The Education-Innovation Gap*

Barbara Biasi† Song Ma‡

July 2, 2020

(Preliminary and Incomplete.)

Abstract

Higher education can play a crucial role in the production of human capital by providing people with "up-to-date" knowledge. We propose a novel approach to quantify the provision of this type of knowledge across US institutions: the education-innovation gap, a measure of the textual distance between the content of higher-education courses and frontier knowledge. We calculate the gap for 3 million US college and university courses by mapping the text of their syllabi to the text of 20 million old and new academic publications. With this new measure we document three main findings. First, instructors (as opposed to fields or schools) explain the largest portion of the total variation in the gap. Second, access to frontier knowledge is unevenly distributed across schools serving students from different backgrounds. Third, the gap is correlated with students' outcomes.

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

Keywords: Education, Innovation, Text Analysis, Inequality

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

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

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

1 Introduction

Human capital is a key driver of innovation and growth (Romer, 1990, 1994). Endogenous growth models describe human capital as a factor-augmenting term in the production of knowledge and ideas, which are non-rival and create positive externalities that fuel growth. Not all human capital, however, is created equal. While earlier studies simply measured it using years of education, more recent works have identified specific types of investments that produce valuable knowledge, including scientific and technical education (Baumol, 2005), social learning and interactions (Lucas Jr and Moll, 2014; Lucas Jr, 2015; Akcigit et al., 2018), and mentorship (Bell et al., 2019).

Naturally, education systems play a crucial role in equipping individuals with valuable knowledge (Biasi et al., 2020). Again, not all education systems are the same. STEM programs, for example, appear most related to invention and growth (Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019). Even within fields, large differences exist across schools and sectors: For instance, most US inventors come from a very small set of elite schools (Bell et al., 2019). What makes some programs stand out for producing valuable knowledge for innovation and growth, however, largely remains a "black box." Yet, understanding how this knowledge is produced and disseminated is very important, especially in light of recent evidence suggesting that the burden of knowledge necessary to innovate has been increasing over time (Jones, 2009), that "ideas are getting harder to find" (Bloom et al., 2020), and that access to invention might be almost impossible for individuals with less advantaged backgrounds (Bell et al., 2019).

In this paper we begin to open this "black box," and we use a novel approach to quantify the extent to which higher education systems are able to equip students with the valuable knowledge that leads to innovation and growth. We do so by developing a new measure: the *education-innovation gap*, which captures the distance between the content of college and university courses (described by course syllabi) and frontier "technologies," such as academic research papers and patents published at different points in time. The intuition behind this measure is the following: A course has a smaller gap if its content is more similar to newer technologies than it is to older ones. For example, an introductory Com-

puter Science course that teaches the programming language *Julia* in 2018 is closer to the technological frontier than one that teaches *Visual Basic*.

We construct this measure using a "text as data" approach applied to course syllabi, academic papers, and patents. Specifically, we compare the text of three million college and university syllabi, corresponding to about 800,000 courses in 62 different fields taught at nearly 2,500 US institutions between 1992 and 2018, with (a) the text of the abstract of 20 million academic publications published between 1975 and 2019; and (b) the text of more than 6 million patents filed between 1976 and 2019. To do this, we first project the text of each document on an existing dictionary (for example, the list of all words listed on *Wikipedia*). This allows us to represent each document as a binary vector, whose elements correspond to dictionary words. Following Kelly et al. (2018), we use weighting techniques to account for the length of each document and the frequency of each word in the document relative to the English language. We then calculate the cosine similarity, a standard measure of vectorial distance, between each syllabus and all the publications and patents released up to 15 years prior to the syllabus year.

Using these cosine similarities, we construct the education-innovation gap of a syllabus as the ratio between the average cosine similarity with technologies released 15 years before a course is taught and the similarity with technologies released one year before. Intuitively, the gap will be larger for syllabi that are more similar to older technologies than to newer ones. Following the example above, a Computer Science course that teaches *Julia* in 2018 should have a smaller gap than one that teaches *Visual Basic*. We validate this measure using the list of required or recommended readings referenced in each syllabus, and we show that syllabi with a larger gap reference older readings. This suggests that our measure performs well in capturing the "novelty" of the content of each course.

Armed with this new measure, we document a set of new findings related to the offering of frontier knowledge in higher education. First, we show that the gap varies significantly across schools, fields, and over time: In order to move a syllabus from the 75th percentile to the 25th percentile of the overall distribution, one would have to replace 17 percent of its content with newer material.

Second, we demonstrate that differences among instructors explain the largest portion

(86 percent) of the total variation in the gap. By comparison, differences among fields only explain 14 percent, and differences across institutions explain 11 percent. Using instructor turnover across courses in an event-study setting (as in Rivkin et al., 2005), we also show that the gap of a given course drops significantly when the instructor changes, which suggests that turnover contributes to making the content of a course more up-to-date.

Third, we demonstrate that institutions disproportionately serving more advantaged students offer significantly more up-to-date courses. For example, the gap is significantly smaller in "Ivy Plus" institutions (Ivy league school plus Duke, MIT, Stanford, and the University of Chicago), compared with all other institutions. In order to make the average syllabus in a non-selective institution equivalent to a syllabus of a course in the same field taught in an Ivy Plus or elite institution, one would have to replace four percent of its content. Similarly, the gap is significantly smaller in schools serving children from more affluent backgrounds (measured by the share of students with parents in the top one percent of the income distribution). The gap is also disproportionately larger for schools primarily enrolling Black and Hispanic students, compared with schools where these students are a small minority. Cross-school differences in the gap are most pronounced for Social Sciences and STEM courses relative to, for example, Humanities courses.

