The Education-Innovation Gap*

Barbara Biasi† Song Ma‡

October 1, 2021

Please click here for the most updated version

Abstract

This paper examines the diffusion of frontier knowledge through higher education courses. Using the text of 1.7 million college and university syllabi and 20 million articles published in top scientific journals since 1975, we construct a new measure: the "education-innovation gap," defined as the ratio of textual similarities between each syllabus and (i) articles published 15 years before and (i) articles published 1 to 3 years before. We use this measure to document four findings. First, the gap varies substantially both across and within schools, with instructors accounting for 40 to 50 percent of the variation. Second, the gap is lower in schools that are more selective and serve fewer disadvantaged and minority students. Third, the gap decreases after the instructor of a course changes, and it is lower for courses taught by research-active faculty. Fourth, the gap is correlated with students' graduation rates and incomes after graduation. These findings are robust to the use of alternative measures of course novelty.

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

Keywords: Education, Innovation, Text Analysis, Inequality

First Draft: September, 2019. We thank Jaime Arellano-Bover, David Deming, David Robinson, Kevin Stange, Sarah Turner, and seminar and conference participants at Yale, Duke, Erasmus University, Maastricht University, Queens, Stockholm School of Economics, NBER (Education, Entrepreneurship, Innovation), AEA, CEPR/Bank of Italy, Junior Entrepreneurial Finance and Innovation Workshop, SOLE, IZA TOM and Economics of Education Conferences, and CESifo Economics of Education Conference for helpful comments. Xugan Chen provided excellent research assistance. We thank the 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.

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

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

1 Introduction

In a knowledge-based economy, new ideas and knowledge – non-rival goods with increasing returns – spur technological innovation and are essential to economic growth (Romer, 1990). It is therefore crucial to understand how ideas and knowledge are produced and disseminated. Education systems (particularly higher education ones) play a crucial role as knowledge providers (Biasi, Deming, and Moser, 2020). Given the upward trend in the "burden of knowledge" required to innovate (Jones, 2009), the importance of these programs is likely to grow.

Not all higher education programs, however, are created equal. Just like there is heterogeneity in the economic returns they produce (Hoxby, 1998; Altonji et al., 2012; Chetty et al., 2019, among others), there might be differences in the extent to which programs equip students with frontier knowledge. The goal of this paper is to quantify these differences by examining the content of higher education instruction. Specifically, we want to measure the distance between the knowledge content of each *course* – as described in its syllabus – and the knowledge frontier, represented by top academic articles recently published in the course's field.

To quantify this distance, we develop a new metric: the *education-innovation gap*, designed to capture the similarity between the content of a course and older knowledge (contained in articles published decades ago) relative to new, frontier knowledge (contained in recently published articles). For example, a Computer Science course that teaches *Visual Basic* (an obsolete programming language) in 2018 would have a larger gap than a course that teaches *Julia* (a recent and updated programming language), because *Visual Basic* is more frequent among old academic articles and *Julia* is more frequent among recent articles.[1]

We construct this measure using a "text as data" approach (Gentzkow et al., 2019). Specifically, we compare the raw text of 1.75 million college and university syllabi, covering about 540,000 courses in 69 different fields taught at nearly 800 US institutions between 1998 and 2018, with the title, abstract, and keywords of 20 million academic publications appeared in top journals since each journal's foundation. We first represent each document as a binary vector, whose elements correspond to words of a dictionary (we use set of all words ever listed on *Wikipedia*) and equal one

[1] First released by Microsoft in 1991, *Visual Basic* is still supported by Microsoft in recent software frameworks. However, the company announced in 2020 that the language would not be further evolved (https://visualstudiomagazine.com/articles/2020/03/12/vb-in-net-5.aspx, retrieved September 30th, 2020). *Julia* is a general-purpose language initially developed in 2009. Constantly updated, it is among the best languages for numerical analyses and computational science. As of July 2021 it was used at 1,500 universities, with over 29 million downloads and a 87 percent increase in a single year (https://juliacomputing.com/blog/2021/08/newsletter-august/, retrieved September 30, 2021).

if the document contains the corresponding word. To account for the importance of a word in the document, its popularity in research at a given point in time, and its use in the English language we weigh each vector element by the ratio between the word's frequency in the document and its frequency in all documents published in previous years (similar to Kelly et al., 2018).

Using these weighted word vectors, we compute the cosine similarity (a measure of vectorial proximity) between each syllabus and each article. We then construct the education-innovation gap of a syllabus as the ratio between the average cosine similarity with articles published 15 years prior and the similarity with articles published one year priot. By construction, the gap is higher 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.

A few empirical regularities 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 (or in addition to) 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 (which we multiply by 100 for simplicity) is equal to 95, which indicates that courses tend to be more similar to newer than to older research. However, a significant amount of variation exists across syllabi. Simulations where we manually change the content of each syllabus indicate that, in order to move a syllabus from the 25th percentile (92) to the 75th percentile (99) of the gap distribution, we would have to replace approximately 74 percent of its content with "newer" knowledge, i.e., words that are most frequent in recent publications. A variance decomposition exercise indicates that differences across fields explain 4 percent of the total variation in the gap; differences across schools explain an additional 2 percent, differences across courses explain 29 percent, and differences across instructors explain 40 percent of the variation.

This decomposition exercise also suggests that students at various schools are differentially exposed to frontier knowledge. More selective schools (such as Ivy-Plus, Chetty et al., 2019) have a significantly lower 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 and Elite schools, 10 per-

cent of its content would have to be replaced with newer knowledge. While the gap is lower for schools with higher endowments and instructional expenditures, it is not related to characteristics such as the share of non-ladder faculty or the designation of Liberal Arts College or Land Grant College.

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. In fact, we find cross-school disparities even within selectivity tiers. Accounting 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 inequality access to frontier knowledge across students from different socio-economic backgrounds.

The decomposition exercise also reveals a crucial role for instructors in determining the content of the courses they teach.[2] To better understand the role of instructors, we leverage faculty turnover within each course. We find that the education-innovation gap of a course declines significantly when the person who teaches it changes, suggesting that instructor who take over a course from someone else update its content. However, not all instructors are equal: Even accounting for course fixed effects, the gap is significantly lower for course taught by faculty with higher research productivity, measured with academic publications and citations.

Research-active instructors might be better updated about the frontier of research and more likely to cover this type of content in their courses, which results in a lower gap. In line with this hypothesis, we find that the gap is lower when the instructor's own interests are closer to the topic of the course. We also find a negative relationship between the gap and research inputs available to the instructor, such as the number and size of government grants. These results indicate that the assignment of instructors to courses can be a powerful tool to expose students to frontier knowledge. They also suggest that public investments in research can generate additional returns in the form of more updated instruction.

Next, we try to understand whether these differences matter for student outcomes. To answer this question, the ideal experiment would randomly allocate students to courses with different gaps. In the absence of this random variation, we set on the more modest goal of characterizing the empirical relationship between the education-innovation gap and student outcomes, such as gradua-

[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).

tion rates, incomes after graduation, and intergenerational mobility. In an attempt to account for endogenous differences across schools, we control for a large set of school observables such as institutional characteristics, various types of expenditure, instructional characteristics, enrollment by demographic groups and by major, selectivity, and parental background. We find that the gap is negatively related to graduation rates and students' incomes, with economically meaningful magnitudes. The relationship with intergenerational mobility is instead indistinguishable from zero.

In the final part of the paper, we probe the robustness of our results to the use of additional measures of novelty of a course' content. We consider three measures: the share of all "new" knowledge contained in a syllabus (designed not to penalize a syllabus that contains old and new knowledge compared with one that only contains new knowledge; a measure of "tail" knowledge, aimed at capturing the presence of the most recent content; and a measure of soft skills, devised to capture the non-academic novelty of a course. All the results are qualitatively unchanged when we use these measures.

This paper contributes to several strands of the literature. First, we characterize 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.

Second, this paper relates to the literature on the "production" of knowledge. Earlier works (Nelson and Phelps, 1966; Benhabib and Spiegel, 2005) have highlighted an important role for human capital and education in the diffusion of ideas and technological advancements. Certain fields, such as STEM, have been shown to be particularly important for innovation (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019).[3] Instead of just looking at differences across fields, here we take a more "micro" approach, and we quantify differences across courses in the provision of frontier knowledge, which might be particularly important for growth.

[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.

Next, our findings contribute to recent studies on the "democratization" (or lack thereof) of access to valuable knowledge. For example, Bell et al. (2019) have shown that US inventors (i.e., people with at least one patent) 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.

Lastly, we use of the text of course 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 calculate cosine similarities between the text of patent documents to measure patent quality, and Gentzkow and Shapiro (2010), who characterize 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.

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 Appendix.

2.1 College and University Course Syllabi

College and university syllabi information was collected by the Open Syllabus Project (OSP).[4] The raw data cover more than seven million English-language syllabi of courses taught in over 80 countries, dating from as early as the 1960s until 2019.

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. We extract four pieces of information from

[4]OSP, originally part of the American Assembly, 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.

the text of each syllabus: (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] Names and codes allow us to construct the course level (introductory, advanced, or graduate).[6] We also use information on the course's field as provided by OSP. Specifically, OSP assigns each syllabus to one of 69 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, 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 1011 words; the median is 660.

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 and Ma (2020)." We successfully identify this portion for 71 percent of all syllabi. For each reference item, we also collect information on 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

[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 AXI.

[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".

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 99.9 percent of all syllabi.

Sample restrictions and description Panel A of Table 1 describes the characteristics of the syllabi sample. To maximize consistency over 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 542,251 courses at 767 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 keywords.

2.3 Course Instructors: Research Productivity and Funding

We use information from Microsoft Academic (MA) to measure the research productivity of all people listed as instructors in the syllabi. MA lists publications, working papers, other manuscripts,

[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 with the highest average similarity.

and patents of each listed researcher, 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 as the number of publications and the number of citations received in the previous five years.[14] On average, instructors publish 1.5 articles per year and 4.6 articles in the most recent five years, receiving 24 citations per year (Table 1, panel B). The distribution of both citations and publication counts are highly skewed: The median instructor in our sample did not publish any article in the previous five years, as a result, received no citations.

We complement publications data from MA with information on government grants received by each researcher, to measure public investment in academic research. We focus on two among the main funding agencies of the U.S. government: the National Science Foundation (NSF) and the National Institute of Health (NIH).[15] Our grant data include 480,633 NSF grants active between 1960 and 2022 (with an average size of $582K in 2019 dollars, Table 1, panel B) and 2,566,358 NIH grants active between 1978 and 2021 (with an average size of $504K). We link grant data to syllabi instructors via a fuzzy matching between the names of the grant investigators and the name of the instructors (more detail can be found in the Data Appendix). Fifteen percent of all syllabi instructors are linked to at least one grant; among these, the average instructor receives 14 grants with an average size of $5,224K.

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).[16] 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

[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] These data are published by each agency, at https://www.nsf.gov/awardsearch/download.jsp and https://exporter.nih.gov/ExPORTER_Catalog.aspx. We accessed these data on May 25, 2021.

[16] 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.

students, enrollment figures for different demographic groups, completion rates, and graduation rates.We link each syllabus to the corresponding IPEDS record as provided by OSP. We are able to successfully link all syllabi in our sample.

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 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 1997-98 to 2007-08, 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 $97,917 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 0.22, with a standard deviation of 0.17. Graduation rates average 61.4 percent in 2018, whereas students' incomes ten years after school entry, for the 2003–04 and 2004–05 cohorts, are equal to $45,035. 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) with hand-collected course catalogues in the

archives of each school.[17] This allows us to compare the syllabi sample to the population for these schools and years.

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 0.01 percent for non-selective private schools to 3.37 percent for highly selective and selective public schools (left columns); these shares are, however, statistically indistinguishable across tiers. The same is true for the 2010-2013 *change* in the share of covered syllabi (right 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.

[17] 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 AXII 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 AXIII).

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 publications. 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 CS229 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 retrospec-

tive 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 $[t(d) - \tau - 2, t(d) - \tau]$.[18]

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

To gauge the extent to which the education-innovation gap is able to capture 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 95.8; 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.

Third, we demonstrate that our measure performs well in capturing the extent to which a syllabus contains old and new knowledge. We do so by constructing a set of 1.7 million fictitious syllabi as sets of knowledge words, each with a given ratio of old to new words (defined, respectively, as those in the top 5 percent in terms of frequency in the new publication corpus between t-3 and t-1 or in the new publication corpus between t-3 and t-1 but not in the old publication corpus between t-15 and t-13; and those in the top 5 percent in terms of frequency in the old publication corpus between t-15 and t-13 or in the old publication corpus between t-15 and t-12 but not in the new publication corpus between t-3 and t-1), and calculating the education-innovation gap for each

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

of them. The gap bears a strongly relationship with the ratio of old to new words, with a correlation of 0.96 (Figure 2, panel a).[19]

Lastly, we simulate how changing the content of a course translates into changes in the education-innovation gap. Specifically, we progressively replace "old" words with "new" words in a randomly selected subsample of 100,000 syllabi and re-calculate the gap for each syllabus as we replace more words. This exercise shows that the gap monotonically decreases as we replace old words with new ones (Figure 2, panel b). This simulation is also useful to gauge the economic magnitude of changes in the gap. In particular, a unit change in the gap requires replacing 10 percent of a syllabus's old words (or 34 old words, compared with 331 words for the median syllabus).

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 95.0, with a standard deviation of 5.9, a 25th percentile of 91.6, and a 75th percentile of 98.8. 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 74 percent of its content (or 245 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.[20] 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

[19] This simulation is described in greater detail in the Online Data Appendix.

[20] 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.

explain 13 percent.[21]

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.

4 The Education-Innovation Gap Across Schools

The decomposition exercise indicates that differences across schools explain approximately 2 percent of the total variation in the gap. Albeit small, cross-school differences might reflect disparities in access to frontier knowledge among students with different backgrounds, if schools with different gaps also serve different student populations. To assess this, we explore whether the education-innovation gap is related to the characteristics of each school and of the students it serves.

4.1 Institutional Characteristics, Finances, and Faculty Composition

We begin by testing how the education-innovation-gap relates to three sets of school attributes: (i) institutional, such as the sector (public or private) and indicators for Liberal Arts Colleges (LAC) and schools classified as R1 ("Very High Research Intensity") according to the Carnegie classification; (ii) financial, such as endowment and spending on instruction, faculty salaries, and research; (iii) and faculty, such as the share of non-ladder faculty, the share of tenure-track (non-tenured) faculty, and the number of academic publications per faculty.

