.104)	(0.105)	(0.152)
5-10%	-0.300***	0.057	-0.265	-0.314	-0.174	-0.319***	-0.140	-0.286
	(0.100)	(0.245)	(0.169)	(0.199)	(0.156)	(0.122)	(0.101)	(0.178)
10-20%	-0.484***	-0.158	-0.284**	-0.549***	-0.426***	-0.257**	-0.211**	-0.759***
	(0.085)	(0.250)	(0.114)	(0.160)	(0.120)	(0.105)	(0.098)	(0.150)
20-40%	-0.147	0.095	-0.056	0.019	-0.162	0.090	-0.087	-0.355**
	(0.091)	(0.243)	(0.123)	(0.176)	(0.144)	(0.117)	(0.105)	(0.153)
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of Gap for Benchmark Group	95.601	93.268	95.660	97.180	94.315	96.576	95.303	94.738
N	1,749,091	178,557	479,071	496,656	412,478	629,079	505,863	614,093
# Inst-by-field	30,219	1,806	8,970	6,392	5,422	22,761	20,844	20,699
Field x Level x Year FE								

Note: OLS coefficients of syllabus-level specifications where the dependent variable is the education-innovation gap and the independent variable are selectivity tiers (panel (a)), quantiles of median parental income at each school (panel (b)), shares of parents with incomes in the top percentile in each school (panel (c)), and share of the school's students who are minority (Black or Hispanic). Column 1 is estimated on all syllabi and includes controls for field-by-course level-by-year fixed effects; columns 2-5 show estimates by macro-field, controlling for course level-by-year fixed effects; and columns 6-8 show estimates by course level, controlling for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 5: The Education-Innovation Gap and Instructor Research Productivity and Fit

Panel (a) Instructor Change	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Instructor change	-0.0505***	-0.0946**	-0.0509*	-0.0281	-0.0616***	-0.0592**	-0.0530**	-0.0432*
	(0.0140)	(0.0376)	(0.0281)	(0.0316)	(0.0235)	(0.0263)	(0.0242)	(0.0223)
N (Course x year)	542,099	56,873	132,006	154,203	134,198	170,200	171,466	192,295
# Courses	167,861	16,719	41,620	47,909	41,211	50,517	52,242	62,171
Panel (b) Instructor Publications Per Year	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Instructor Research Productivity (Publications)	-0.402***	-0.236***	-0.269***	-0.542***	-0.222***	-0.370***	-0.245***	-0.375***
	(0.021)	(0.067)	(0.046)	(0.038)	(0.035)	(0.042)	(0.042)	(0.031)
N (Course x year)	632,450	64,079	141,407	199,361	167,845	208,249	191,750	226,960
# Course	273,064	23,596	63,747	85,140	72,758	80,751	83,039	105,325
Panel (c) Instructor Citations Received Per Year	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Instructor Research Productivity (Citations)	-0.084***	-0.019	-0.076***	-0.106***	-0.054***	-0.073***	-0.030**	-0.082***
	(0.006)	(0.016)	(0.013)	(0.013)	(0.009)	(0.012)	(0.012)	(0.009)
N (Course x year)	632,450	64,079	141,407	199,361	167,845	208,249	191,750	226,960
# Course	273,064	23,596	63,747	85,140	72,758	80,751	83,039	105,325
Panel (c) Instructor's Fit with the Course	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Instructor-Course Fit	-2.623***	-1.786***	-5.801***	-0.995***	-4.519***	-3.066***	-0.777*	-2.001***
	(0.198)	(0.517)	(0.684)	(0.348)	(0.278)	(0.503)	(0.403)	(0.254)
N (Course x year)	632,450	64,079	141,407	199,361	167,845	208,249	191,750	226,960
# Course	273,064	23,596	63,747	85,140	72,758	80,751	83,039	105,325
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS coefficients of course-level specifications where the dependent variable is the education-innovation gap and the dependent variable are indicators for an instructor change (panel (a)), instructor's publications per year (panel (b)), instructor's citations per year (panel (c)), and a measure of instructor's research fit, defined as the cosine similarity between the syllabus and the instructor's publications. Column 1 is estimated on all syllabi; columns 2-5 show estimates by macro-field; and columns 6-8 show estimates by course level. All specifications control for course ands school-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 6: The Education-Innovation Gap and Student Outcomes