Lastly, we find that the gap strongly correlates with a range of students' outcomes, including graduation rates, income, and measures of intergenerational mobility (Chetty et al., 2019). For example, a 0.1 smaller gap (i.e., that of a syllabus that has 25 percent more "knowledge" terms compared with the mean syllabus) is associated with a 4.4 percentage point higher graduation rate, or 10 percent compared with the mean. The same difference in the gap is also correlated with a 6 percent higher income and with a 5 percent larger probability that students with parents in the bottom income quintile reach the top quintile during adulthood.

Importantly, these patterns are almost entirely driven by schools serving students from low-income families. In fact, the correlation between the gap and either graduation rates, income, and mobility is small and indistinguishable from zero for schools serving more wealthy students. The correlation is instead negative and large for schools with a more economically disadvantaged student body. For example, while a 0.1 smaller gap bears a

zero correlation with students' incomes in schools with more than 5 percent of students with parental income in the top percentile, it is associated with a 40 higher student income in schools serving less than 0.1 percent of top-parental income students. While not sufficient to establish a causal link, these findings suggest that the content of higher education is particularly crucial for students who do not have access to the network and resources provided by elite schools.

Taken together, our findings reveal large differences in the provision of up-to-date knowledge across institutions serving different types of students. At the same time, they suggest a possibly large role for these differences in explaining later-life outcomes. Our results also highlight an important role for instructors in shaping the content of higher education, a result analogous to what has been found in the literature on teachers in K-12 education (Rockoff, 2004; Chetty et al., 2014).

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

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

to the frontier of knowledge.

Third, this paper relates to the literature on the "production" of innovation. Earlier works (Nelson and Phelps, 1966; Benhabib and Spiegel, 2005) have highlighted an important role of human capital for technology diffusion. More recently, Jones (2009) has shown how innovators are increasingly required to update their knowledge and skills to keep up with a fast-paced world; failure to do so delays the correlation between researcher/inventor life-cycle productivity (Jones, 2010; Jones and Weinberg, 2011). Technical and scientific education has been associated with more innovation and growth (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019).[1] Our paper contributes to this body of evidence by taking a more "micro" approach to quantify the extent to which higher education is able to provide students with up-to-date knowledge, necessary to innovate.

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

2 Data

Our empirical analysis combines different types of data, including the text of course syllabi; the abstract of academic publications; information on US higher education institutions; and measures of labor market outcomes for the students at these institutions. .

[1]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.

2.1 College and University Course Syllabi

We obtained the text of college and university course syllabi from the American Assembly's Open Syllabus Project (OSP).[2] The main data set includes more than seven million English-language syllabi from institutions in over 80 countries, with a few syllabi dating as far back as the 1960s. To maximize coverage, we focus our attention on 3,048,530 syllabi taught at 3,186 US institutions between 1992 and 2018.

A typical syllabus contains several sections, including a course's basic identifying information (such as the name and the code), a description of its content, a list of references and recommended readings, course requirements (such as assignments and exams), an outline of the grading policy, and a brief description of other policies regarding absences, misconduct, plagiarism, etc.

To obtain information on a course's content, central to our analysis, we proceed as follows. First, we parse the full text of each syllabus to extract the name of the institution, the course's name and code, its level (either introductory, advanced, or graduate-level), the name of the instructor, the academic year and, when available, the term (i.e.m Fall, Winter, Spring, or Summer).[3] Information on course codes is particularly useful because it allows us to compile a panel of courses taught at a given institution over time. Combined with information on the instructors, this it allows us to observe how the syllabus of a given course evolves when the person in charge of teaching the course changes. OSP assigns each syllabus to one of 62 detailed fields, such as English Literature, Computer Science, or Economics. In most of our analyses we aggregate these fields into six macro-categories: STEM, Humanities, Social Sciences, Business, and Vocational. This aggregation is outlined in Appendix Table AI.

Second, we extract the full text of the parts of each syllabus that are likely to contain information on its "knowledge" content. The first is the course description, which typically describes the basic structure of the course, the key concepts that are covered, and (in

[2]OSP collects its 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.

[3]We extract these pieces of information using named-entity recognition (NER) techniques.

many cases) a timeline of the content and the materials for each lecture. We identify this part of the syllabus searching for section titles such as "Course Summary" and "Course Description." On average, this section contains 575 words.

The second portion is the list of references, which contains bibliographic information on the required and recommended readings for the course. We identify this section for 62 percent of all syllabi searching for section titles such as "References", "Readings", and "Textbooks." We also search for and collect other in-text citations (such as "Biasi and Ma (2020)"). We complement the bibliographic information of each reference item included in the syllabus (such as the year of publication or the textbook edition) with information from Elsevier's SCOPUS database (see Section 2.2 below for additional details). In Section 3.3 we use the list of references to conduct a validation exercise for our measure of the education-innovation gap.

Sample Construction and Selection Our sample of syllabi corresponds to a subgroup of all college and university courses taught in the US between 1992 and 2018. The sample coverage improves across the years, with the number of covered syllabi, syllabi per instructor, and institutions increasing rapidly over time (Appendix Figure AI). In an ideal situation, the group of syllabi we observe in each year would represent a randomly selected sample of the syllabi of all courses taught in US institutions in that year. A potential concern for our analysis, however, is that the selection into the sample might be driven by unobservable, time-varying attributes of a syllabus that are also correlated with other characteristics of the course or the institution.

To obtain a better understanding of these selection patterns, we perform a series of checks. First, we examine whether the increase in sample size across time disproportionately affects some fields. Appendix Figure AII shows the macro-field composition of the syllabus sample in five-year intervals between 1995 and 2015. The field composition appears stable over time, with STEM courses representing between 30 and 35 percent of the sample, Humanities representing 25 percent, and the Social Sciences representing between 20 and 30 percent of all syllabi in each year. .

Second, we test whether our working sample over- or under-represents specific geo-

graphic areas of the US. Appendix Figure AIII shows the number of institutions (panel (a)) and of courses taught syllabi between 2016 and 2018 (panel (b)) in each state. The distribution of these two variables is very similar across states, indicating that our sample does not cover certain areas more than others.