To estimate these correlations accounting for field, course level, and year of the syllabus we estimate the following specification:

$$\text{Gap}_i = X_i\boldsymbol{\beta} + \phi_{f(i)l(i)t(i)} + \varepsilon_i,$$

where Gap_i measures the education-innovation gap of syllabus i, taught in school $s(i)$ in year $t(i)$. The variable X_i is the institutional characteristic of interest, and field-by-level-by-year fixed effects

[21] 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.

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.

Estimates of β for each school characteristics are shown in Figure 4. Public schools have a slightly higher gap, but the difference is indistinguishable from zero. No differences in the gap emerge between LACs and other schools; R1 schools have instead a 0.2 lower gap.

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 indicates that, in order to close the difference in the gap between R1 and other institutions, we would have to replace approximately 2 percent of the knowledge content in syllabi of non-selective schools (7 terms). The difference between R1 and other institutions, although significant, is therefore quite small.

A statistically and economically significant relationship exists between the gap and financial characteristics, such as endowment and spending on instruction, faculty salary, and research. For example, a 10-percent increase in instructional spending is associated with a 3.5 lower gap, or a 35 percent change in the syllabus; a 10-percent increase in research spending is associated with a unit lower gap or a 10 percent change in the syllabus.

Perhaps surprisingly, schools employing a higher share of non-ladder faculty (such as clinical and adjunct professors and lecturers) do not appear to have a higher gap. Instead, schools where a higher share of faculty is on the tenure track but still untenured (such as assistant professors and untenured associate professors) have a higher gap. This could be due to a tenure process that places more emphasis on research than on teaching quality. Lastly, the average number of publications per faculty member is also negatively related to the gap, suggesting that more research-active professors are more likely to include frontier knowledge in their course.

Taken together, these findings reveal that wealthier institutions and those with a higher focus on research offer courses closer to the research frontier. We will revisit this finding in Section 5, where we explore the relationship between the gap and individual characteristics of the instructors.

4.2 School Selectivity

Next, we study how the gap differs across schools that admit different shares of applicants. Following Chetty et al. (2019), we bin schools in five "tiers" according to sector and selectivity in admissions, as measured by Barron's 2009 ranking. "Ivy Plus" include Ivy League universities and the University of Chicago, Stanford, MIT, and Duke. "Elite" schools are all the other schools classi-

fied 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.[22]

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 the vector $\mathbf{S}_i'$ contains indicators for selectivity tiers, and everything is as before.

Point estimates of the coefficients vector $\boldsymbol{\beta}$ in equation (2), shown in panel (a) of Figure 5, 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 and Elite schools have the smallest gap, at 94.9. The gap increases up to to 95.9 for non-selective schools. These estimates imply that, in order to close the difference in the gap between Ivy-Plus and non-selective schools, one would have to replace approximately 10 percent of the knowledge content in syllabi of non-selective schools (34 terms).

In Appendix Table AI (panel a) we re-estimate equation (2) for syllabi in different fields and course levels, using non-selective schools as the reference tier. Columns 1-4, estimated by macro-fields, reveal that differences in the gaps across selectivity tiers are most pronounced for Business, STEM, and Social Science. Columns 5-7, estimated by level of the course (undergraduate basic, advanced, and graduate), indicate that differences are largest for basic undergraduate and graduate courses.

4.3 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

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

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 0.16 lower gap (Figure 6, 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 5, 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 5, 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 7 percent of the total knowledge content of the average syllabus, or 24 knowledge terms.

In panel b of Table AI we re-estimate these specifications for different subgroups of syllabi, using schools in the bottom half 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 half is most pronounced for STEM (1.04, column 3). The difference is present across all course levels, but it is largest for basic undergraduate courses (0.77, column 5).

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 6 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 -8.6 (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 0.9 lower gap, equivalent to a 9 percent difference in the syllabus syllabus content, or 31 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 5, 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 31 knowledge terms to the average syllabus, or 9 percent of its content.

4.4 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 3.2 and significant at 1 percent (Figure 6, panel (c). This indicates that a ten-percent increase in the share of minority students in each school is associated with a 3.3 percent difference in the content of the average syllabus, or 11 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 5, 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 10 knowledge words in the average syllabus, or 3.1 percent. These estimates are robust to controls for school selectivity (Figure 5, panel (d), lighter series).

These patterns are confirmed by the estimates in panel c of Table AI, where we use schools with more than 70 percent minority students as the reference group. Differences across schools by share of minority student are larger for courses in Humanities (-0.27, column 3) and STEM (-0.26, column 3). They are also larger for graduate courses (column 7).

5 The Role of Instructors

Our decomposition indicates that instructors explain most of the variation in the gap not only across, but also within schools. To better understand how instructors impact the content 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 exploit turnover of instructors across courses over time.

5.1 The Education-Innovation Gap When The Instructor Changes

We begin by studying how the content of a course changes when a new person starts teaching it. We estimate an event study of the gap in a 8-years window around the time of the instructor change:

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

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.[23] We restrict our attention to courses taught by a maximum of two persons in each year and we set $t(i) - T_c = 0$ for all courses without an instructor change, which thus serve as the 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 7, indicate that a change in a course's instructor is associated with a sudden decline in the education-innovation gap. Estimates are indistinguishable from zero and on a flat trend in the years leading to an instructor change; the year of the change, the

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

gap is 0.09 lower. This decline is equivalent to replacing 1 percent of the content of a syllabus, or 3 knowledge words.

In Table 4 (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 by about 0.1 on average (3.4 additional words or 1 percent of a course's content, column 1, significant at 1 percent). The decline is largest for Business, Humanities, and STEM courses (columns 2, 3, and 4), as well as for and graduate courses (column 8).

These results confirm that instructors play a crucial role in shaping the content of the courses they teach. They also suggest that people 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' Characteristics

While on average a course experiences a decline in the gap when the instructor changes, this average could mask substantial differences across instructors. For example, the decline could be larger if the instructor is more research-active, and thus better informed on frontier knowledge. Similarly, the gap could be lower if the instructor is an expert on the topics covered by the course, i.e., if their research interests are in line with the course. We now explore these possibilities.

Research productivity We begin by studying how the instructor's research productivity, measured using individual counts of citations and publications in the last five years, relates to the gap. We obtain information on publications and citations 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. The 90th percentile of the publications distribution ranges from a minimum of 0 for Humanities to a maximum of 12 for STEM; the same percentile of the citations distribution ranges from 0 for Humanities to 221 for Social Sciences.

Panels a and b of Figure 8 show a binned scatterplot of the gap and either citations (panel a) or publications (panel c) in the prior 5 years, controlling for field effects.[24] The relationship between the gap and instructors' productivity is significantly negative for both measures of productivity.

This negative relationship is confirmed in Table 5 (column 1), which shows estimates of the education-innovation gap (measured at the course-year level) as a function of within-field quartiles

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

of instructor publications (panel a) and citations (panel b); the omitted category are courses whose instructors do not have any publications or citations. In these specifications we control for course, field-by-year fixed effects, to account for unobserved determinants of the gap that are specific to a course in a given field and year. This implies that these estimates are obtained out of changes in instructors for the same course over time.

Estimates on the full sample of syllabi indicate that the gap progressively declines as the number of instructor publications and citations grows. In particular, a switch from an instructor without publications and one with a number of publications in the top quartile of the field distribution is associated with a 0.1 decline in gap (equivalent to changing 3.4 terms or 1 percent of a course's syllabus, Table 5, panel a, column 1, significant at 5 percent). Similarly, a switch from an instructor without citations to one with citations in the top quartile is associated with a 0.1 lower gap (panel b, column 1, significant at 5 percent). These relationships are stronger for Humanities and Social Science courses (columns 3 and 5) and for courses at the graduate level (column 8).

Fit between the instructor and the course These findings indicate that instructors who produce more and better cited research teach courses with a lower gap. A possible explanation for this finding is 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 for courses closer in terms of topics to the instructor's own research.

To test for this possibility, we construct a measure of "fit" between the course and the instructor's research, defined as the cosine similarity between the set of all syllabi from the same course across schools and the instructor's research in the previous 5 years.[25] One attractive property of this measure is that it is does not uniquely reflect the content of the syllabus itself, which is of course directly shaped by the instructor; rather, it aims at capturing the content of all courses on the same topic. We then correlate this measure with the education-innovation gap, controlling for course and field-by-year fixed effects. Estimates of this relationship indicate that a one-standard deviation increase in instructor-course fit is associated with a 0.09 decline in the gap (Table 4, panel b, significant at 5 percent). This relationship is particularly strong for STEM and Social Science courses (column 4) and those at the advanced undergraduate level (column 6).

[25] Constructing this measure requires obtaining a unique identifier for courses on the same field or topic (e.g. Machine Learning) across schools. The Online Appendix details the procedure we use to perform this.

Research funding Our results so far indicate a positive relationship between research output and the education-innovation gap. We now test whether the same relationship holds for research inputs, such as government grants. Data on the number and size of NSF and NIH grants received by each instructor reveals a negative relationship between the gap and these two measures of research inputs (Figure 8, panel c for the number of grants and panel d for the grant amount).

This relationship is confirmed by the estimates in Table 6. Controlling for course and field-by-year effects, a switch from an instructor who never received a grant to one with a number of grants in the top quartile of the field distribution is associated with a 0.08 reduction in the gap (panel a, column 1, p-value equal to 0.10). The size of the grants matters too: Courses taught by instructors with a total average grant size in the top quartile have a 0.12 lower gap compared with instructors who never received a grant, and the gap progressively declines as the grant amount increases (panel b, column 1, significant at 5 percent). These findings suggest that public investments in academic research can yield additional private and social returns in the form of more up-to-date instruction, which – as we show next – is associated with better student outcomes.

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. 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.

6 The Education-Innovation Gap and Students' Outcomes

We have shown that 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 study whether these differences are related to students' outcomes. We focus on three outcomes: graduation rates, income, and intergenerational mobility. Graduation rates are from IPEDS and cover the years 1998 to 2018. Data on students' incomes ten years after graduation are from the College Scorecard, and cover students who graduated between 1998 and 2008. We complement this information with cross-sectional data on average and median incomes and the odds of reaching top income percentiles of all students who graduated from each school between 2002 and 2004, calculated by Chetty et al. (2019) using data from tax records. Chetty et al. (2019) also provide a measure of intergenerational mobility, defined as the probability that students with parental incomes in the bottom quintile of the distribution reach the top quintile during adulthood.

All these outcomes are measured at the school level, whereas the education-innovation gap is at

the syllabus level. To construct a school-level measure 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. Because outcome measures refer to students who complete undergraduate programs at each school, we construct θ_s using only undergraduate syllabi; our results are robust to the use of all syllabi. Appendix Figure AVI shows the distribution of θ_s; the standard deviation is 0.85, corresponding to a 5 percent change in the average syllabus.

In the remainder of this section, we present estimates of the parameter δ in the following equation:

$$Y_{st} = \delta\theta_s^z + X_{st}\gamma + \tau_t + \varepsilon_{st} \tag{4}$$

where Y_{st} is the outcome for students who graduated from school s in year t, θ_s^z the school fixed effect in equation (3) standardized to have mean zero and variance one, X_{st} is a vector of school observables, and τ_t are year fixed effects. We calculate bootstrapped standard errors, clustered at the level of the school, to account for the fact that θ_s^z is an estimated quantity.

The possible existence of unobservable attributes of schools and students, related to both the content of a school's courses and student outcomes, prevents us from interpreting the parameter δ as the causal effect of the gap on these outcomes. Nevertheless, we attempt to get as close as possible to a causal effect by accounting for a rich set of school observables from IPEDS, and we show how our estimates change when we control for them. We include seven groups of controls, including institutional characteristics (control, selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification); instructional characteristics (student-to-faculty ratio and the share of ladder faculty); financials (total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student); enrollment (share of undergraduate and graduate enrollment, share of white and minority students); selectivity (indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, indicators for schools not using either SAT or ACT in admission); major composition (share of students with majors in Arts and Humanities, Business, Health, Public

and Social Service, Social Sciences, STEM, and multi-disciplinary fields); and family background, measured as the natural logarithm of parental income. Panel a of Table 7 shows the unconditional correlations between each outcome and the school-level education-innovation gap (i.e., estimates of δ in equation (4)); panel b shows the same correlations controlling for these school characteristics.

6.1 Graduation Rates

Column 1 of Table 7 shows the relationship between the gap (measured in standard deviations) and graduation rates. An estimate of -0.05 in panel a, significant at 1 percent, indicates that a one-standard deviation decline in the gap (or a 10 percent change in the content of a syllabus) is associated with a 5 percentage point higher graduation rates. Compared with an average of 0.61, this corresponds to a 8 percent increase in graduation rates.

The estimate of δ declines as we control for observable school characteristics, indicating that part of this correlation can be explained by other differences across schools. However, it remains negative and significant at -0.007, indicating that that a one-standard deviation reduction in the gap is associated to a 1.1 percent increase in graduation rates (panel b, column 1, significant at 5 percent).

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, in columns 2-8 of Table 7 we study the relationship between the gap and various income statistics.

Column 2 shows estimates on the natural logarithm of mean student income from the College Scorecard. While imprecise, this estimate indicates that a one-standard deviation in the gap is associated with a 0.7 percent increase in income controlling for the full set of observables (panel b, p-value equal to 0.17). The College Scorecard also reports mean incomes for students with parental incomes in the bottom tercile of the distribution; for these students, the relationship is slightly larger at 0.8 percent (column 3, significant at 10 percent). Estimates are largely unchanged when we use median instead of mean income (column 4).

Information on mean student incomes at the school level is also reported by Chetty et al. (2019), calculated using tax records for a cross section of students. Unconditional estimates (which omit year effects due to the cross-sectional structure of the data) indicate that a one-standard deviation

in the gap is associated with a 7 percent increase in students' mean income (panel a, column 5, significant at 1 percent). This estimate is smaller, at 1.4 percent, when controlling for institutional characteristics (panel b, column 5, significant at 1 percent).

Lastly, in columns 6 through 8 of Table 7 we investigate the relationship between the gap and the probability that students' incomes reach the top echelons of the income distribution. Estimates with the full set of controls indicate that a one-standard deviation decline in the gap is associated with a 0.84 percentage-point increase in the probability of reaching the top 20 percent (2.2 percent, panel b, column 6, significant at 1 percent), a 0.53 percentage-point increase in the probability of reaching the top 10 percent (2.5 percent, column 7, significant at 5 percent), and a 0.31 percentage-point increase in the probability of reaching the top 5 percent (2.7 percent, column 8, significant at 10 percent). Taken together, these results indicate a positive relationship between the school-level education-innovation gap and students' average and top incomes.