	Graduation rates		log(Income)		Intergenerational mobility	
	(1)	(2)	(3)	(4)	(5)	(6)
Gap	-0.0602***	-0.0134***	-0.0569***	-0.0110*	-0.0306***	-0.0036
	(0.0079)	(0.0044)	(0.0094)	(0.0063)	(0.0058)	(0.0043)
Median parental income		0.0055***		0.0065***		0.0040***
		(0.0003)		(0.0004)		(0.0003)
Share parents in top income quintile		-1.4422***		-2.3973***		-1.2322***
		(0.2015)		(0.2984)		(0.2127)
Share minority students		0.0776***		0.2681***		0.1885***
		(0.0258)		(0.0427)		(0.0272)
Year FE	Yes	Yes	Yes	Yes	No	No
Tier FE	No	Yes	No	Yes	No	Yes
R-squared	0.0809	0.6994	0.0597	0.5597	0.0349	0.4997
Mean of dep var	0.5637	0.5637	–	–	0.2914	0.2914
N	16143	15882	4697	4629	778	776

Note: OLS estimates of student outcomes on the gap and school-level controls. Columns 1-4 are at the school-by-year level and control for year fixed effects; columns 5-6 are at the school level. Columns 2, 4, and 6 control for a school's selectivity tier. Standard errors are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 7: The Education-Innovation Gap and Student Outcomes: Heterogeneity By School Characteristics

	Graduation rates		log(Income)		Intergenerational mobility	
	(1)	(2)	(3)	(4)	(5)	(6)
Gap * Ivy Plus/Elite	0.0204		0.0133		0.0315*	
	(0.0153)		(0.0292)		(0.0169)	
Gap * (Highly) Selective Private	-0.0134***		-0.0019		-0.0018	
	(0.0052)		(0.0120)		(0.0070)	
Gap * (Highly) Selective Public	-0.0161**		-0.0191**		-0.0095	
	(0.0079)		(0.0080)		(0.0063)	
Gap * Non-selective	-0.0222*		-0.0227*		-0.0045	
	(0.0128)		(0.0124)		(0.0121)	
Gap * Parent Income Q4 (High)		-0.0033		0.0102		0.0161*
		(0.0069)		(0.0151)		(0.0092)
Gap * Parent Income Q3		-0.0156*		-0.0122		-0.0116
		(0.0086)		(0.0142)		(0.0079)
Gap * Parent Income Q2		-0.0137		-0.0273***		-0.0098
		(0.0113)		(0.0099)		(0.0087)
Gap * Parent Income Q1 (Low)		-0.0116		-0.0051		-0.0010
		(0.0072)		(0.0104)		(0.0082)
Median parental income	0.0055***		0.0065***		0.0040***	
	(0.0003)		(0.0004)		(0.0003)	
Share parents in top income quintile	-1.3867***	0.4064***	-2.3274***	0.1003	-1.1661***	0.1230
	(0.2075)	(0.0974)	(0.3030)	(0.1904)	(0.2173)	(0.1410)
Share minority students	0.0782***	-0.0570**	0.2677***	0.0994**	0.1885***	0.0795***
	(0.0256)	(0.0237)	(0.0431)	(0.0451)	(0.0272)	(0.0233)
Year FE	Yes	Yes	Yes	Yes	No	No
Tier FE	Yes	Yes	Yes	Yes	Yes	Yes
Parent income quartile FE	No	Yes	No	Yes	No	Yes
N	15882	15882	4629	4629	776	776
Mean of dep var	0.5637	0.5637	–	–	0.2914	0.2914
R-squared	0.7008	0.7192	0.5613	0.5510	0.5030	0.5181