While the evidence in Appendix Figures AII ad AIII is useful to describe our working sample, it is not informative of how our sample compares to the population of all syllabi, nor whether this comparison varies over time. To better compare our sample to the population of all syllabi we compiled the full list of courses offered between 2010 and 2019 in a subsample of 161 US institutions, by hand-collecting information on the course catalogues maintained in each school's archive.[4] Figure 1 shows the trend in the share of courses taught in these institutions that are included in our sample. This share is approximately equal to five percent and is relatively flat throughout the time period. This indicates that, at least in the subsample of schools for which catalogue information is available and throughout this time period, the increase in the number of syllabi shown in Appendix Figure AI is likely driven by an increase in the number of courses that are offered, rather than by an increase in sample coverage.

As an additional piece of evidence of the sample selection patterns, in Table 2 we test whether the share of syllabi included in the sample is correlated with a range of institutional characteristics such as selectivity, financials, and enrollment. Panel (a) shows means and standard errors of the share of covered syllabi across selectivity tiers. In 2013, this share ranged from 0.03 percent for non-selective schools to 4.42 percent for highly selective and selective public schools (right columns); these shares are, however, statistically indistinguishable across tiers. The same is true for the 2010-2013 change in the share of covered syllabi (left columns). Panel (b) shows instead the correlation between the share of syllabi included in the sample and a set of financials (such as expenditure on instruction, endowment per capita, sticker price, and average salary of all faculty), enrollment, the share of students in different categories (Black, Hispanic, alien), and the share of students graduating in Arts and Humanities, STEM, and the Social Sciences. These correlations are all

[4]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.

statistically indistinguishable from zero. While we cannot rule out that sample selection is driven by unobservables, these results reassuringly indicate that selection is not associated with a school's observable characteristics.

2.2 Academic Publications

To map the content of each syllabus with frontier research, we collected data on academic publications that appeared in top journals within each field. We define top journals as those which were ranked among the top 10 by Impact Factor in each field at least once since 1975. We then extracted information on the articles published in these journals since their foundation from Elsevier's SCOPUS dataset.[5] Our final list of publications includes 20 million peer-reviewed articles in the same fields as our syllabi, corresponding to approximately 100,000 articles per year.[6] We capture the knowledge content of each article using its title, abstract, and keywords.

2.3 Information on US Higher Education Institutions

The last component of our dataset includes institution-level information on all universities and colleges for which syllabi texts are available. Our primary source of data is the the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES).[7] Available information includes institutional characteristics (such as name and address, control or affiliation, levels of awards offered, types of programs, tuition and fees), institutional prices, SAT and ACT scores of admitted students (when applicable), enrollment and completion rates. We link syllabi and IPEDS records using a fuzzy matching algorithm based on institution names; we are able to successfully link 89 percent of all syllabi.

We complement this data set with two additional sources of institution-level informa-

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

[6] SCOPUS categorizes 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.

[7] IPEDS has collected data through a series of interrelated surveys from all postsecondary institutions since 1993. The completion of these surveys is mandatory for all institutions that participate in, or are applicants for participation in, any federal financial assistance program.

tion on schools' and students' characteristics. The first one is the dataset used (and made available) by Chetty et al. (2019), assembled through a variety of sources (including tax records). This dataset includes a set of institutional characteristics, such as school selectivity, defined using Barron's selectivity scale; students' race and ethnicity; the incomes of students and their parents, obtained from tax records for the years 1996 to 2012; and a measure of intergenerational mobility, defined as the share of students with income in the top quintile and parental income in the bottom quartile.

The second source is the College Scorecard Database of the US Department of Education, an online tool made available to consumers to compare the cost of higher education in the country. We use this database as a source of information on the incomes of graduates ten years after they started college and on graduation rates of students by cohort. The data is available for the academic years 1996-97 to 2017-18.

2.4 Sample Description

Panel (a) of Table 1 summarizes our syllabi sample. Out of 2,929,346 US syllabi taught between 1992 and 2018 in 2,417 institutions, we successfully link 2,576,191 syllabi from 2,396 institutions to IPEDS records. We further link 2,570,962 syllabi to data from Chetty et al. (2019). Our final sample contains 91,407 syllabi per year, with 944 syllabi in 1992, 37,106 in 2001, and 271,955 in 2017. These correspond to an average of 0.416 syllabi per FTE instructor per year, including 0.0140 in 1992, 0.132 in 2001, and 0.9896 in 2017. The fields with the most syllabi are Mathematics (with 214,912 syllabi across all years), English Literature (with 186,809 syllabi), and Business (with 138,639 syllabi).

3 Measuring the Education-Innovation Gap

To construct the education-innovation gap we combine textual information from course syllabi with information that captures frontier knowledge, such as academic publications. In this section we describe this measure in detail, provide the intuition behind it, and perform validation checks.

3.1 Measuring Similarities in Text

The first step for the construction of the gap consists of computing textual similarities between pairs of syllabi and technologies. To do so we begin by representing each text document d, i.e., a syllabus or a paper, 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 $w \in W$. A standard measure of textual similarity between two documents d and k is the cosine similarity between $\tilde{V}_d$ and $\tilde{V}_k$, defined as

$$\rho_{dk} = \frac{\tilde{V}_d}{\|\tilde{V}_d\|} \cdot \frac{\tilde{V}_k}{\|\tilde{V}_k\|} \tag{1}$$

This metric captures the proximity between d and k in the space of words W. We make several adjustments to this simple measure, which we describe below.

Accounting for term frequency and relevance First, we want to overweight terms that best capture the knowledge content of a given document, and underweight those that are used frequently but do not necessarily capture content. To do so, we need to account for the frequency of a term in the English language and in the body of all texts. For example, a term such as "genome editing" is more rarely used both in the English language and across all syllabi compared with a term such as "assignment." It will therefore be more informative of the content of a given syllabus and, as such, it should receive more weight.