6.3 Intergenerational Mobility

Using data from Chetty et al. (2019), in column 9 of Table 7 we also study the association between the gap and intergenerational mobility, defined as the probability that students born in families in the top income quintile reach the top quintile when they enter the labor market. The unconditional correlation between these two variables is equal to -0.0293, indicating that a one-standard deviation lower gap is associated with a 2.9 percentage-points increase in intergenerational mobility (9.9 percent, panel a, column 9, significant at 1 percent). This correlation, however, becomes smaller and indistinguishable from zero when we control for school observables, reaching -0.0047 when we include the full set of controls (column 9, panel b, p-value equal to 0.15).

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, such as graduation rates and incomes after graduation. 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 the inclusion of controls for a large set of school and student characteristics, indicating that these correlations are unlikely to be driven by cross-school differences in spending, selectivity, major composition, or parental background. Thee findings point to a potentially important role for up-to-date instruction on the outcomes of students as they exit college and enter the

labor market.

7 Alternative Measures of Course Novelty

In spite of its desirable properties, our measure of the education-innovation gap has some limitations. For example, the gap penalizes courses that include old *and* new content, relative to courses that include exactly the same new content but no old content. Being devised to measure the "average" age of content, the gap is also unable to distinguish courses with extremely novel content among those with the same gap. Lastly, the gap only captures the novelty of academic content. A course with relatively old academic content, though, could still be novel in other dimensions, for example by teaching skills in high demand in the labor market.

In this section, we probe the robustness of our results using alternative measures for the novelty of a course's content, aimed at (i) capturing the presence of new content regardless of older one; (ii) capturing the presence of extremely new content; and (iii) capturing novelty of the skills the course develops, rather than academic content. We briefly describe the results here; more detail can be found in the Online Appendix.

7.1 Presence of New Content

The education-innovation gap measures the presence, in a syllabus, of new content relative to older one. Consider two syllabi which both cover the same frontier research in a given field; the first syllabus is shorter and only contains this new content, while the second one is longer also contains older one. Our measure would assign a lower gap to the first syllabus compared to the second, even if both do an equal job in terms of covering frontier knowledge. To address this limitation of the education-innovation gap, we construct an alternative metric which measures the *share of old knowledge* of each syllabus, defined as one minus the ratio between the number of "new words" in each syllabus (defined as knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$) and the number of all new words. The correlation between the share of old knowledge and the education-innovation gap is 0.22 (Figure 9, panel a), and all our results hold if we use the former as an alternative measure of novelty of a syllabus's content (see Appendix Figure AVII, lighter series, for the school-level analysis; Appendix Tables AII and AIII for the analysis on the instructors; and Appendix Table AIV for the relationship with student outcomes).

7.2 Right Tail of Academic Novelty

Our education-innovation gap captures the "average" novelty of a syllabus. It is possible for two syllabi to have the same gap when one of them only covers content from five years prior while the other covers mostly material from fifteen years prior, but also a small amount of material from the previous year. To construct a measure that captures the presence of "extremely" new material in a syllabus, we proceed as follows. First, we draw 100 "sub-syllabi" from each syllabus, defined as subsets of 20 percent of the syllabus's words, and calculate the corresponding education-innovation gap. We then recalculate the average gap among all sub-syllabi in the bottom 5 percent of the gap distribution of a given syllabus.[26] We refer to this as a "tail measure" of novelty.

The tail measure is positively correlated with the education-innovation gap, with a correlation of 0.67. All our results hold when using the tail measure as a metric for syllabus novelty (see Appendix Figure AVII, darker series, for the school-level analysis; Appendix Tables AV and AVI for the analysis on the instructors; and Appendix Table AVII for the relationship with student outcomes).

7.3 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).

The measure of soft-skills intensity is negatively correlated with the education-innovation gap (with a correlation of -0.14, Figure 9, panel c). The cross-school differences in the skill intensity of the courses display the same patterns we found for the education-innovation gap: The prevalence of soft skills increases with school selectivity (Figure AVIII, panel a), it is larger for schools where the median parental income is in the top portion of the distribution (panel c), and for those enrolling

[26] Our results are robust to the use of the top 10 and one percent.

a higher share of minority students (panel d). Soft skills are also more prevalent for courses taught by the most research-productive instructors (Table AVIII).

In closing, we examine the relationship between courses' soft-skills intensity and student outcomes. Controlling for the full set of school observables used in Table AX, a one-standard deviation increase in the soft-skills intensity of a school's courses is associated to a 1.2 percentage-point increase in graduation rates (2 percent, column 1, significant at 1 percent); a 1.7 percent higher mean income (column 2, significant at 1 percent); and a 1.2 percent higher chances of reaching the top income quintile for students with parental income in the bottom quintile (18 percent, column 9, significant at 1 percent).

Taken together, these findings indicate that the variation across and within schools in the extent to which courses are up-to-date, and its relationship with student outcomes, are not unique to academic "novelty." They also hold when we capture novelty with the skills that students are most likely to acquire during a course. We interpret this as additional evidence for the importance of accounting for differences in content across courses when considering the heterogeneity of educational experiences of students across different schools and their consequences for short- and long-run outcomes.

8 Conclusion

This paper has studied the diffusion of frontier knowledge through higher education with an in-depth analysis of the content of college and university courses. Our approach centers around a new measure, the "education-innovation gap," defined as the textual similarity between syllabi of courses taught in colleges and universities and the frontier knowledge published in academic journals. Using text analysis techniques, we estimate this measure comparing the text of 1.7 million course syllabi with that of 20 million academic publications.

Using our measure, we document a set of new findings about the dissemination of new knowledge in US higher-education institutions. First, a significant amount of variation exists in the extent to which this knowledge is offered, both across and within schools. 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 and incomes after graduation. Taken together, our results suggest that the education-innovation gap

can be an important measure to study how frontier knowledge is produced and disseminated.

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.

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, *The Quarterly Journal of Economics* 132, 1593–1640.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

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.

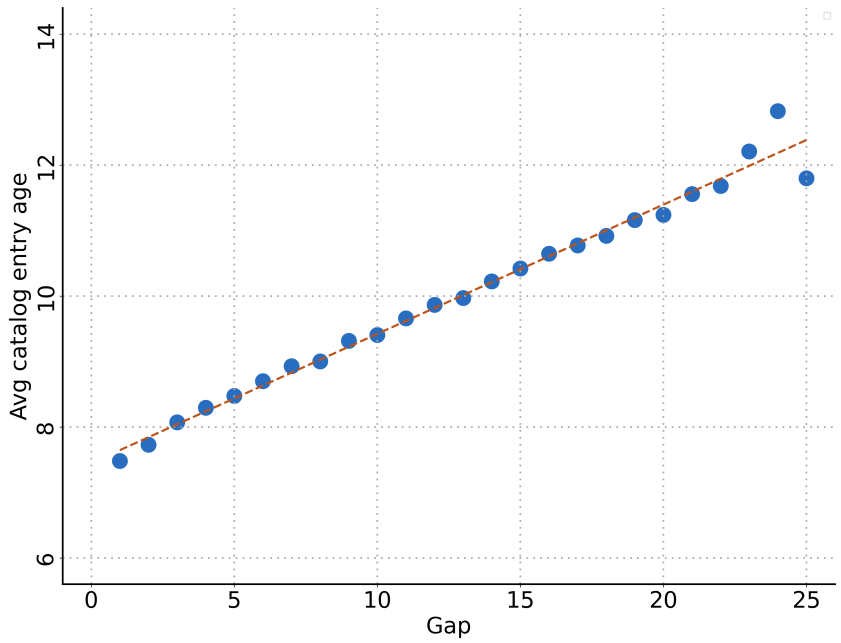
Rockoff, Jonah E, 2004, The impact of individual teachers on student achievement: Evidence from panel data, *American Economic Review* 94, 247–252.

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.

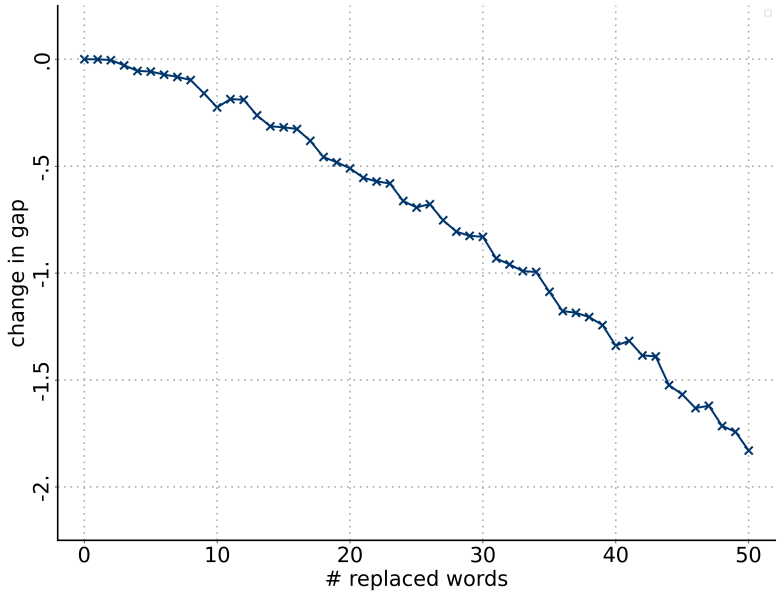
Figure 1: Validating the Education-Innovation Gap Measure With Syllabi References



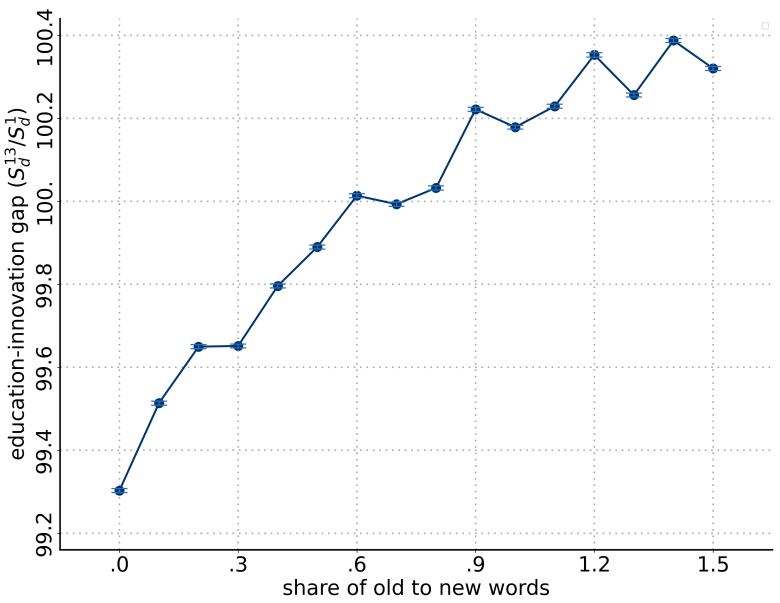
Note: This figure shows the correlation between the gap and the reference age of each syllabus. The reference age is defined as the average difference between the year of the syllabus and the year of each reference listed in the syllabus as a required or recommended reading. We divide syllabi in 25 equally-sized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