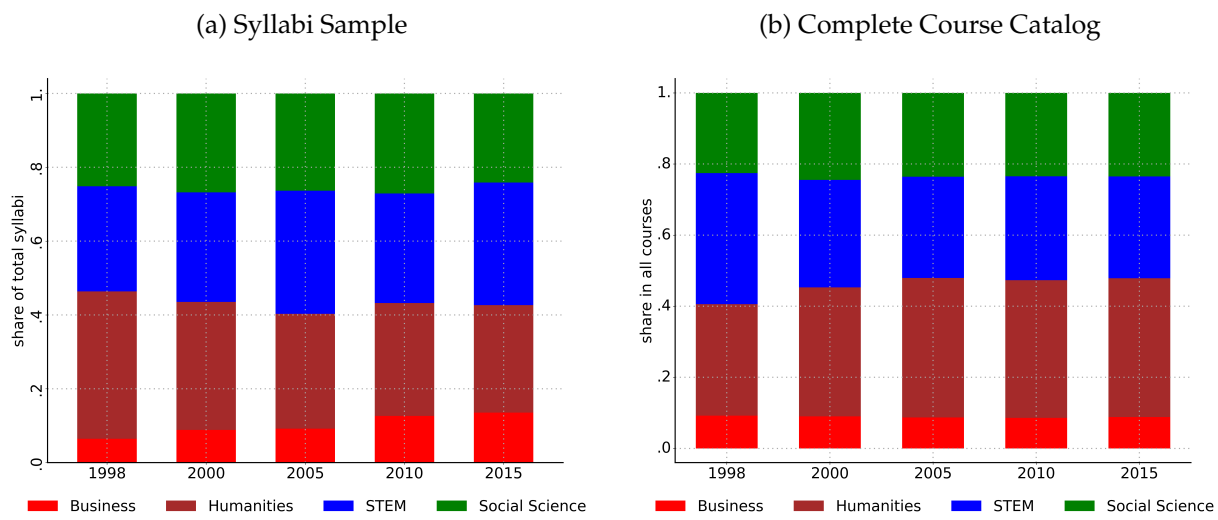
Note: OLS estimates of student outcomes on the gap, interacted with selectivity tiers (columns 1, 3, and 5) and quantiles of the parental median income distribution (columns 2, 4, and 6), and school-level controls. Columns 1-4 are at the school-by-year level and control for year fixed effects; columns 5-6 are at the school level. Columns 2, 4, and 6 control for a school's selectivity tier. Standard errors are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Appendix