We account for the relevance of each term in two ways. First, we construct our document vectors focusing only 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. We use two dictionaries: (1) the list of all unique terms ever used as keywords in academic publications from the beginning of our publication sample to 2019 (the "keywords" dictionary); and (2) the list of all terms that have an English Wikipedia webpage as of 2019 (the "Wikipedia" dictionary). Our main analyses uses the keywords dictionary; our results are robust to using the Wikipedia dictionary.

Second, we use a standard approach in the textual analysis literature, the "term-frequency-

inverse-document-frequency (TFIDF)" transformation of word counts, to weigh each term by its true "relevance" in capturing the content of a text. The intuition behind the TFIDF approach is as follows: A term w will receive a higher weight in document d if (a) it appears more frequently in document d and (b) it is used less frequently across all documents of the same type as d.

In practice, this approach can be implemented by weighing each term w in document d by the quantity

$$TFIDF_{wd} = TF_{wd} \times IDF_w \tag{2}$$

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

$$IDF_w \equiv \log\left(\frac{|D|}{\sum_d \mathbb{1}(d \text{ contains word } w)}\right) \tag{3}$$

is the inverse document frequency of term w in the set D of all documents of the same type as d. Intuitively, the term TF_{wd} overweighs terms that are used frequently in document d, while the term IDF_d underweighs terms that are common across all documents in D. As a result, a term w has a high $TFIDF_{wd}$ if it appears relatively frequently in d but is not too common among other documents in D, which implies that w is a distinguishing feature of the knowledge content of document d.

Accounting for changes in term relevance over time The weighting approach described so far calculates IDF by pooling together documents published in different years. This is not ideal for our analysis, because it ignores the temporal ordering of syllabi and technologies. As we explain below, we are instead interested in the novelty of the content of a syllabus d relative to technologies published in the years prior to d, without taking into account the content of future technologies. To see this consider, for example, course CS299 at Stanford University, taught by Andrew Ng in the early 2000 and one of the first entirely focused on *Machine Learning*. Pooling together documents from different years would result in a very low $TFIDF_{wd}$ for the term "machine learning" in the course's syllabus: Since the term has been used very widely in the last years, its frequency across all documents

would be very high and its IDF very low. Not accounting for changes in the frequency of this term over time would then lead us to misleadingly underestimate the course's path-breaking content.

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

$$BIDF_{wt} \equiv \log\left(\frac{\sum_d \mathbb{1}(d \text{ published before } t)}{\sum_d \mathbb{1}(d \text{ published before } t) \times \mathbb{1}(d \text{ contains word } w)}\right) \tag{4}$$

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 $TFIDF$, using $BIDF$ in lieu of IDF:

$$TFBIDF_{wd} = TF_{wd} \times BIDF_{wt(d)} \tag{5}$$

where $t(d)$ is the publication year of document d. Continuing with our example above, the term "machine learning" would have a higher $BIDF$ in the early 2000s than inn 2018 (since in the early 2000s the term was not yet widely used) and, in turn, attribute a higher $TFBIDF$ to course CS299. This would allow us to better characterize the content of the course.

Building the weighted cosine similarity Having calculated $TFBIDF_{wd}$ for each term w and text d, we can obtain a weighted version of our initial vector $\tilde{V}_d$, denoted as V_d; each element w of V_d is equal to $TFBIDF_{wd}$. We can then re-define the cosine similarity between two texts d and k, accounting for term relevance, as

$$\rho_{dk} = \frac{V_d}{\|V_d\|} \cdot \frac{V_k}{\|V_k\|}. \tag{6}$$

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 overlapping terms, $\rho_{dk} = 0$.

3.2 Calculating the Education-Innovation Gap

We calculate cosine similarities between each syllabus and each technology, such as academic publications in the same field, using both dictionaries (SCOPUS keywords and Wikipedia). With these measures we are now ready to construct the education-innovation gap. Our goal is to devise a measure that will be smaller for syllabi that are more similar to recent technologies, relative to older ones.

For each syllabus d we first define the average similarity of a syllabus with all technologies published in a given time period:

$$S_d^\tau = \sum_{k \in \Omega_\tau(d)} \rho_{dk} \tag{7}$$

where ρ_{dk} is the cosine similarity between syllabus d and a technology k, defined in equation (6), and $\Omega_\tau(d)$ is the set of all technologies published in the three-years time interval $[t(d) - \tau - 2, t(d) - \tau]$, where $t(d)$ is the year of publication of syllabus d.[8] We then define the education-innovation gap as the ratio between the average similarity of a syllabus with older technologies and the similarity with more recent ones:

$$Gap_d \equiv \left(\frac{S_d^{13}}{S_d^1} \right) \tag{8}$$

It follows that a syllabus published in t has a lower education-innovation gap if its text is more similar to the text of technologies published between $t-3$ and $t-1$ than to the text of technologies published between $t-15$ and $t-13$.

At a first glance, one might be tempted to simply define that education-innovation gap as S_d^1, the similarity with the most recent technologies. This measure, however, would be sensitive to idiosyncratic differences in the "style" of language across syllabi in different fields, or even within the same field. Being defined as a ratio of similarities *to the same syllabus*, our measure is essentially free of any time-invariant, syllabus-specific components of S.

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

3.3 Validating The Education-Innnovation Gap

To validate our measure and to confirm that it captures the novelty of the content of a syllabus, we use information from each syllabus' list of references. Specifically, we check whether syllabi with a lower gap list newer references. Figure 2 plots the relationship between the average reference "age," defined as the difference between the publication year of each syllabus and the publication year of each reference. Reassuringly, this relationship is strong and almost linear, with a correlation of 0.83. .