Figure 2: Validating the Education-Innovation Gap with Two Simulations

(a) Gap for Fictitious Syllabi, By Ratio of Old to New Words

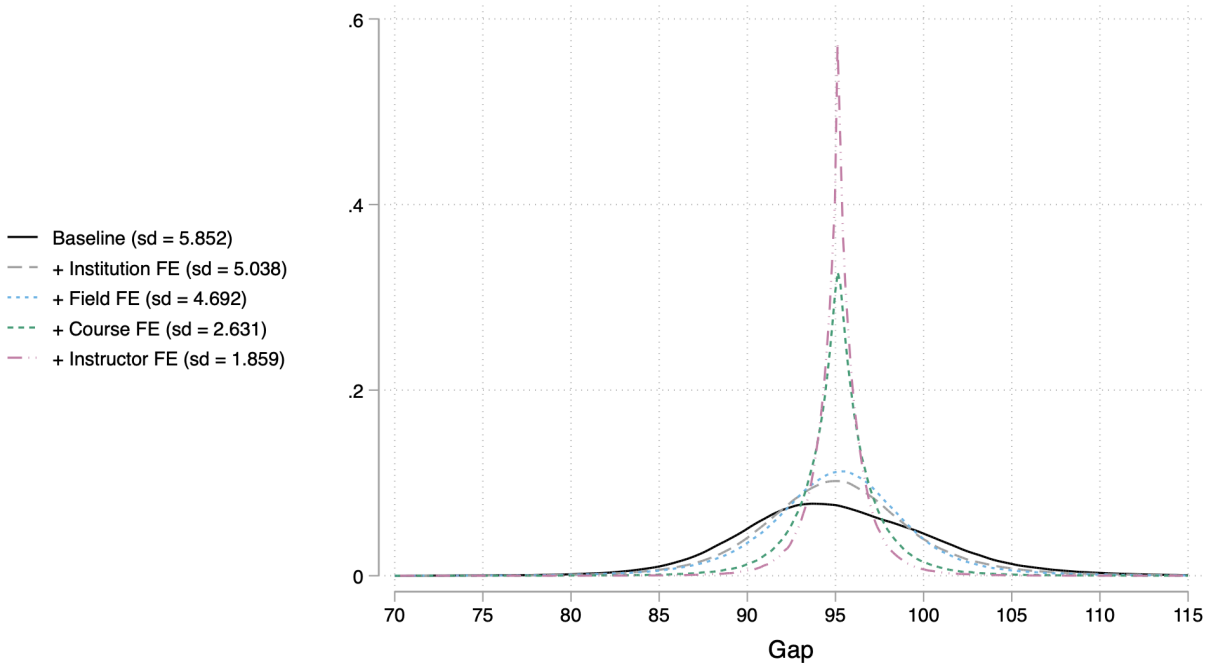


(b) Gap As We Replace Syllabi's Old Words Newer Words



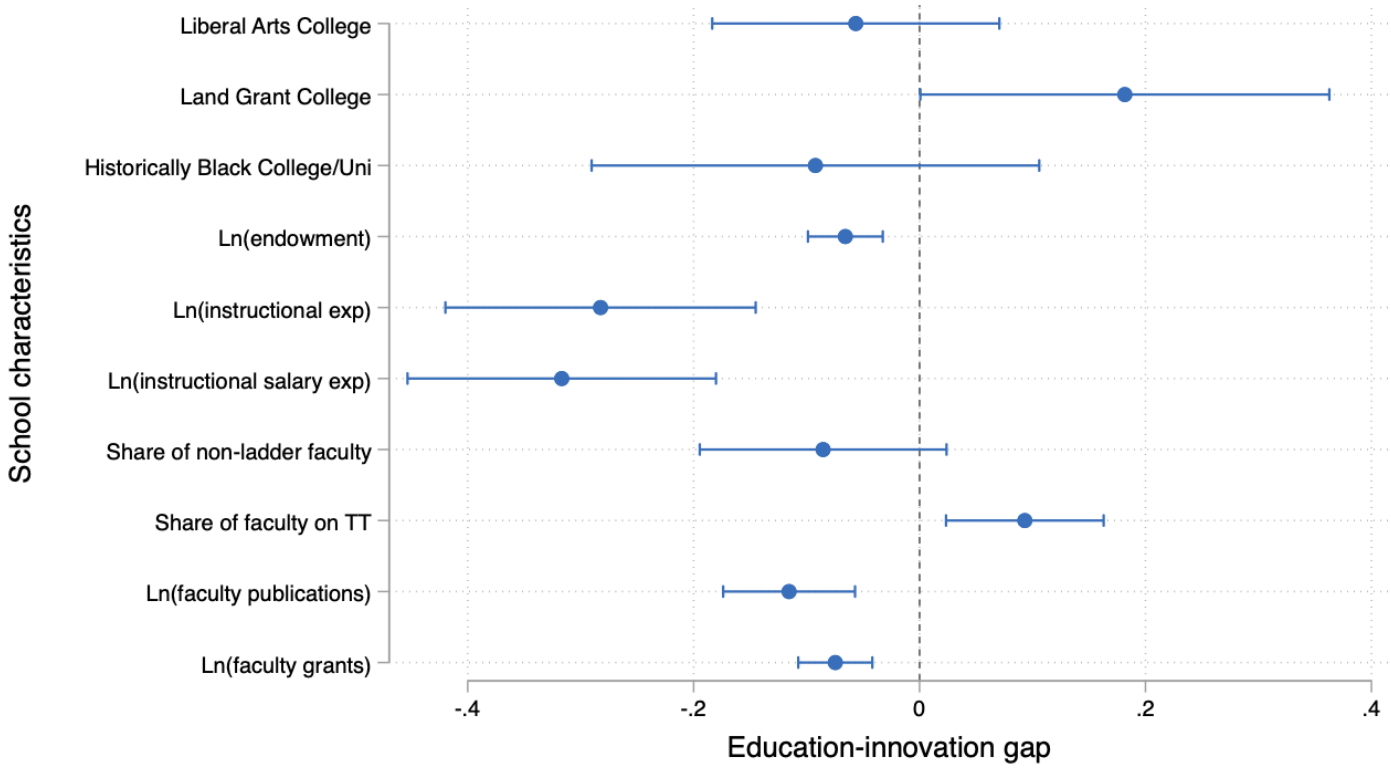
Note: Panel a shows the relationship between the ratio of "old" to "new" words (defined, respectively, as those in the top 5 percent in terms of frequency in the old publication corpus between t-15 and t-13 or in the old publication corpus between t-15 and t-12 but not in the new publication corpus between t-3 and t-1; and those in the top 5 percent in terms of frequency in the new publication corpus between t-3 and t-1 or in the new publication corpus between t-3 and t-1 but not in the old publication corpus between t-15 and t-13) and the education-innovation gap for a set of 1.7 million fictitious syllabi. Panel b shows how the education-innovation gap changes as we replace older words with newer ones. 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.

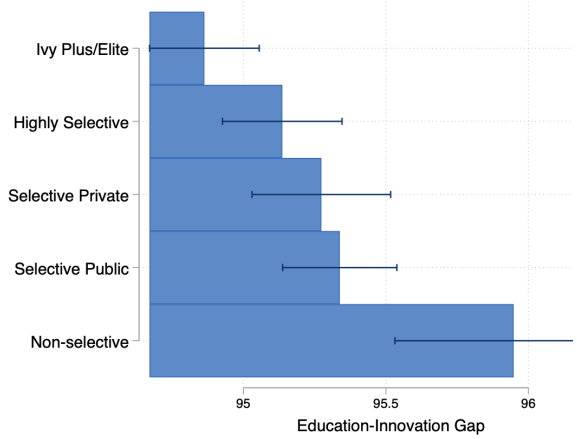
Figure 4: The Education-Innovation Gap and School Characteristics



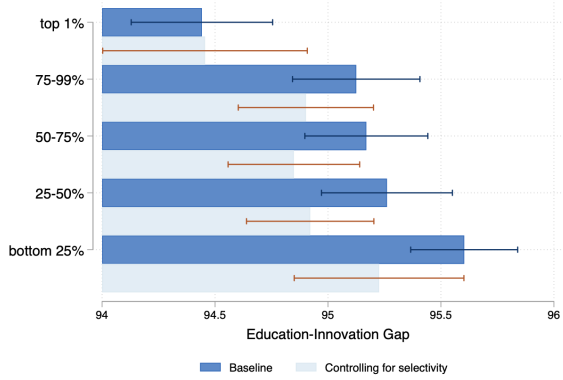
Notes: The figure shows point estimates and 95-percent confidence intervals of the relationship between the gap and various school-level characteristics, controlling for field-by course level-by-year fixed effects (corresponding to the parameter β in equation (2). The gap is defined as the ratio between 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. Expenditures and faculty composition refer to the year 2018 and are from IPEDS. Variables referring to faculty shares are standardized to have mean zero and standard deviation one. Estimates are obtained pooling syllabi data for the years 1998 to 2018. Standard errors are clustered at the school level.

Figure 5: The Education-Innovation Gap, School Selectivity, and Student Composition

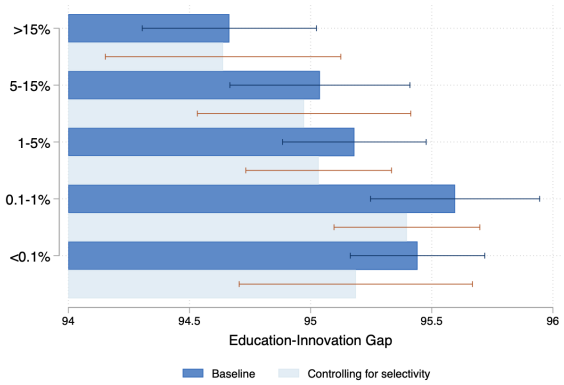
(a) By school tier



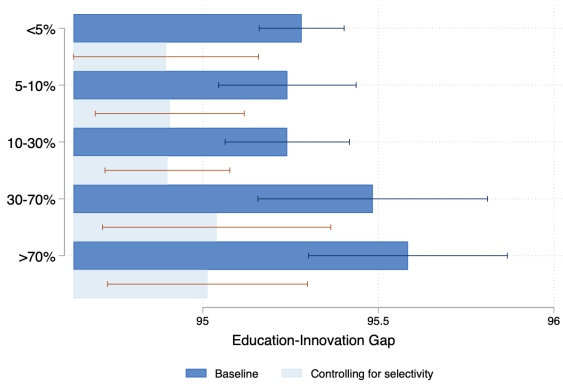
(b) By percentile of parental income



(c) By % of parents in top income percentile



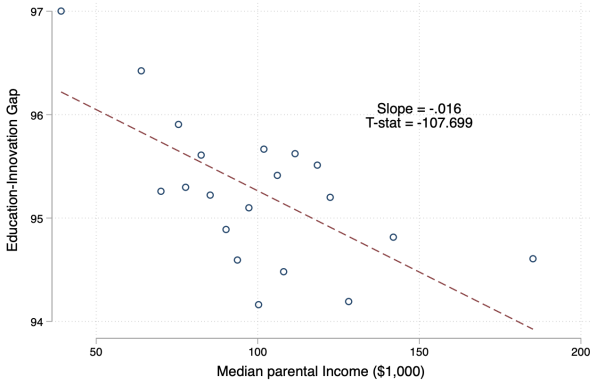
(d) By % of minority students (Black/Hispanic)



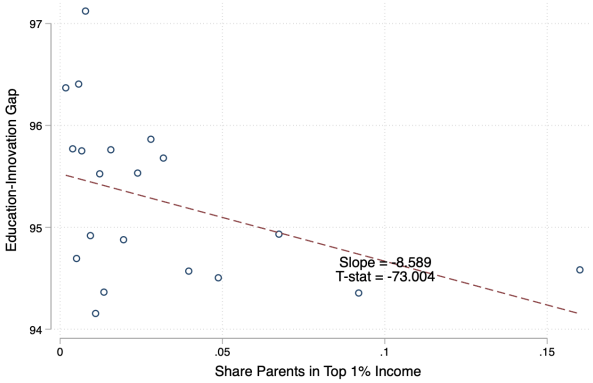
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by school tier (panel a), percentile of median parental income in the school (panel b), share students with parents in the top income percentile in the school (panel c), and share of students 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-d, the lighter series also control for selectivity tiers. Standard errors are clustered at the school level.

Figure 6: The Education-Innovation Gap And Student Demographics

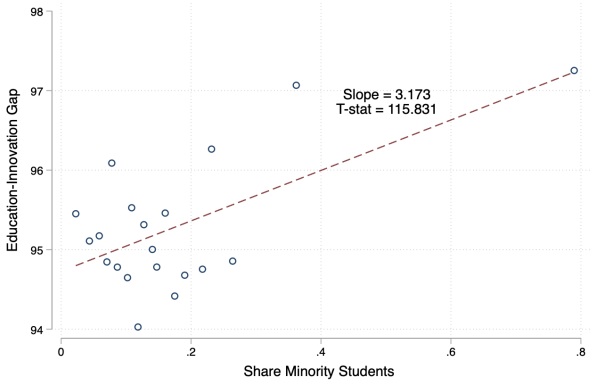
(a) Median parental income



(b) % of parents in top income percentile

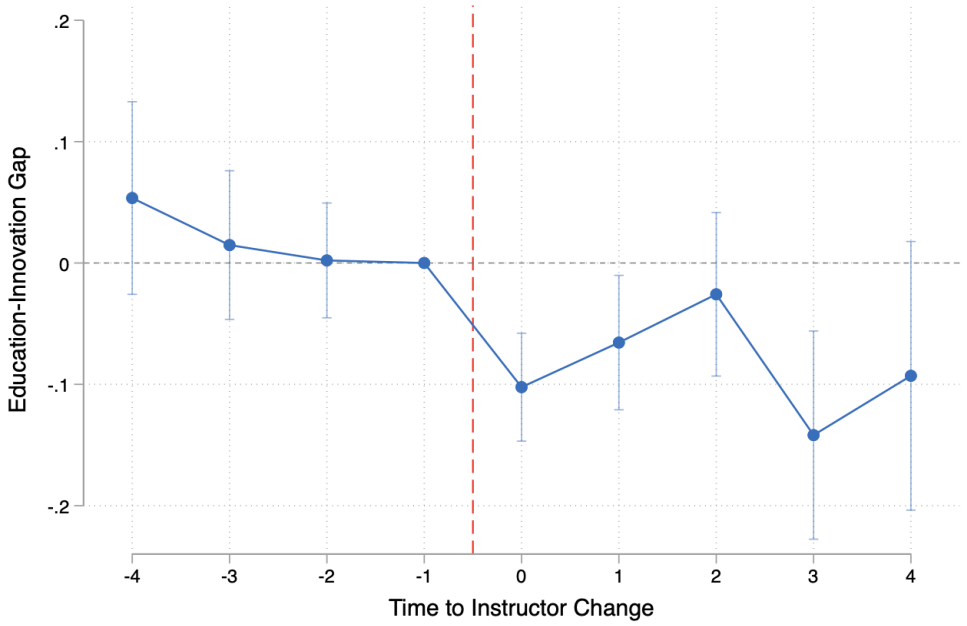


(c) % of minority students (Black/Hispanic)



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)).

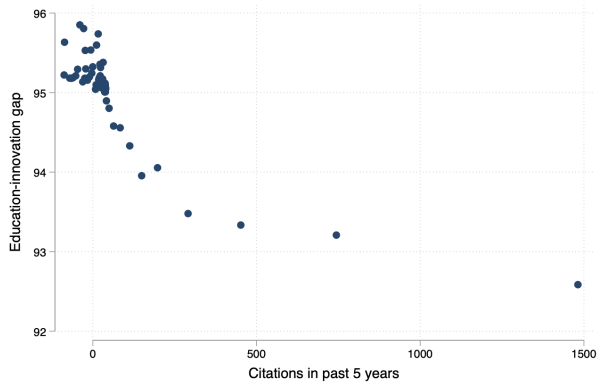
Figure 7: Education-Innovation Gap Around The Time of An Instructor Change



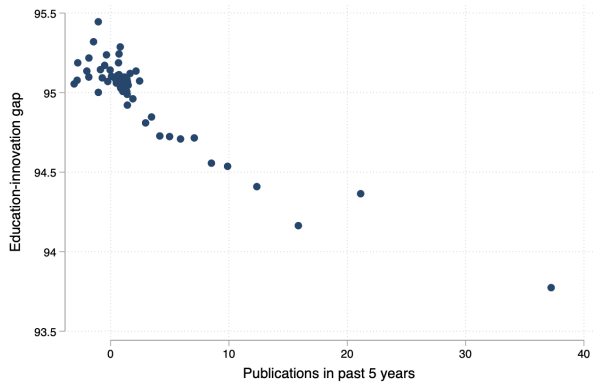
Notes: Estimates and standard deviations of the parameters δ_k, corresponding to an event study of the gap around an instructor change and specified in equation (2). Standard errors clustered at the course level.