For online publication only

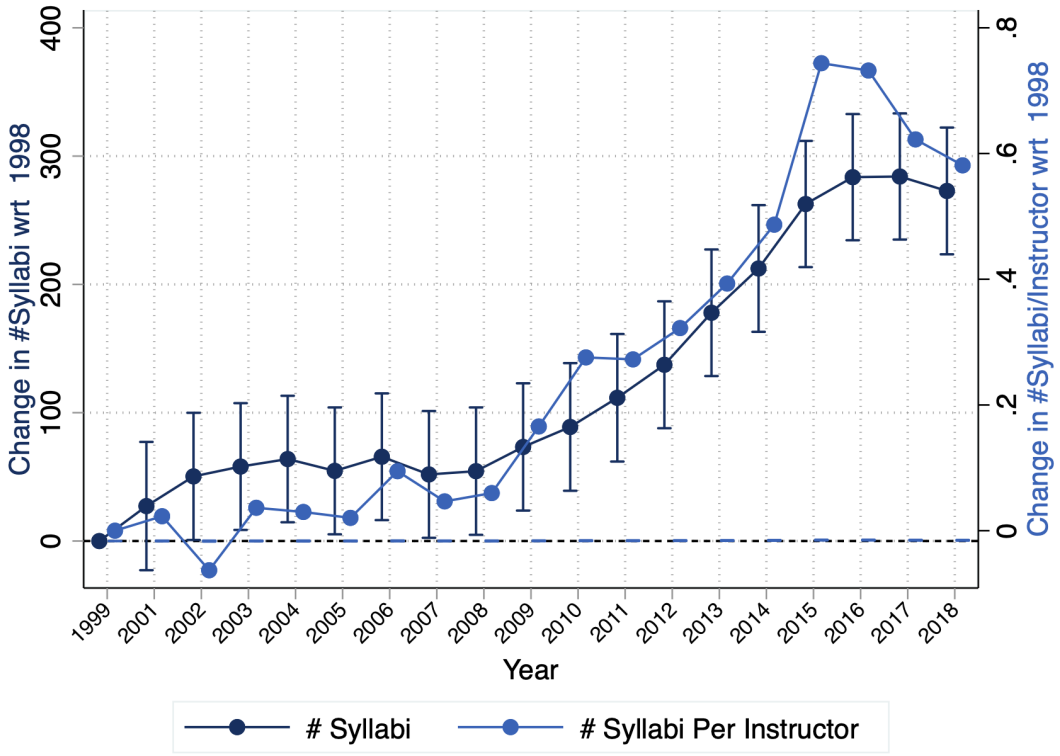
Additional Tables and Figures

Figure AI: Stable Field Coverage of the Syllabi Data



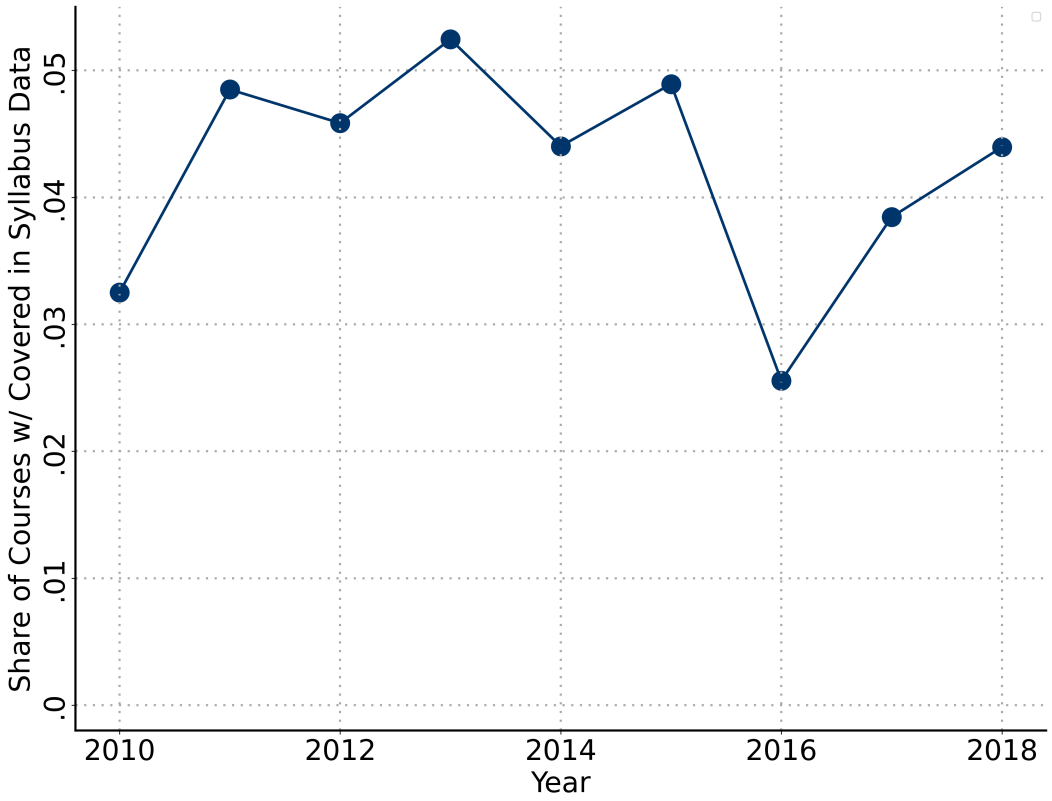
Note: Syllabi field composition by five-year periods. Panel (a) is for the syllabi sample. Panel (b) is for all courses collected from course catalog.