3.4 Interpreting the Economic Magnitude

To illustrate how a unit change in the gap translates into differences in the content of a course, we perform a simple simulation exercise. The thought experiment asks the following question: How many knowledge terms do we need to replace in a given syllabus to generate a given change in the education-innovation gap? We answer this question by randomly selecting a subsample of 100,000 selected syllabi and by replacing "old" terms with "new" ones in each syllabus, where "old" and "new" are defined based on the frequency of each term among all publications in the same field in the year prior to the one of the syllabus (old are those in the bottom 5 percent of the frequency distribution, new are those in the top 5 percent). We then recalculate the gap for each syllabus as we gradually replace more words.

This exercise is illustrated in Figure 3, which shows the relationship between the number of replaced terms (horizontal axis) and the average change in the gap (vertical axis). Replacing 20 older terms with 20 newer terms is associated with a 0.01 reduction in the gap. On average, each syllabus includes 480 terms; a 0.01 reduction in the gap requires replacing approximately 4.2 percent of the content of the syllabus.

4 Decomposing The Education-Innovation Gap

We begin our analysis describing how the education-innovation gap varies across fields, across institutions within the same field, across courses within the same institution, and

within the same course as the instructor changes over time.

The overall variation in the gap across all syllabi between 1992 and 2018 is illustrated in Figure 4 (solid line). The average course has a gap of 0.96, with a standard deviation of 0.051, a 25th percentile of 0.927, and a 75th percentile of 0.987. To better quantify the extent of this variation we make use of the relationship illustrated in Figure 3: In order to move a syllabus from the 75th to the 25th percentile one would have to replace approximately 80 of its knowledge terms (or 17 percent of the average syllabus, which contains 480 terms).

Figure 4 also shows the distribution of the gap within fields, institutions, and instructors. The distribution of the gap within instructors is much less dispersed compared with the distribution within fields and within institutions (with a standard deviation of 0.039 within instructor, compared with 0.049 within field and 0.049 within institution). This suggests that most of the variation in the gap occurs across courses taught by different people in the same field and institution, rather than across schools or fields.[9]

To more rigorously quantify the contribution of the different attributes to the variation in the gap, we estimate an OLS regressions of the gap as a function of year fixed effects. We then report how the R-squared of this "baseline" regression (which captures the "unexplained" portion of the variation in the gap) changes as we gradually add institution, field, course, and instructor fixed effects to the baseline model. These R-squared are reported in column 1 of Table 3; in column 2 we express each R-squared as a share of the R-squared of the baseline regression. This share represents the proportion of the total variation explained by the additional fixed effects included in each regression. This exercise reveals that differences among institutions explain 11 percent of the variation in the gap; differences among fields explain 14 percent, differences among courses explain 66 percent, and differences among instructors explain 62 percent. Combined, instructors and courses explain 86 percent of the total variation in the gap.

[9]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, and instructors fixed effects. We then added the mean gap to each set of residuals.

4.1 Instructors and the Education-Innovation Gap

The results of the decomposition indicate, perhaps not surprisingly, that instructors play the biggest role in shaping the content of the courses they teach.[10] As an additional test of this hypothesis, we follow the literature on the role of teachers for student achievement (Rivkin et al., 2005; Chetty et al., 2014) and study the evolution of the education-innovation gap of a course when its instructor changes over time. To do so we perform an event study in a 7-years window around the time of the instructor change. We estimate:

$$Gap_{ct} = \sum_{k=-5}^{5} \mathbb{1}(t - C_c = k)\delta_k + \theta_c + \tau_t + \varepsilon_{ct} \tag{9}$$

where Gap_{ct} is the gap of course c in year t and C_c denotes the year of the instructor change.[11] Course fixed effects θ_c allow us to study changes in the gap for the *same* course over time, and year fixed effects τ_t account for secular changes in the gap that are common across all courses. In this equation, the parameters δ_k capture the differences between the gap k years after an instructor change relative to the year of the change.

OLS estimates of δ_k, shown in Figure 5, are indistinguishable from zero and on a flat trend in the years leading to an instructor change. After the change, however, the gap drops significantly and continues to decline up to three years after the change. Three years post-change, the gap is 0.002 lower compared with the year of the change. This decline is equivalent to replacing 10 knowledge terms to the average syllabus, or 3 percent of its content.

In Table 4 we re-estimate equation (9) pooling together the years before and after an instructor change. This exercise indicates an average variation in the gap equal to -0.0006 in the three years following the change (column 1, significant at 1 percent). This estimate is robust to the inclusion of institution-by-year fixed effects (to account for changes common across all courses in a given school, column 2) and field-by-year fixed effects (to account for secular changes in the gap that are common across all courses in a given field, column

[10] This result is analogous to the large role of teachers for student achievement illustrated by many studies, including Rockoff (2004), Kane and Staiger (2008), and Chetty et al. (2014), among others.

[11] We use the first change observable in our sample; results are robust to this choice.

3). Table 5 shows estimates by macro-fields. After an instructor change, the education-innovation gap declines the most for courses in Math, Economics, Psychology, and History. The decline is the smallest for Engineering and English Literature.[12]

Taken together, these estimates confirm the importance of instructors in determining the novelty of the content of the courses they teach. They also indicate that, when a new instructor takes over an existing course, they update the material to make it significantly more up-to-date and closer to the knowledge frontier.

5 The Education-Innovation Gap Across Schools

Our decomposition exercise indicates that differences across instructors explain the largest portion of the variation in the gap. Yet, differences across schools are non-negligible and account for 10 percent of the total variation. Cross-school differences are particularly important if they appear related to the characteristics of the students who attend these institutions.

We now describe how the education-innovation gap varies among different types of institutions. We focus our attention on three school characteristics: Selectivity, students' parental income, and the racial and ethnic makeup of the student population. The ultimate goal of our exercise is to assess whether access to up-to-date educational content differs among students with different backgrounds, who attend different types of schools.