Figure 8: Instructors' Research Productivity and Funding and The Education-Innovation Gap

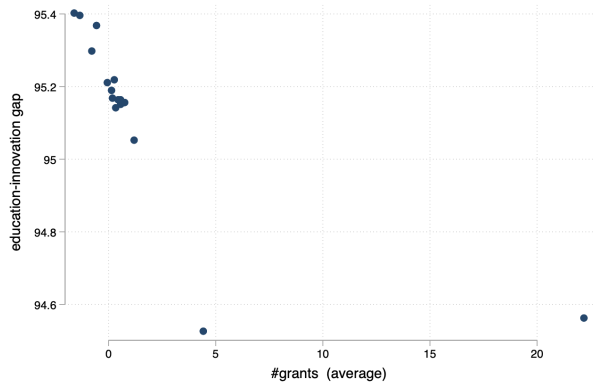
(a) Citations, Last 5 Years



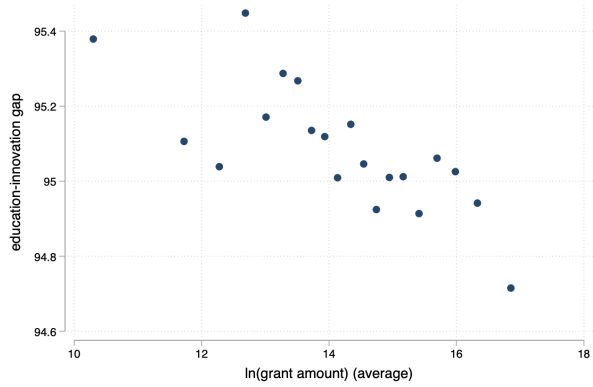
(b) Publications, Last 5 Years



(c) Total #Grants



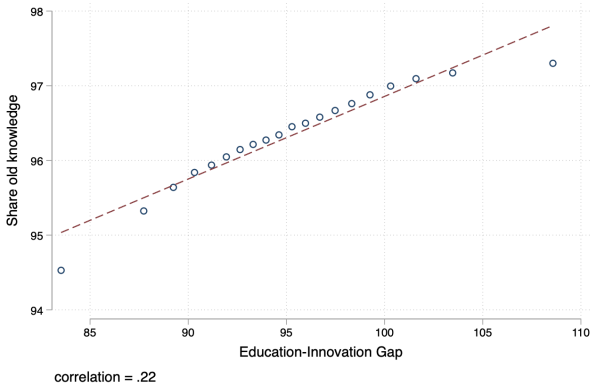
(d) Average Grant Amount if #Grants>0 (Logs)



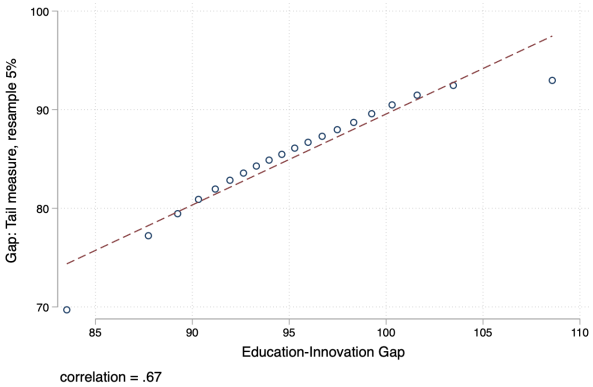
Notes: Binned scatterplot of the gap (vertical axis) and measures of research productivity and funding (horizontal axis): number of citations in the last 5 years (panel a), number of article publications in the last 5 years (panel b), total number of NSF and NIH grants ever received, and average size of grants received for instructors with at least one grant (in logs). All relationships are obtained controlling for field fixed effects.

Figure 9: The Education-Innovation Gap and Alternative Measures of Novelty: Binned Scatterplots

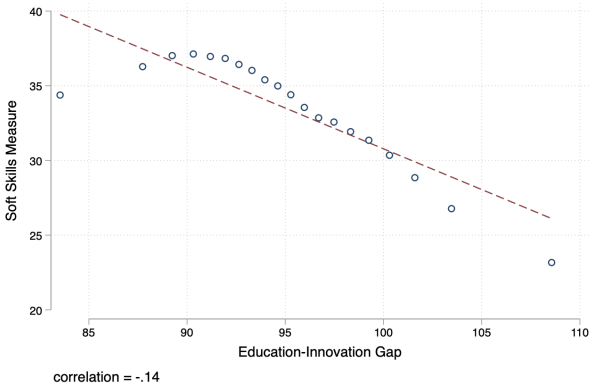
(a) Share of old knowledge



(b) Tail measure



(c) Soft-skills intensity



Notes: Binned scatterplots of the education-innovation gap and three alternative measures of novelty of each syllabus: a measure of old knowledge, defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$, panel a); a "tail measure," calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution (panel b); and a measure of soft skills, defined as the share of words in the assignment portion of a syllabus which refer to soft skills (panel c).

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%
Ever Published?	1,706,319	0.329	0.470	0	0	1
# Publications per year	682,286	1.455	1.696	1	1	1.286
# Publications, last 5 years	682,286	4.575	12.238	0	0	3
# Citations per year	682,286	24.146	82.553	0	1.333	15.286
# Citations, last 5 years	682,286	103.961	621.203	0	0	17
Ever Grant?	1,706,319	0.112	0.315	0	0	0
# Grants	190,738	13.828	26.294	3	6	13
Grant amount ($1,000)	190,738	5,223.871	20,243.181	467.297	1,464.919	4,385.383

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

	count	mean	std	25%	50%	75%
Median parental income ($1,000)	767	97.917	31.054	78	93.5	109.850
Share parents w/income in top 1%	767	0.03	0.041	0.006	0.013	0.032
Share minority students	760	0.221	0.166	0.116	0.166	0.267
Graduation rates (2012–13 cohort)	758	0.614	0.188	0.473	0.616	0.764
Income (2003–04, 2004–05 cohorts)	762	45,035.433	10,235.2	38,200	43,300	49,775
Intergenerational mobility	767	0.294	0.138	0.183	0.28	0.375
Admission rate	715	0.642	0.218	0.534	0.683	0.799
SAT score	684	1,104.395	130.493	1,011.75	1,079.5	1,181.5

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	.
+ School 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, Instructor Changes, and Fit Between Instructors' Research and Course Content

Panel (a): Instructor change	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
After change	-0.0924***	-0.0972	-0.1283***	-0.0875**	-0.0086	-0.0741	-0.0719	-0.1133***
	(0.0244)	(0.0696)	(0.0481)	(0.0436)	(0.0417)	(0.0455)	(0.0450)	(0.0374)
N (Course x year)	379538	35577	97464	134095	94557	125494	112143	137843
# Courses	126369	11551	33234	40137	31581	43536	35385	46244
Panel (b): Instructor's fit w/course	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
fit w/top course (sd)	-0.0877**	0.1480	0.0017	-0.0836	-0.0849	-0.0637	-0.1428*	-0.0611
	(0.0398)	(0.1001)	(0.1723)	(0.0608)	(0.0656)	(0.0832)	(0.0790)	(0.0558)
N (Course x year)	54591	3199	2218	33119	12587	16743	16224	21139
# Courses	17077	1011	761	10267	3909	5208	4833	6883
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is the education-innovation gap. In panel (a), the dependent variable is an indicator for years following an instructor change, for courses with only one instructor and at most two instructor changes over the observed time period. In panel (b), the dependent variable is a measure of fit between the instructor's research and the content of the course, defined as the cosine similarity between the set of all syllabi from the same course across schools and the instructor's research in the previous 5 years.. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

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

Panel a): #publications	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	-0.0218	0.0515	-0.0830**	0.0727*	-0.0692**	-0.0163	-0.0244	-0.0281
	(0.0175)	(0.0479)	(0.0338)	(0.0421)	(0.0292)	(0.0293)	(0.0319)	(0.0299)
2nd quartile	-0.0523*	0.0322	-0.0454	-0.0245	-0.0485	-0.0528	-0.0048	-0.0844*
	(0.0273)	(0.0677)	(0.0708)	(0.0548)	(0.0428)	(0.0479)	(0.0487)	(0.0443)
3rd quartile	-0.0023	0.0634	-0.0799	0.1251**	-0.1083**	0.0316	0.0127	-0.0417
	(0.0290)	(0.0758)	(0.0812)	(0.0536)	(0.0456)	(0.0539)	(0.0539)	(0.0440)
4th quartile	-0.0969**	0.1003	-0.1938	-0.0311	-0.1869***	-0.0112	-0.1387*	-0.1437**
	(0.0393)	(0.1148)	(0.1524)	(0.0617)	(0.0652)	(0.0783)	(0.0726)	(0.0571)
Panel b): #citations	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	0.0237	0.0047	0.0305	0.1251**	-0.0358	0.0351	0.0240	0.0141
	(0.0237)	(0.0638)	(0.0521)	(0.0515)	(0.0367)	(0.0404)	(0.0428)	(0.0398)
2nd quartile	0.0021	0.0253	-0.0896	0.1469***	-0.0617	0.0504	0.0062	-0.0417
	(0.0274)	(0.0662)	(0.0757)	(0.0535)	(0.0421)	(0.0505)	(0.0489)	(0.0429)
3rd quartile	-0.0484	-0.0560	-0.1030	0.0256	-0.1163**	0.0010	-0.0349	-0.0994**
	(0.0327)	(0.0855)	(0.1031)	(0.0579)	(0.0501)	(0.0618)	(0.0605)	(0.0489)
4th quartile	-0.0960**	0.0974		-0.0662	-0.1395**	-0.0661	-0.0935	-0.1436**
	(0.0425)	(0.1074)		(0.0664)	(0.0683)	(0.0835)	(0.0785)	(0.0619)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is the education-innovation gap; the independent variables are indicators for quartiles of the number of publications (panel (a)) and citations (panel (b)) of a course's instructors in the previous five years. The omitted category are courses with instructors with no publications or citations. For courses with more than one instructor, we consider the mean number of publications and citations across all instructors. All specifications control for course and field-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 Instructor Research Resources: Grant numbers and amount

Panel (a): #grants	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	-0.0488*	-0.0672	-0.0793	-0.0070	-0.0936**	-0.0775*	-0.0639	0.0072
	(0.0272)	(0.0811)	(0.0563)	(0.0503)	(0.0440)	(0.0430)	(0.0516)	(0.0471)
2nd quartile	-0.0550	0.0790	-0.0906	-0.1233**	-0.0529	-0.0184	-0.1072	-0.0609
	(0.0377)	(0.1227)	(0.0868)	(0.0613)	(0.0665)	(0.0610)	(0.0706)	(0.0645)
3rd quartile	-0.0372	-0.1093	-0.0160	-0.0430	-0.0698	0.0232	-0.0494	-0.1007
	(0.0448)	(0.1416)	(0.0921)	(0.0706)	(0.0867)	(0.0750)	(0.0813)	(0.0760)
4th quartile	-0.0819	0.0102	-0.0551	-0.1188	-0.0790	-0.0601	-0.0119	-0.1525*
	(0.0498)	(0.1447)	(0.1136)	(0.0825)	(0.0949)	(0.0853)	(0.0919)	(0.0816)
Panel (b): grant amount ($)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0133	-0.0827	-0.1393	-0.0045	-0.0606	-0.0956*	0.0303	0.0834
	(0.0357)	(0.1142)	(0.1138)	(0.0490)	(0.0689)	(0.0540)	(0.0668)	(0.0677)
2nd quartile	-0.0635*	-0.1335	-0.0823	-0.0723	-0.0639	-0.0401	-0.0426	-0.1075*
	(0.0342)	(0.1003)	(0.0625)	(0.0727)	(0.0548)	(0.0554)	(0.0656)	(0.0571)
3rd quartile	-0.0506	0.0360	-0.0016	-0.1172*	-0.0090	-0.0032	-0.0663	-0.0826
	(0.0390)	(0.1315)	(0.0806)	(0.0697)	(0.0655)	(0.0649)	(0.0733)	(0.0648)
4th quartile	-0.1235**	0.1087	-0.0960	-0.1461*	-0.2057**	-0.0513	-0.1956**	-0.1435*
	(0.0489)	(0.1217)	(0.1087)	(0.0820)	(0.0928)	(0.0847)	(0.0879)	(0.0809)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is the education-innovation gap; the independent variables are indicators for quartiles of the number of grants (panel (a)) and total grant amount (panel (b)) ever obtained by a course's instructors. The omitted category are courses with instructors with no grants. For courses with more than one instructor, we consider the mean number and amount of grants. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 7: The Education-Innovation Gap and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2019)				
Panel (a): no controls	Grad rate (1)	Mean (2)	$P_y \leq 33$ pctile (3)	Median (4)	Mean (5)	P(top 20%) (6)	P(top 10%) (7)	P(top 5%) (8)	P(top 20%| $P_y \leq 20$ pctile) (9)
Gap (sd)	-0.0513***	-0.0555***	-0.0645***	-0.0512***	-0.0722***	-0.0333***	-0.0265***	-0.0187***	-0.0293***
	(0.0070)	(0.0099)	(0.0102)	(0.0086)	(0.0123)	(0.0054)	(0.0050)	(0.0036)	(0.0054)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): with controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap (sd)	-0.0073**	-0.0067	-0.0083*	-0.0090*	-0.0137***	-0.0084***	-0.0053**	-0.0031*	-0.0047
	(0.0030)	(0.0044)	(0.0049)	(0.0048)	(0.0050)	(0.0028)	(0.0022)	(0.0016)	(0.0030)
Mean dep. var.	0.5816	–	–	–	–	0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

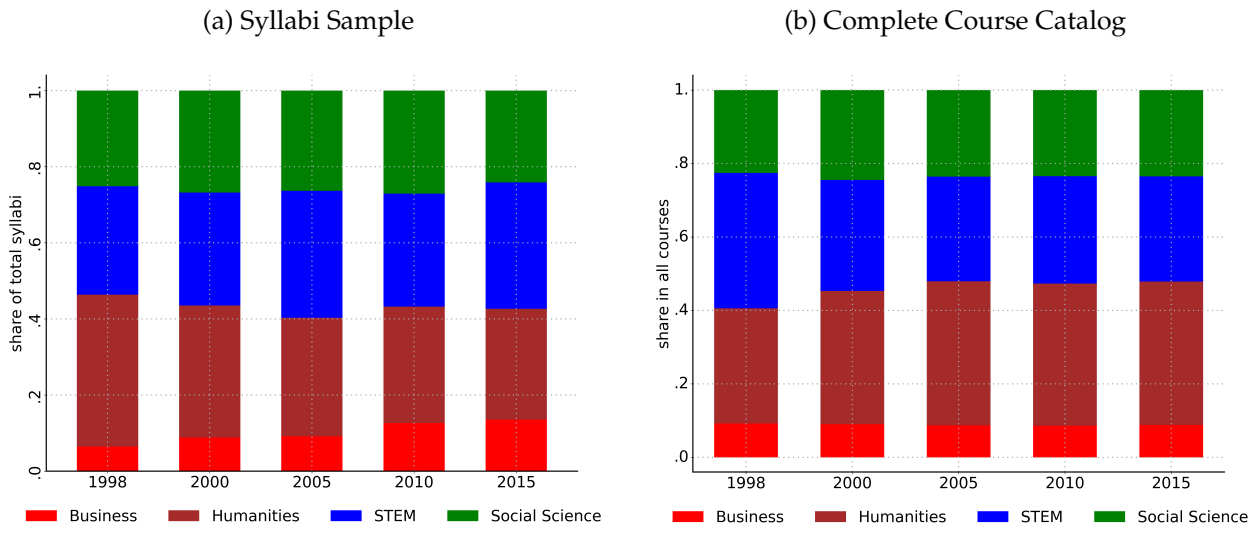
Note: OLS estimates of the coefficient δ in equation (4). The variable *Gap (sd)* is a school-level education-innovation gap (estimated as $\theta_{s(i)}$ in equation (3)), standardized to have mean zero and variance one. The dependent variable are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2019), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2019), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Appendix