Figure AII: Syllabi Per Year and Syllabi Per Instructor Per Year



Note: Trends in the number of syllabi per year (solid line) and syllabi per instructors per year (dashed line), controlling for institution, and relative to 1993. The number of instructors for each institution is taken from IPEDS.

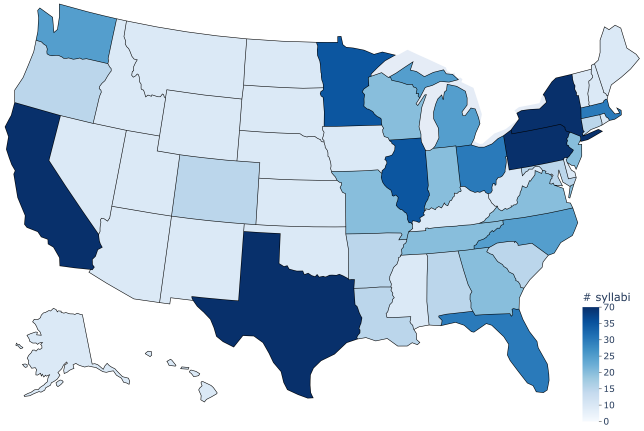
Figure AIII: Selection Into the OSP: Share of Covered Syllabi, Catalogue Data



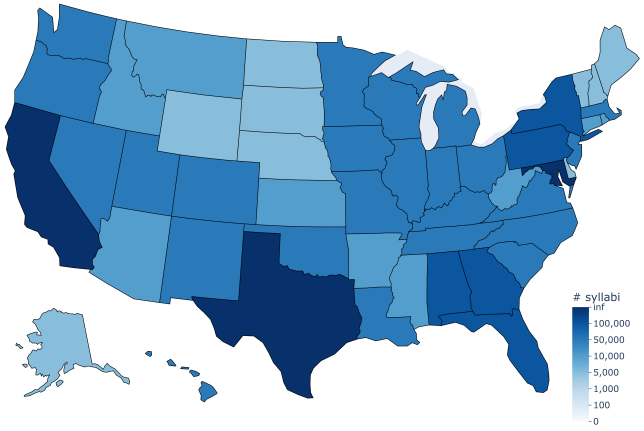
Note: Share of syllabi from the full catalogue of 161 selected institutions that are included in our sample. Catalogue data are collected from university archives.