5.1 School Selectivity

To study the relationship between the gap and school selectivity, we divide schools in five "tiers." We follow the classification of Chetty et al. (2019), based on (i) Barron's 2009 selectivity ranking, (ii) control (public/private), and (iii) type of degree offered (four-year or two-year). Our tiers are as follows: "Ivy Plus" include Ivy League colleges and the University of Chicago, Stanford, MIT, and Duke; "Elite" include all other schools classified as Tier 1 in Barron's ranking; "Highly selective and selective public" include public schools

[12]Tables 4 and 5 show OLS estimates of the parameter δ in the equation $Gap_{ct} = \delta Change_{ct} + \theta_c + \tau_t + \varepsilon_{ct}$, where $Change_{ct}$ is the share of instructors of course c who are teaching the course for the first time in our time period. Standard errors are clustered at the course level.

in Barron's Tiers 2 to 4; "Highly selective and selective private" include private schools in Tiers 2 to 4; "Non-selective" includes Barron's Tiers 5 to 9 and all four-year colleges not included in Barron's classification; and "Two-year" include all two-year institutions.

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

$$\begin{aligned} Gap_i &= \beta_1 \text{Ivy}_{s(i)} + \beta_2 \text{Elite}_{s(i)} + \beta_3 \text{SelPub}_{s(i)} + \beta_4 \text{SelPriv}_{s(i)} \\ &\quad + \beta_5 \text{NonSel}_{s(i)} + \beta_6 \text{TwoYear}_{s(i)} + \phi_{f(i)} + \tau_{t(i)} + \varepsilon_i \end{aligned} \tag{10}$$

where Gap_i measures the education-innovation gap of syllabus i, taught in school $s(i)$ in year $t(i)$. The indicator variables Ivy_s, Elite_s, SelPub_s, SelPriv_s, NonSel_s, and TwoYear_s equal one if school s belongs to the Ivy Plus, elite, highly selective and selective public, highly selective and selective private, non-selective, and two-year tiers, respectively. Field fixed effects ϕ_f control for systematic, time-invariant differences in the gap across fields. Year fixed effects τ_t control flexibly for secular trends in the gap that are common across all syllabi. We cluster standard errors at the institution level.

Point estimates of the coefficients $\beta_1 - \beta_6$ in equation (10), which represent conditional mean gaps for schools in each tier, are shown in panel (a) of Figure 6 along with confidence intervals. These estimates indicate that the gap is significantly lower for more selective schools, and it progressively increases as selectivity declines. Ivy Plus have the smallest gap, at 0.950, followed by Elite schools at 0.952. The gap increases to 0.958 for non-selective schools and 0.960 for two-year schools. Combined with the calculation in Figure 3 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 20 knowledge terms to the average syllabus in non-selective schools (or 4 percent of its content).

The evidence in panel (a) of Figure 6 is confirmed by the estimates shown in Table 6, where we re-estimate equation (10) grouping together Ivy Plus and elite and selective public and private schools, respectively, and we use non-selective and two-year schools as the reference tier. In column 1 we omit field fixed effects: Relative to an average of 0.97 for non-selective and two-years schools, the gap is 0.012 smaller for Ivy Plus and elite (equivalent to 22 additional knowledge terms, significant at 1 percent), and 0.007 smaller

for other highly selective and selective schools (or 15 knowledge terms, significant at 1 percent). These estimates are only slightly smaller when controlling for field-by-year fixed effects (Table 6, columns 2 and 3).

Differences Across Fields In columns 3-7 of Table 6 we re-estimate the specification in column 1 separately by macro-fields. The differences in the gap across schools in different tiers are most pronounced for Social Sciences courses. For example, compared with non-selective and two-years schools, Ivy-plus schools have a 0.015 lower gap for Social Sciences courses (or 25 knowledge words, significant at 1 percent), and a 0.008 lower gap for STEM, Humanities, and Business courses (15 knowledge terms, significant at 1 percent). The difference in gap across selectivity tiers is instead smaller and insignificant for syllabi of Vocational courses.

5.2 Parental Income

Ivy-Plus and elite schools, whose courses have the smallest education-innovation gap, are disproportionately attended by students from wealthier backgrounds (Chetty et al., 2019). This suggests that access to up-to-date educational content might be unequally distributed across more and less advantaged students. For a more direct test of this hypothesis we describe how the gap differs across schools serving students with different socio-economic backgrounds. We measure backgrounds with two statistics of parental income, meant to capture different "portions" of the income distribution: The median parental income of all students in the school and the share of students with parental income in the top percentile of the national distribution.

Median Parental Income Panel (a) of Figure 7 shows the relationship between schools' median parental income and the education-innovation gap (we bin schools in 50 quantiles by median parental income and plot the average gap for schools in each bin on the vertical axis). The relationship between these two variables is negative, with a slope coefficient of -0.0001 (significant at 1 percent). This implies that a $20,000 increase in median parental income is is associated with a 0.002 smaller gap, or roughly 5 additional knowledge terms

per syllabus.

To more directly investigate how the gap varies across schools with different parental incomes we divide schools in five bins based on their median parental income, grouping together schools in the bottom 25 percent, 25-50 percent, 50-75 percent, 75-99 percent, and top 1 percent of distribution of median income across all schools. We then estimate a specification similar to equation (10), using indicators for the five median parental income bins in place of the indicators for the six school tiers.

Point estimates and confidence intervals of the coefficients in this specification, shown in panel (b) of Figure 6, show no strong differences in the gap across schools with different median parental incomes, except for schools in the top percentile and bottom quartile. Specifically, schools with median parental income in the bottom 25 percent have a gap equal to 0.959. Schools in the middle of the distribution (25 to 99 percent) have a significantly smaller gap, between 0.956 and 0.957. Schools with median parental income in the top percentile of the distribution across all schools have a much smaller gap, at 0.948. 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 45 knowledge terms, or nearly 10 percent of the total knowledge content of the average syllabus.