For online publication only

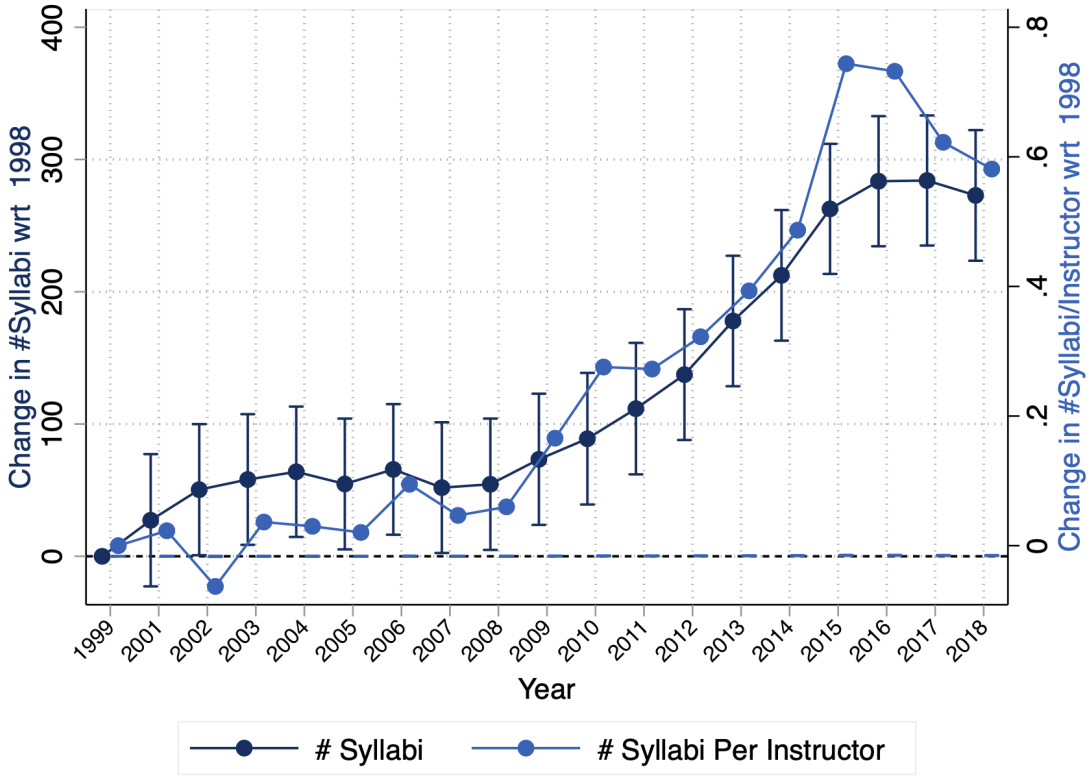
Additional Tables and Figures

Figure AI: Stable Field Coverage of the Syllabi Data



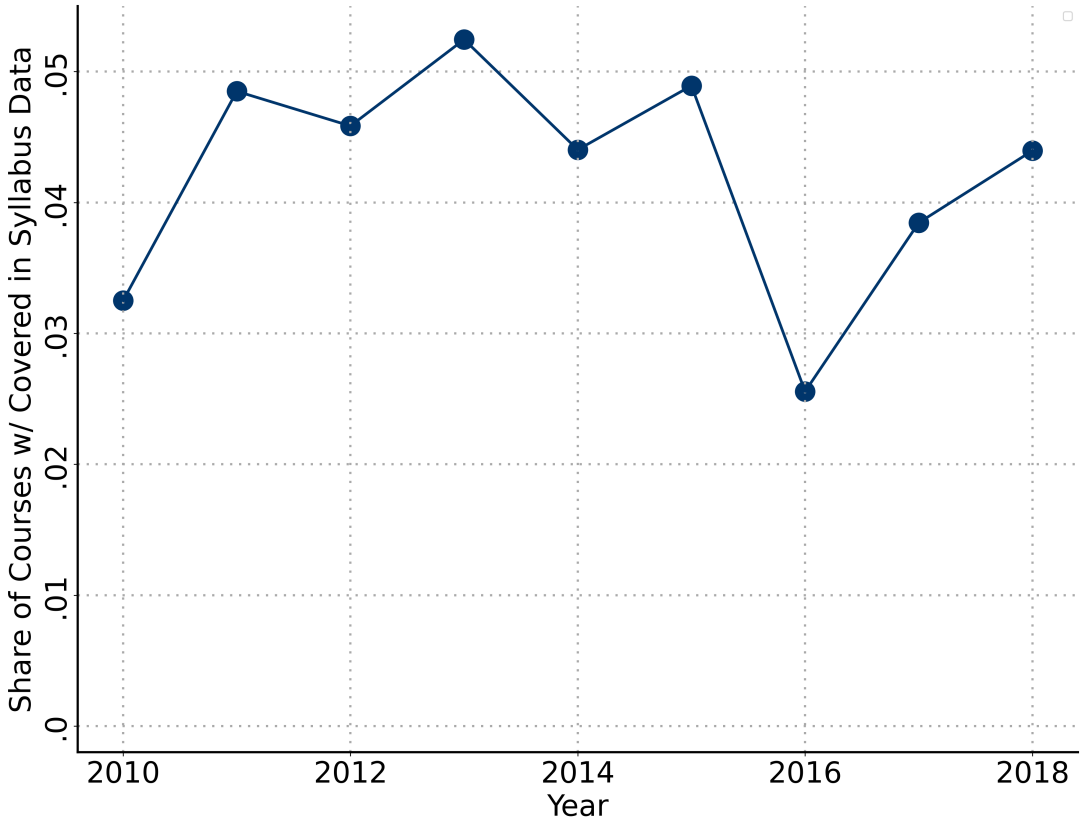
Note: Syllabi field composition by five-year periods. Panel (a) is for the syllabi sample. Panel (b) is for all courses collected from course catalog.

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.

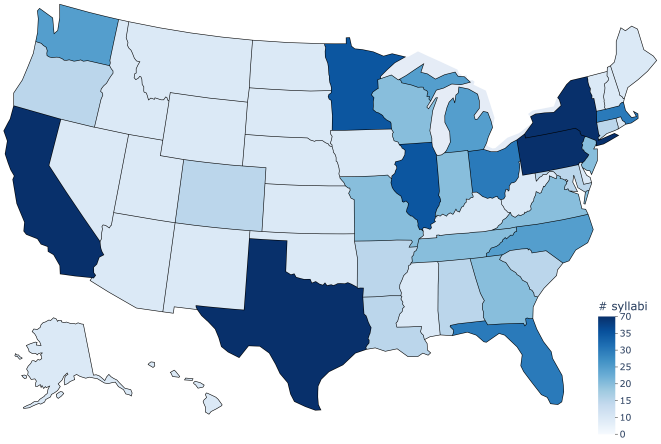
Figure AIII: Selection Into the OSP: Share of Covered Syllabi, Catalogue Data



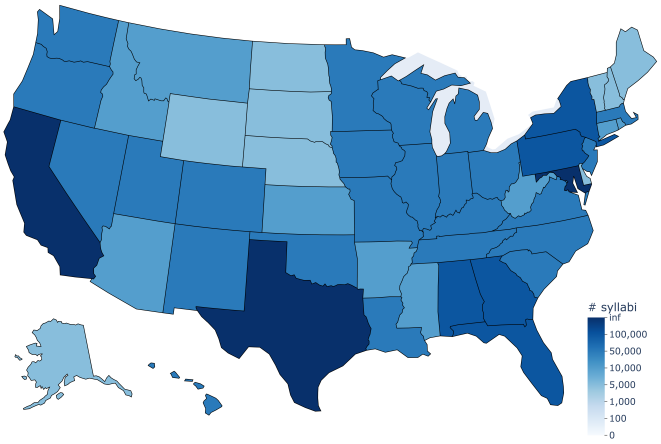
Note: Share of syllabi from the full catalogue of 161 selected institutions that are included in our sample. Catalogue data are collected from university archives.

Figure AIV: Syllabi Across The United States

Panel a) Number of Institutions in Each State

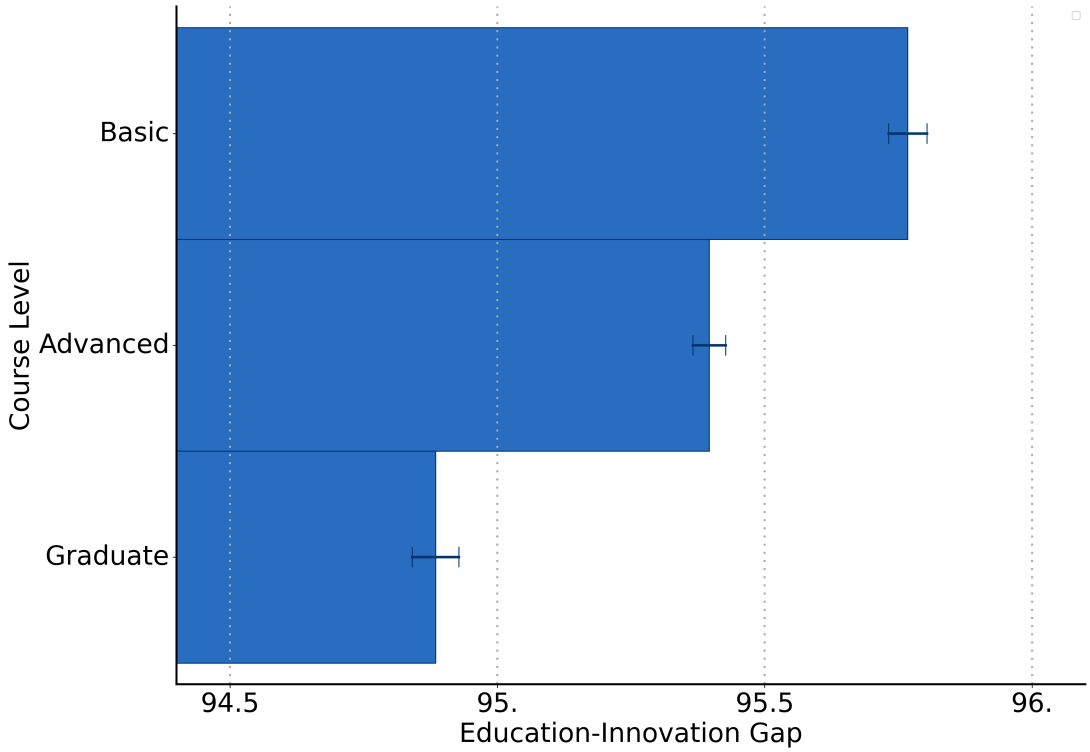


Panel b) Number of Syllabi in Each State



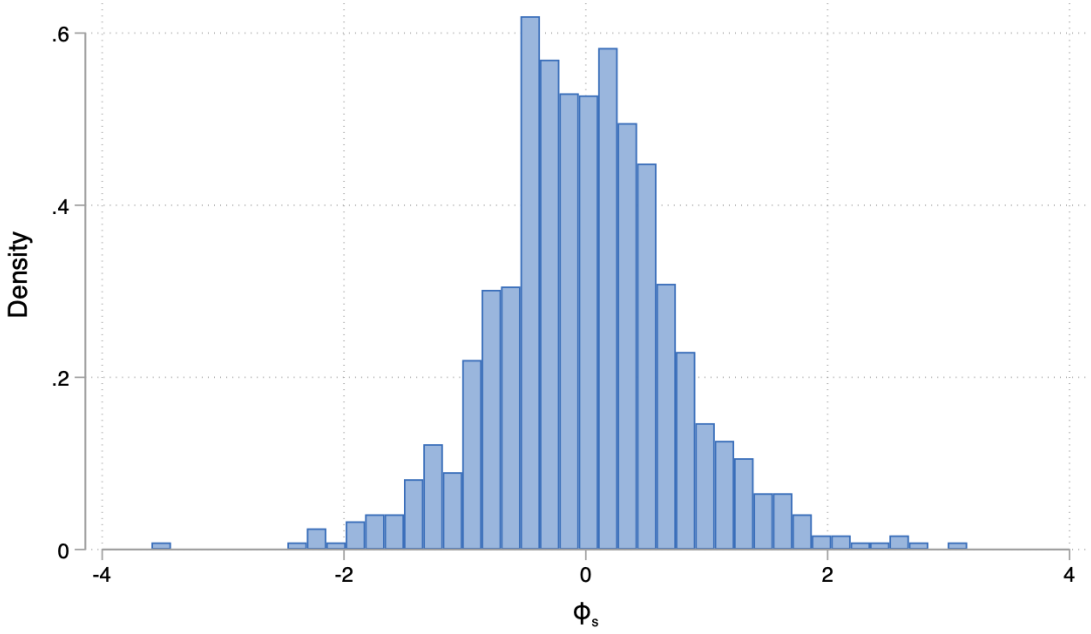
Note: The map plots the number of IPEDS institution (top panel) and the number of syllabi (bottom panel) from each state.

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.