Figure AIV: Syllabi Across The United States

Panel a) Number of Institutions in Each State

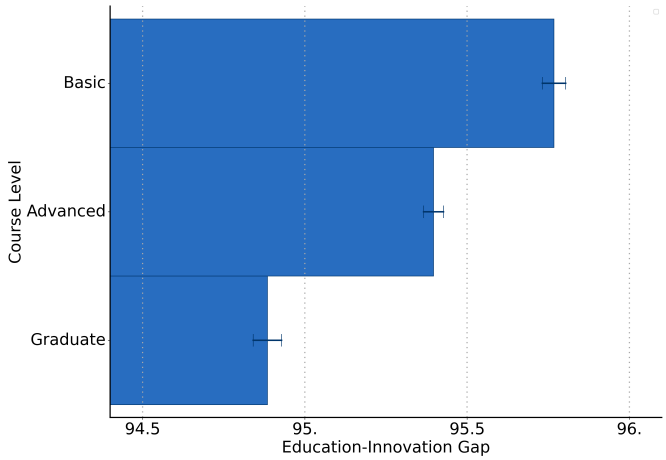


Panel b) Number of Syllabi in Each State



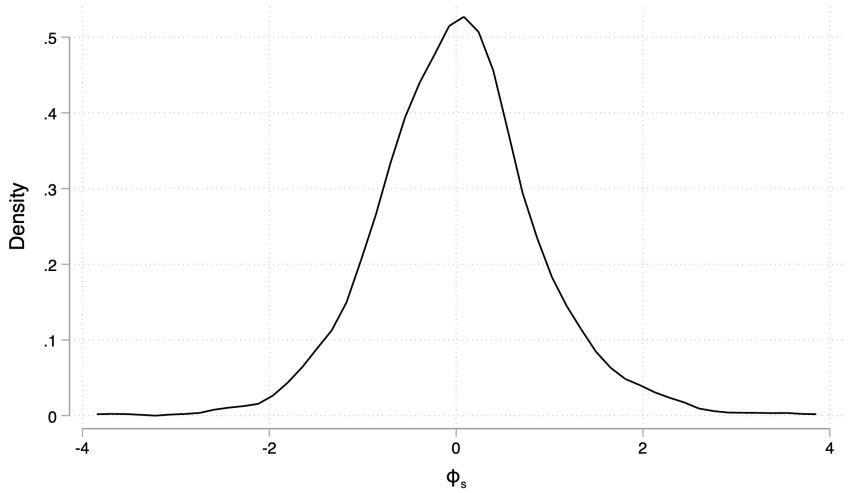
Note: The map plots the number of IPEDS institution (top panel) and the number of syllabi (bottom panel) from each state.

Figure AV: Education-Innovation Gap Across Different Course Levels



Note: Average education-innovation gap across courses at different levels: basic, advance (undergraduate), and graduate level.

Figure AVI: Distribution of School-Level Gap



Distribution of ϕ_s, the school-level component of the gap, as defined in equation (3).

Table AI: Categorization of Course (Macro-)Fields

Macro-field	Fields
Business	Business, Accounting, Marketing
Humanities	English Literature, Media / Communications Philosophy, Theology, Criminal Justice Library Science, Classics, Women's Studies Journalism, Religion, Sign Language Music, Theatre Arts, Fine Arts, History Film and Photography, Dance, Anthropology
STEM	Mathematics, Computer Science, Biology Engineering, Chemistry, Physics Architecture, Agriculture, Earth Sciences Basic Computer Skills, Astronomy, Transportation Atmospheric Sciences
Social Sciences	Psychology, Political Science, Economics Law, Social Work, Geography Linguistics, Sociology Education
Vocational	Fitness and Leisure, Basic Skills Mechanic / Repair Tech, Cosmetology Culinary Arts, Health Technician, Public Safety

Note: Mapping between the "macro-fields" used in our analysis and syllabi's "fields" as reported in the OSP dataset.