Share of Parents in the Top Income Percentile To capture an even more extreme measure of inequality in parental incomes across schools, we repeat our analysis using the share of parents with incomes in the top percentile as a proxy for students' background. Panel (c) of Figure 7 shows the relationship between schools' 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 -0.0572 (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 0.006 lower gap, equivalent to replacing 10 knowledge terms in the average syllabus.

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. Estimates and

confidence intervals, shown in panel (c) of Figure 6, 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 0.951 for schools where more than 15 percent of students are in the top percentile, whereas it is much larger at 0.962 for schools where less than 0.1 percent of students have parental incomes at the very top of the distribution. These estimates also imply that, in order to close the gap between schools with almost no students and with 15 percent or more students with parental incomes in the top percentile, one would have to replace approximately 20 knowledge terms to the average syllabus, or 4 percent of its content.

This pattern is confirmed by the estimates in Table 7 (columns 1-2), where the reference group are schools with less than 0.1 percent of students whose parents have incomes in the top percentile. An estimate of -0.012 for $> 15\%$ implies that, to close the difference in the gap between these schools and those where over 15 percent of students have parents with top incomes, one would have to replace 45 knowledge terms in the average syllabus, or 9.3 percent of its content. These differences appear most pronounced for Social Sciences and STEM courses (Table 7, columns 3 and 5), whereas they are close to zero for Vocational courses (column 7).

5.3 Students' Race and Ethnicity

Lastly, we investigate whether schools serving a larger portion of Black or Hispanic students have significantly different gaps for their courses. The relationship between the share of minority students in each school and the average education-innovation gap is positive, suggesting that the gap is larger in schools serving more minority students (with a slope coefficient equal to 0.0125, significant at 1 percent, Figure 7, panel (c)).

To better 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 (10), using indicators for each of these bins as independent variables. Estimates and 95 percent confidence intervals of these coefficients confirm that schools with more than 40 percent of students who are minority have a significantly larger gap, equal to 0.960. By comparison, schools with a share of minority students between 10 and 20 percent have a gap of 0.957,

and schools with less than 5 percent minority students have a gap of 0.952. These estimates imply that, in order to close the difference in the gap between schools with more than 40 percent and less than 5 percent minority students, one would have to replace 20 knowledge words in the average syllabus of the former group of schools, or 4 percent.

These patterns are confirmed by the estimates in Table 8, where we use schools with less than 5 percent minority students as the reference group. Compared with the reference group, schools with more than 40 percent minority students have a 0.011 larger gap. These estimates are robust to controlling for field-by-year fixed effects (column 2). Differently from other school characteristics, this relationship appears stable across all field macro-classes, and it is stronger for Humanities and Vocational subjects (columns 4 and 7).

6 Student Outcomes and the Education-Innovation Gap

The findings in the previous section show significant differences in access to up-to-date knowledge across schools with different selectivity and serving different populations of students. We now study whether these differences also bear a relationship with students' outcomes. We focus our attention on three outcomes: graduation rates, income, and rates of intergenerational mobility. Graduation rates and income figures are from the College Scorecard and cover years from 1997 to 2018. The measure of intergenerational mobility is taken from Chetty et al. (2019) and is available for a cross-section of students born in 1980-1982.

6.1 Graduation Rates

Figure 8 (panel (a)) shows the relationship between a school's graduation rate in a given year and the average gap across all courses for the same school in prior four years. This relationship is negative and significant: a 0.01 smaller gap is associated with a 0.44 percentage points higher graduation rate, or 1 percent compared with an average graduation rate of 46 percent across the whole time period. This correlation is robust to controlling for year fixed effects (Table 11, column 1, significant at 1 percent).

Correlation by Selectivity Tier Next, we study whether the negative relationship between graduation rates and the education-innovation gap varies across selectivity tiers. We do so by estimating the following equation:

$$\begin{aligned} Y_{st} &= \beta_1 \text{Ivy}_{s(i)} \bar{G}_{st} + \beta_2 \text{Elite}_{s(i)} \bar{G}_{st} + \beta_3 \text{SelPub}_{s(i)} \bar{G}_{st} + \beta_4 \text{SelPriv}_{s(i)} \bar{G}_{st} + \beta_5 \text{NonSel}_{s(i)} \bar{G}_{st} \\ &+ \beta_6 \text{TwoYr}_{s(i)} \bar{G}_{st} + \delta_1 \text{Ivy}_{s(i)} + \delta_2 \text{Elite}_{s(i)} + \delta_3 \text{SelPub}_{s(i)} + \delta_4 \text{SelPriv}_{s(i)} \\ &+ \delta_5 \text{NonSel}_{s(i)} + \delta_6 \text{TwoYr}_{s(i)} + \tau_t + \varepsilon_{st} \end{aligned} \tag{11}$$

where Y_{st} is the graduation rate in school s and year t, $\bar{G}_{st}$ is the average gap for courses in school s taught in the four years prior to t, and everything is as before. In this specification, the parameters $\beta_1 - \beta_6$ capture the correlations between the gap and graduation rates for schools in each tier.

Figure 9 (panel (a)) shows point estimates and 95-percent confidence intervals of the parameters $\beta_1 - \beta_6$ in equation (11). Estimates are negative for all school tiers, suggesting that a lower education-innovation gap is associated with higher graduation rates in all schools. Some differences, however, exist across tiers. For instance, an estimate of -1 for β_5 indicates that a 0.01 lower gap is associated with a one percentage point higher graduation rate for non-selective schools, or 3 percent compared with a mean rate of 36 percent for these schools. The estimate of β_1 is instead equal to 0.37, which implies that a 0.01 lower gap in Ivy Plus schools is only associated with a 0.37 percentage point (or 0.4 percent) higher rate in these schools. These patterns are confirmed by the estimates in Table 9 (column 2).