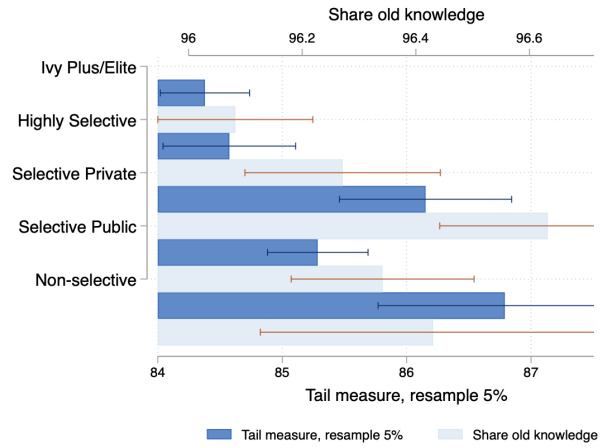
Figure AVI: Distribution of School-Level Gap



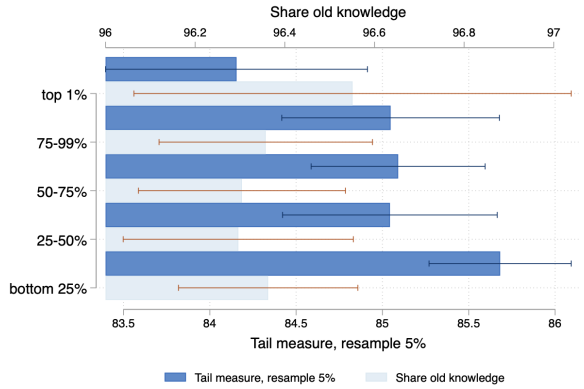
Note: Distribution of ϕ_s, the school-level component of the gap, corresponding to $\theta_{s(i)}$ in equation (3).

Figure AVII: Share of Old Knowledge and Tail Measure, By School Characteristics

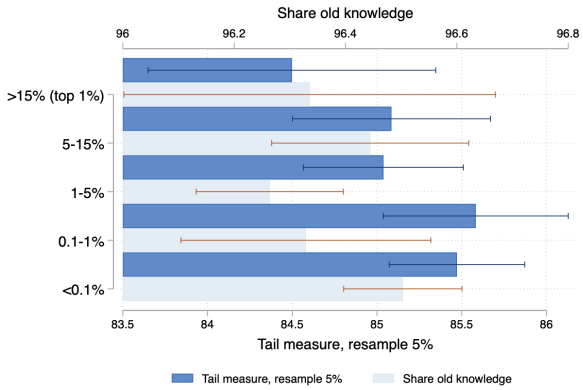
(a) By school tier



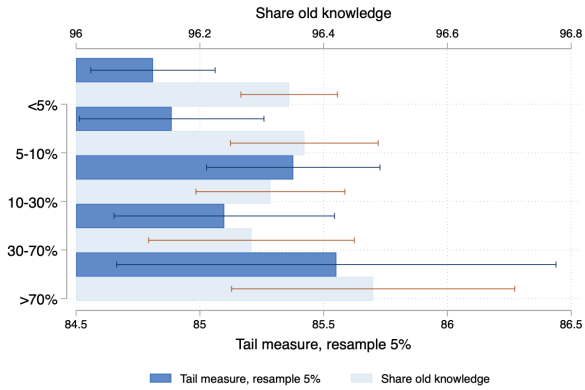
(b) By percentile of parental income



(c) By % of parents in top income percentile



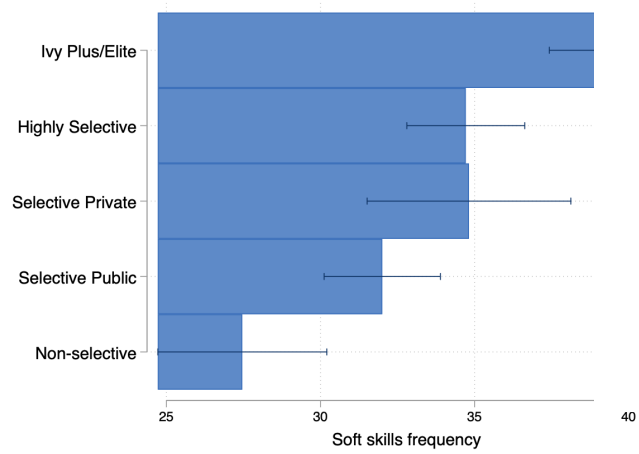
(d) By % of minority students (Black/Hispanic)



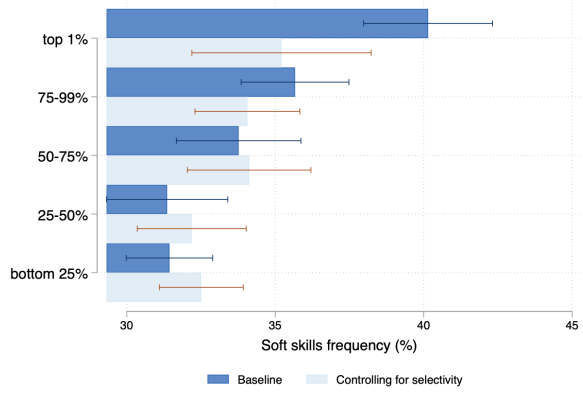
Notes: The figure shows averages and 95-percent confidence intervals of two alternative measures of the education-innovation gap between syllabi and publications, by school tier (panel a), percentile of median parental income in the school (panel b), share students with parents in the top income percentile in the school (panel c), and share of students who are either Black or Hispanic (panel d). The "share of old knowledge" is defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$). The "tail measure" is calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution. 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-d, we also control for selectivity tiers. Standard errors are clustered at the school level.

Figure AVIII: Soft Skills, By School Characteristics

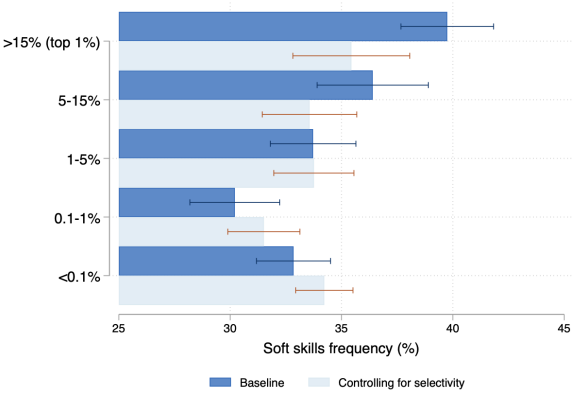
(a) By school tier



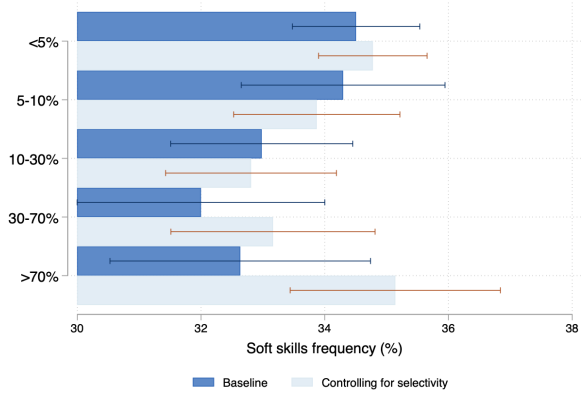
(b) By percentile of parental income



(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 a measure of soft skills, defined as the share of words in the assignment portion of a syllabus which refer to soft skills. The measure is shown 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). 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. The lighter series in panels b-d also controls for selectivity tiers. Standard errors are clustered at the school level.

Table AI: The Education-Innovation Gap By School Characteristics

Panel (a): selectivity	Business (1)	Humanities (2)	STEM (3)	Social Science (4)	Basic (5)	Advanced (6)	Graduate (7)	
Ivy Plus/Elite	-1.6865***	-0.1037	-1.3885***	-1.2059***	-0.9340***	-0.7020***	-0.9673***	
	(0.2807)	(0.4462)	(0.3633)	(0.2193)	(0.2104)	(0.1734)	(0.3459)	
Highly Selective/Selective	-0.9985***	0.3014	-1.3265***	-0.7073***	-0.5042**	-0.3943**	-0.7744***	
	(0.2472)	(0.4444)	(0.2549)	(0.2169)	(0.2055)	(0.1611)	(0.2385)	
Observations	173269	466728	486332	406602	670227	468913	567104	
Panel (b): median parent income	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
top 1%	-0.4556	-0.4582	-1.0401***	0.0055	-0.7735***	-0.5475**	-0.3439	
	(0.5547)	(0.3213)	(0.3819)	(0.1943)	(0.2428)	(0.2192)	(0.2591)	
75-99%	-0.0750	-0.2121*	-0.2942*	0.0524	-0.2693***	-0.0448	-0.0978	
	(0.1559)	(0.1175)	(0.1565)	(0.1284)	(0.1009)	(0.0919)	(0.0975)	
50-75%	-0.1807	-0.2627*	-0.2818**	-0.1268	-0.3643***	-0.0391	-0.1287	
	(0.1750)	(0.1361)	(0.1348)	(0.0975)	(0.1089)	(0.0722)	(0.0883)	
Observations	173269	466728	486332	406602	670227	468913	567104	
Panel (c): share minority	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
¡5%	-0.0407	-0.1121	-0.0902	-0.0534	-0.0878	-0.1383	-0.3484*	
	(0.2853)	(0.1836)	(0.2210)	(0.1405)	(0.1428)	(0.1209)	(0.2046)	
5-30%	-0.0894	-0.1284	-0.0380	-0.0387	0.0907	-0.0755	-0.5097**	
	(0.2744)	(0.1977)	(0.2028)	(0.1373)	(0.1456)	(0.1078)	(0.1980)	
30-70%	-0.4390	0.0428	0.4170*	0.0041	0.3454**	-0.0734	-0.3449	
	(0.2886)	(0.2054)	(0.2366)	(0.1814)	(0.1463)	(0.1555)	(0.2306)	
N (syllabi)	173269	466728	486332	406602	670227	468913	567104	
# Schools	1760	8530	6137	5257	22021	19747	19534	
Field, Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates of indicators for selectivity tiers (panel a), percentiles of median parental income in the school (panel b), and share of students who are either Black or Hispanic (panel c). 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-d, the lighter series also control for selectivity tiers. Standard errors are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AII: Share of Old Knowledge and Instructor Research Productivity: Publications and Citations

Panel (a): #publications	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	-0.0008	-0.0346	0.0108	0.0442*	-0.0227	0.0122	0.0121	-0.0290
	(0.0117)	(0.0323)	(0.0239)	(0.0267)	(0.0193)	(0.0195)	(0.0207)	(0.0201)
2nd quartile	-0.0485***	-0.0304	-0.0376	-0.0739**	-0.0392	-0.0635**	-0.0428	-0.0375
	(0.0179)	(0.0482)	(0.0495)	(0.0348)	(0.0277)	(0.0316)	(0.0315)	(0.0297)
3rd quartile	-0.0224	-0.0556	-0.0427	0.0195	-0.0779**	-0.0393	0.0098	-0.0305
	(0.0187)	(0.0505)	(0.0562)	(0.0328)	(0.0303)	(0.0341)	(0.0335)	(0.0296)
4th quartile	-0.0286	-0.0234	0.0462	-0.0108	-0.0853**	0.0045	-0.0138	-0.0634*
	(0.0249)	(0.0794)	(0.0936)	(0.0380)	(0.0418)	(0.0498)	(0.0425)	(0.0382)
Panel (b): #citations	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0075	0.0241	-0.0073	-0.0039	-0.0389	0.0320	-0.0209	-0.0479*
	(0.0162)	(0.0434)	(0.0396)	(0.0327)	(0.0255)	(0.0272)	(0.0284)	(0.0280)
2nd quartile	-0.0292*	-0.0698	-0.0246	-0.0038	-0.0441*	0.0137	-0.0570*	-0.0475*
	(0.0177)	(0.0488)	(0.0508)	(0.0341)	(0.0261)	(0.0323)	(0.0322)	(0.0274)
3rd quartile	0.0033	-0.0584	-0.0506	0.0692**	-0.0347	-0.0128	0.0005	0.0167
	(0.0204)	(0.0557)	(0.0656)	(0.0351)	(0.0320)	(0.0387)	(0.0361)	(0.0317)
4th quartile	-0.0739***	-0.0514		-0.0675*	-0.0977**	-0.1055**	-0.0231	-0.0738*
	(0.0274)	(0.0790)		(0.0408)	(0.0454)	(0.0524)	(0.0479)	(0.0428)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$). The independent variables are indicators for quartiles of the number of publications (panel (a)) and citations (panel (b)) of a course's instructors in the previous five years. The omitted category are courses with instructors with no publications or citations. For courses with more than one instructor, we consider the mean number of publications and citations across all instructors. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AIII: Share of Old Knowledge and Instructor Research Resources: Grant numbers and amount

Panel (a): #grants	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	0.0239	-0.0362	0.0189	0.0341	0.0057	0.0347	0.0088	0.0181
	(0.0174)	(0.0605)	(0.0397)	(0.0300)	(0.0302)	(0.0271)	(0.0323)	(0.0320)
2nd quartile	0.0112	0.1138	0.0960	-0.0294	0.0392	-0.0081	0.0158	0.0395
	(0.0243)	(0.0762)	(0.0611)	(0.0385)	(0.0417)	(0.0408)	(0.0422)	(0.0422)
3rd quartile	-0.0213	-0.0331	-0.0810	0.0137	0.0161	0.0339	-0.0741	-0.0494
	(0.0288)	(0.1072)	(0.0683)	(0.0439)	(0.0517)	(0.0495)	(0.0487)	(0.0497)
4th quartile	0.0145	0.0120	-0.0756	0.0473	0.0326	0.0301	-0.0800	0.0783
	(0.0325)	(0.0855)	(0.0800)	(0.0517)	(0.0644)	(0.0558)	(0.0566)	(0.0546)
Panel (b): grant amount ($)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	0.0114	0.0135	-0.0169	0.0162	-0.0405	-0.0225	0.0238	0.0594
	(0.0224)	(0.0811)	(0.0694)	(0.0301)	(0.0442)	(0.0353)	(0.0392)	(0.0422)
2nd quartile	-0.0008	0.0253	0.0382	-0.0345	0.0184	0.0392	-0.0531	-0.0063
	(0.0220)	(0.0735)	(0.0436)	(0.0434)	(0.0370)	(0.0354)	(0.0386)	(0.0397)
3rd quartile	0.0146	-0.0320	-0.0748	0.0546	0.0643	0.0412	-0.0047	-0.0008
	(0.0254)	(0.0757)	(0.0636)	(0.0431)	(0.0401)	(0.0432)	(0.0459)	(0.0422)
4th quartile	0.0260	0.0268	0.0175	0.0360	0.0491	0.1111**	-0.0934	0.0349
	(0.0322)	(0.0884)	(0.0700)	(0.0532)	(0.0633)	(0.0549)	(0.0570)	(0.0542)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$). The independent variables are indicators for quartiles of the number of grants (panel (a)) and total grant amount (panel (b)) ever obtained by a course's instructors. The omitted category are courses with instructors with no grants. For courses with more than one instructor, we consider the mean number and amount of grants. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AIV: Share of Old Knowledge and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2019)				
Panel (a): no controls	Grad rate (1)	Mean (2)	$P_y \leq 33$ pctile (3)	Median (4)	Mean (5)	P(top 20%) (6)	P(top 10%) (7)	P(top 5%) (8)	P(top 20%| $P_y \leq 20$ pctile) (9)
Share old knowledge	-0.0424***	-0.0594***	-0.0678***	-0.0499***	-0.0755***	-0.0338***	-0.0303***	-0.0226***	-0.0310***
	(0.0074)	(0.0108)	(0.0124)	(0.0104)	(0.0134)	(0.0063)	(0.0056)	(0.0035)	(0.0064)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): with controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Share old knowledge	-0.0040	-0.0034	-0.0027	-0.0018	-0.0109**	-0.0048	-0.0041*	-0.0032**	-0.0004
	(0.0035)	(0.0045)	(0.0059)	(0.0053)	(0.0046)	(0.0032)	(0.0024)	(0.0016)	(0.0034)
Mean dep. var.	0.5816	10.8281	10.7605	10.7096		0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