Table AII: List of Institutions in the Catalog Data

Institution	Institution
Aiken Technical College	Minnesota State University Moorhead
Alabama Agricultural and Mechanical University	Mississippi College
Alabama State University	Mississippi Community College Board
Alexandria Technical and Community College	Missouri State University
Arkansas Tech University	Mitchell Technical Institute
Asnuntuck Community College	Montgomery College
Bay Path University	Morehead State University
Benedictine University	Mountain Empire Community College
Bentley University	Mountwest Community and Technical College
Bluegrass Community and Technical College	Mt. San Antonio College
Briar Cliff University	New Mexico State University Alamogordo
Brown University	Niagara University
Bryan College	Nichols College
California Baptist University	North Carolina State University
California Lutheran University	North Florida Community College
California Polytechnic State University	Northwest Arkansas Community College
Camden County College	Oakwood University
Campbell University	Oral Roberts University
Cardinal Stritch University	Orangeburg-Calhoun Technical College
Carlow University	Oregon State University
Catawba College	Oxnard College
Cecil College	Penn State New Kensington
Cedarville University	Plymouth State University
Center for Creative Studies	Princeton University
Cerritos College	Richland Community College
Coe College	Robeson Community College
College of Alameda	Rocky Mountain College
College of Southern Nevada	SUNY College at Old Westbury
College of the Siskiyous	SUNY Oneonta
Columbia University	SUNY Orange
Concordia University Texas	San Diego Mesa College
Copiah-Lincoln Community College	San Diego Miramar College
County College of Morris	San Diego State University
Dartmouth College	South Arkansas Community College
Daytona State College	Southern University at New Orleans
Dominican University	Spring Arbor University
Duke University	Spring Hill College
Eastern Nazarene College	Stanford University
ENMU-Ruidoso Branch Community College	State University of New York at Potsdam
Elmhurst College	Suffolk County Community College
Florida Gulf Coast University	Texas Lutheran University
Florida Institute of Technology	The University of Texas Rio Grande Valley
Fresno Pacific University	Three Rivers Community College
Frostburg State University	Trevecca Nazarene University

(continued)

Table AII. Continued

Institution	Institution
George Mason University	Trocaire College
Georgia State University	University of Akron
Glendale Community College	University of Central Oklahoma
Grays Harbor College	University of Chicago
Green River Community College	University of Colorado Denver
Grossmont College	University of Evansville
Helena College University of Montana	University of Louisville
Herkimer County Community College	University of Maine at Presque Isle
Hibbing Community College	University of Missouri-St. Louis
Hood College	University of Montana
Hudson County Community College	University of North Carolina at Chapel Hill
Indiana University Northwest	University of North Dakota
Iowa Central Community College	University of North Texas
Jackson State Community College	University of Notre Dame
Jefferson State Community College	University of Pennsylvania
Kankakee Community College	University of Pittsburgh
Kellogg Community College	University of South Carolina Aiken
Kettering University	University of South Florida Sarasota-Manatee
Keystone College	University of Wisconsin-River Falls
King's College - Pennsylvania	Upper Iowa University
Kutztown University	Vanderbilt University
Lake Forest College	Virginia Highlands Community College
Las Positas College	Wayne State College
Lassen Community College	Weber State University
Leeward Community College	Webster University
Lincoln University - Missouri	Wenatchee Valley College
Long Beach City College	Wentworth Institute of Technology
Los Medanos College	Wesleyan University
Louisiana State University in Shreveport	Western Dakota Technical Institute
Macmurray College	Western State Colorado University
Marian University - Indiana	William Jewell College
Marian University - Wisconsin	William Woods University
Marietta College	Yale University
Martin Luther College	Youngstown State University
Martin Methodist College	Yuba College
Millsaps College	

f schools for which we collected course catalog data.

Table AIII: School Characteristics of Schools In and Out of Catalog Data

	Mean for Institutions In the Sample # Institutions = 158	Mean for Institutions Out of the Sample # Institutions = 1,956	t-statistics	p-values
ln Expenditure on instruction (2013)	8.693	8.601	-1.725	0.085
ln Endowment per capita (2000)	6.857	6.483	-1.304	0.193
ln Sticker price (2013)	9.197	9.153	-0.520	0.603
ln Avg faculty salary (2013)	8.890	8.850	-1.897	0.058
ln Enrollment (2013)	8.708	8.634	-0.685	0.494
Share Black students (2000)	0.109	0.112	0.153	0.879
Share Hispanic students (2000)	0.063	0.065	0.183	0.855
Share alien students (2000)	0.025	0.022	-1.030	0.303
Share grad in Arts & Humanities (2000)	7.581	7.958	0.382	0.703
Share grad in STEM (2000)	14.861	14.050	-0.772	0.440
Share grad in Social Sciences (2000)	21.068	19.202	-1.342	0.180

Note: Balance test of universities in and out of the catalog sample.