Note: OLS estimates of the coefficient δ in equation (4). The variable *Share old knowledge (sd)* is a school-level measure of the share of old knowledge, defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$), standardized to have mean zero and variance one. The dependent variable are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2019), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2019), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AV: Tail Measure and Instructor Research Productivity: Publications and Citations

Panel a): #publications	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	0.0215	0.0808	0.0572	0.0851	-0.0544	0.0293	0.0129	0.0089
	(0.0287)	(0.0708)	(0.0571)	(0.0711)	(0.0459)	(0.0502)	(0.0504)	(0.0470)
2nd quartile	-0.0643	0.0676	0.0805	-0.1265	-0.0480	-0.0368	0.0595	-0.1730**
	(0.0453)	(0.1001)	(0.1157)	(0.0955)	(0.0677)	(0.0832)	(0.0775)	(0.0719)
3rd quartile	0.0252	0.1763	0.1802	0.1000	-0.1163	0.1088	0.0631	-0.0688
	(0.0494)	(0.1192)	(0.1374)	(0.0941)	(0.0738)	(0.0965)	(0.0895)	(0.0719)
4th quartile	-0.0696	0.3777**	0.1791	-0.0716	-0.1932*	0.1203	-0.1951	-0.1373
	(0.0665)	(0.1813)	(0.2606)	(0.1061)	(0.1018)	(0.1353)	(0.1199)	(0.0954)
Panel b): #citations	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0034	-0.0649	0.0244	0.0853	0.0003	0.0454	0.0194	-0.0847
	(0.0404)	(0.0929)	(0.0869)	(0.0937)	(0.0604)	(0.0713)	(0.0718)	(0.0655)
2nd quartile	-0.0150	0.0210	0.0150	0.1274	-0.1159*	0.0521	0.0363	-0.1036
	(0.0464)	(0.1056)	(0.1256)	(0.0949)	(0.0659)	(0.0912)	(0.0806)	(0.0677)
3rd quartile	-0.0154	-0.0336	0.3280*	0.0017	-0.1040	0.0724	0.0023	-0.1007
	(0.0540)	(0.1306)	(0.1755)	(0.0963)	(0.0814)	(0.1063)	(0.0948)	(0.0807)
4th quartile	-0.1364*	0.2698		-0.1407	-0.2208**	-0.1332	-0.1588	-0.1465
	(0.0718)	(0.1721)		(0.1136)	(0.1074)	(0.1459)	(0.1263)	(0.1047)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is a tail measure of the education-innovation gap, calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution. The independent variables are indicators for quartiles of the number of publications (panel (a)) and citations (panel (b)) of a course's instructors in the previous five years. The omitted category are courses with instructors with no publications or citations. For courses with more than one instructor, we consider the mean number of publications and citations across all instructors. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AVI: Tail Measure and Instructor Research Resources: Grant numbers and amount

Panel (a): #grants	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	-0.0025	-0.1370	0.0323	0.0208	-0.0929	0.1008	-0.1772**	0.0229
	(0.0465)	(0.1378)	(0.0991)	(0.0870)	(0.0717)	(0.0752)	(0.0897)	(0.0774)
2nd quartile	-0.0292	0.1476	-0.0565	-0.1000	-0.0208	0.1156	-0.2210*	-0.0389
	(0.0649)	(0.1830)	(0.1413)	(0.1077)	(0.1104)	(0.1060)	(0.1226)	(0.1102)
3rd quartile	-0.0272	-0.5496***	0.2776*	-0.0284	-0.1821	0.2055	-0.1439	-0.1845
	(0.0765)	(0.2027)	(0.1519)	(0.1233)	(0.1317)	(0.1338)	(0.1358)	(0.1246)
4th quartile	-0.1386*	-0.4701**	0.0912	-0.1727	-0.0718	-0.0075	-0.2999*	-0.1367
	(0.0827)	(0.2078)	(0.1754)	(0.1415)	(0.1520)	(0.1417)	(0.1548)	(0.1339)
Panel (b): grant amount ($)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0162	-0.1819	-0.1930	-0.0136	-0.1026	0.0465	-0.1348	0.0404
	(0.0633)	(0.1829)	(0.1987)	(0.0873)	(0.1127)	(0.0976)	(0.1222)	(0.1125)
2nd quartile	0.0075	-0.1525	0.1264	0.0164	-0.1020	0.1325	0.0079	-0.1574*
	(0.0559)	(0.1571)	(0.0997)	(0.1220)	(0.0863)	(0.0938)	(0.1024)	(0.0932)
3rd quartile	-0.0579	-0.1507	-0.0242	-0.1032	0.0253	0.1082	-0.3365***	0.0031
	(0.0656)	(0.1726)	(0.1367)	(0.1200)	(0.1018)	(0.1092)	(0.1242)	(0.1081)
4th quartile	-0.1029	-0.4753**	0.1247	-0.0887	-0.1298	0.2088	-0.5267***	-0.0721
	(0.0823)	(0.2211)	(0.1670)	(0.1410)	(0.1489)	(0.1409)	(0.1516)	(0.1359)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is a tail measure of the education-innovation gap, calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution. The independent variables are indicators for quartiles of the number of grants (panel (a)) and total grant amount (panel (b)) ever obtained by a course's instructors. The omitted category are courses with instructors with no grants. For courses with more than one instructor, we consider the mean number and amount of grants. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AVII: Tail Measure and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2019)				
Panel (a): no controls	Grad rate (1)	Mean (2)	$P_y \leq 33$ pctile (3)	Median (4)	Mean (5)	P(top 20%) (6)	P(top 10%) (7)	P(top 5%) (8)	P(top 20%| $P_y \leq 20$ pctile) (9)
Gap, tail measure (5	(0.0087)	(0.0101)	(0.0117)	(0.0093)	(0.0138)	(0.0057)	(0.0049)	(0.0035)	(0.0054)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): with controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap, tail measure (5	(0.0035)	(0.0039)	(0.0049)	(0.0041)	(0.0045)	(0.0028)	(0.0019)	(0.0014)	(0.0034)
Mean dep. var.	0.5816	10.8281	10.7605	10.7096		0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

Note: OLS estimates of the coefficient δ in equation (4). The variable *Gap (tail, 5%) (sd)* is a school-level tail measure of novelty, calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution, standardized to have mean zero and variance one. The dependent variable are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2019), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2019), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AVIII: Soft Skills and Instructor Research Productivity: Publications and Citations

Panel a): #publications	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	0.1946***	-0.2299	0.6370***	0.1592	0.1712	0.2453**	0.2820**	0.0536
	(0.0642)	(0.2063)	(0.1395)	(0.1118)	(0.1230)	(0.1001)	(0.1192)	(0.1170)
2nd quartile	-0.1037	-0.7033**	-0.0748	0.0205	-0.0757	0.1338	-0.1671	-0.2709
	(0.0934)	(0.2848)	(0.2798)	(0.1388)	(0.1761)	(0.1510)	(0.1661)	(0.1679)
3rd quartile	-0.0695	0.2617	0.3822	-0.2015	-0.1340	-0.0618	0.0396	-0.1196
	(0.1000)	(0.3479)	(0.3349)	(0.1392)	(0.1942)	(0.1607)	(0.1865)	(0.1718)
4th quartile	0.2485*	0.8475*	0.6286	0.0071	0.4117	-0.1622	0.6034**	0.3276
	(0.1307)	(0.4934)	(0.6174)	(0.1645)	(0.2827)	(0.2255)	(0.2392)	(0.2152)
Panel b): #citations	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0873	-0.3166	0.3188	-0.0637	-0.1899	-0.0561	0.0653	-0.2348
	(0.0835)	(0.2642)	(0.2220)	(0.1280)	(0.1569)	(0.1292)	(0.1551)	(0.1530)
2nd quartile	-0.1411	-0.5172*	0.3753	-0.1989	-0.1551	0.1961	-0.2140	-0.3464**
	(0.0945)	(0.3026)	(0.3072)	(0.1388)	(0.1776)	(0.1549)	(0.1742)	(0.1624)
3rd quartile	-0.0402	0.1846	0.2835	-0.2097	-0.0855	-0.3766**	0.3093	-0.0004
	(0.1093)	(0.3691)	(0.4106)	(0.1479)	(0.2141)	(0.1834)	(0.2007)	(0.1837)
4th quartile	0.2250	0.6495		0.1192	0.1768	0.0553	0.4696*	0.1897
	(0.1390)	(0.4954)		(0.1732)	(0.2893)	(0.2318)	(0.2549)	(0.2332)
N (Course x year)	581657	59750	144842	168743	149506	209994	171776	199611
# Courses	153708	14885	39726	44809	38852	55555	43451	54622
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is the percentage of words contained in the assignment portion of each syllabus which refer to soft skills. The independent variables are indicators for quartiles of the number of publications (panel (a)) and citations (panel (b)) of a course's instructors in the previous five years. The omitted category are courses with instructors with no publications or citations. For courses with more than one instructor, we consider the mean number of publications and citations across all instructors. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AIX: Soft Skills and Instructor Research Resources: Grant numbers and amount

Panel (a): #grants	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
1st quartile	0.0820	-0.6054	0.0443	0.1170	0.1679	0.0781	0.0400	0.0969
	(0.0951)	(0.3844)	(0.2373)	(0.1329)	(0.1999)	(0.1388)	(0.1815)	(0.1836)
2nd quartile	0.1134	0.3906	0.1789	0.0619	-0.0307	-0.1807	0.5571**	0.1177
	(0.1223)	(0.5075)	(0.3431)	(0.1577)	(0.2968)	(0.1782)	(0.2367)	(0.2306)
3rd quartile	0.0209	-2.6437***	0.1450	0.0956	0.2737	-0.0803	0.1352	0.0605
	(0.1421)	(0.6125)	(0.3758)	(0.1846)	(0.3585)	(0.2206)	(0.2565)	(0.2652)
4th quartile	0.0966	-1.2181*	1.1869***	-0.0406	0.1053	-0.0229	0.5728*	-0.1698
	(0.1617)	(0.6588)	(0.4060)	(0.2172)	(0.3799)	(0.2498)	(0.3126)	(0.2841)
Panel (b): grant amount ($)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	0.1262	-0.7400	0.0444	0.1122	0.3440	0.1023	0.0437	0.2541
	(0.1078)	(0.4824)	(0.4349)	(0.1267)	(0.2988)	(0.1544)	(0.2058)	(0.2164)
2nd quartile	0.0088	-0.7438	0.2977	-0.0939	0.0324	-0.2534	0.3109	0.0833
	(0.1183)	(0.4669)	(0.2509)	(0.1885)	(0.2366)	(0.1786)	(0.2263)	(0.2172)
3rd quartile	0.1652	-1.0022*	0.5972*	0.0408	0.1466	-0.0082	0.5456**	0.0364
	(0.1300)	(0.5221)	(0.3279)	(0.1823)	(0.2897)	(0.1983)	(0.2358)	(0.2441)
4th quartile	0.1127	-0.7600	0.3951	0.1793	0.0302	0.1955	0.3891	-0.1947
	(0.1621)	(0.6699)	(0.4055)	(0.2207)	(0.3725)	(0.2507)	(0.3123)	(0.2828)
N (Course x year)	581657	59750	144842	168743	149506	209994	171776	199611
# Courses	153708	14885	39726	44809	38852	55555	43451	54622
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates, one observation is a course. The dependent variable is the percentage of words contained in the assignment portion of each syllabus which refer to soft skills. The independent variables are indicators for quartiles of the number of grants (panel (a)) and total grant amount (panel (b)) ever obtained by a course's instructors. The omitted category are courses with instructors with no grants. For courses with more than one instructor, we consider the mean number and amount of grants. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AX: Soft Skills and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2019)				
Panel (a): no controls	Grad rate (1)	Mean (2)	$P_y \leq 33$ pctile (3)	Median (4)	Mean (5)	P(top 20%) (6)	P(top 10%) (7)	P(top 5%) (8)	P(top 20%| $P_y \leq 20$ pctile) (9)
Soft skills (%)	0.0982***	0.0935***	0.0966***	0.0818***	0.1125***	0.0497***	0.0394***	0.0293***	0.0521***
	(0.0066)	(0.0093)	(0.0118)	(0.0078)	(0.0116)	(0.0053)	(0.0043)	(0.0034)	(0.0050)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): no controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Soft skills (%)	0.0116***	0.0172***	0.0096*	0.0209***	0.0125*	0.0103***	0.0028	0.0007	0.0119***
	(0.0035)	(0.0057)	(0.0058)	(0.0056)	(0.0064)	(0.0037)	(0.0028)	(0.0019)	(0.0043)
Mean dep. var.	0.5816	10.8281	10.7605	10.7096		0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

Note: OLS estimates of the coefficient δ in equation (4). The variable *Soft skills (%)* is the school-level share of words in the assignment portion of a syllabus which refer to soft skills, estimated as $\theta_{s(i)}$ in equation (3)), standardized to have mean zero and variance one. The dependent variable are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2019), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2019), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table AXI: 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 AXII: 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 AXII. 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	

Note: List of schools for which we collected course catalog data.

Table AXIII: 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.