Correlation by Parental Income In Figure 10 (panel (a)) we investigate whether the relationship between the gap and graduation rates varies across schools serving students with different parental incomes. To do so we re-estimate equation (11), substituting the indicators for school tiers with indicators for a school's share of students with parental income in the top percentile. These estimates indicate that the correlation between the gap and graduation rates is small and insignificant for schools serving more than 5 percent of students with parental incomes in the top percentile, and it becomes progressively more negative as the share of students with top parental incomes declines. In particular, it is equal to -0.4 for schools with 1 to 5 percent of students with parental income at the top and to -1 for schools

with less than 0.1 percent of students with parents' incomes in the top percentile. This last estimate implies that a 0.01 lower gap is associated with a 1 percentage point higher graduation rate in these schools (or 3 percent percent compared with an average rate of 35 percent). These findings are summarized in Table 9 (column 3).

6.2 Students' Incomes

Next, we study whether the gap is related to the income of students who graduate from each school. Figure 8 (panel (b)) shows a negative relationship between a school's average gap over four years and the median income of the students who graduate from that school at the end of the four years, measured ten years after graduation. An estimated slope of -29.03 indicates that a 0.01 lower gap is associated with a $290 higher income, or 0.6 percent compared with an average income of $45,754.

As before, we explore whether this negative relationship varies across selectivity tiers. We do so by re-estimating equation (11) using the natural logarithm of student income as the dependent variable. The point estimates and confidence intervals of the parameters in this equation, shown in panel (b) of Figure 9, indicate that the relationship between the gap and students' incomes is negative and statistically significant in all schools (except for Ivy Plus schools), but it is significantly higher in Elite schools. These estimates imply, for instance, that a 0.01 lower gap is associated with a 2 percent higher income rate for Elite schools and with a one percent higher income in non-selective schools (with estimated correlations equal to -2.1 and -0.96, respectively). The correlation is istead smaller for highly selective and selective schools, and for Ivy Plus. These patterns are confirmed by the estimates in Table 10 (column 2).

Next, we investigate whether the relationship between the gap and students' incomes is different across schools serving students with different parental incomes. We re-estimate equation (11) using the natural logarithm of income as the dependent variable and indicators for each school's share of students with parental income in the top percentile as the explanatory variables. Estimates of this equation, shown in panel (b) of Figure 10, indicate that the correlation between the gap and students' incomes is small and insignificant for schools serving more than 5 percent of students with parents' incomes in the top percentile,

whereas it is much larger for schools with fewer students in the top parental income percentile. In schools where virtually no students have parental income in the top percentile, the correlation is -4; this estimate implies that a 0.01 lower gap is associated with a four percent higher income. By comparison, this correlation is equal to -1 for schools with 1 to 5 percent of students with top parental income. These findings are summarized in Table 10 (column 3).

6.3 Gap and Intergenerational Income Mobility

The last outcome we investigate is intergenerational mobility, a measure of equality of economic opportunity across students from different parental backgrounds. Following Chetty et al. (2019), we define mobility as the probability that a student with parental income in the bottom quintile reaches the top income quintile during adulthood.

Figure 8 (panel (c)) shows a binned scatterplot of this measure of mobility, measured using 2014 incomes for students who graduated between 2002 and 2004, and the education-innovation gap, calculated for each school using data for the years 1998 and 2004 (i.e., the first four years of college for these cohorts). The relationship between these two variables is negative and significant: A 0.01 lower gap is associated with a 0.1 percentage points lower probability of reaching the top quintile, or, nearly 0.5 percent compared with an average probability of 23 percent.

In panel (c) of Figure 9 we estimate this correlation separately by school tiers, with the strategy employed in the previous subsection. In Ivy Plus schools, a 0.01 reduction in the gap is associated with a 0.5 percentage point higher chance that students from the bottom parental quintile reach the top quintile in adulthood, or 1 percent compared with an average of 50 percent for students in these schools. The same estimate is 2.4 percent for Elite schools (with an estimated correlation of -1.2). By comparison, a 0.01 reduction in the gap is associated with an increase in mobility between 0.002 and 0.005 percent in selective annd non-selective schools. These findings are summarized in Table 11 (column 2).

In panel (c) of Figure 10 we also investigate whether the relationship between the gap and intergenerational mobility is different across schools serving students with different parental incomes. The correlation is small and insignificant for schools with a higher share

of students with parental income in the top percent, and it becomes progressively more negative as this share declines. For example, a 0.01 reduction in the gap is associated with a 0.9 percentage point increase in intergenerational mobility in schools with less than 0.1 percent of students with parental income in the top percentile (or 6 percent). These findings are summarized in Table 11 (column 3).

6.4 Summing Up

Taken together, our results indicate a robust relationship between the education-innovation gap and students' future outcomes. Importantly, this relationship differs across schools belonging to different selectivity tiers and serving different populations of students. While the patterns across schools with different selectivity are mixed (for example, we find that the correlation between the gap and graduation rates is strongest in non-selective schools, whereas the correlation with income and intergenerational mobility is strongest in Elite schools), we consistently find that the relationship between the gap and all outcomes at study is the strongest in schools serving students from more disadvantaged backgrounds, whereas it is almost zero in institutions serving a wealthier population. While not enough to establish a causal link, these patterns do suggest that the content of higher education courses might be especially important for students who lack a strong socio-economic background, perhaps because they have to rely more strongly on the knowledge they acquire in schools in order to succeed in the labor market.

7 Conclusion

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

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

References

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

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

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

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

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

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

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.

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.

Kane, Thomas J, and Douglas O Staiger, 2008, Estimating teacher impacts on student achievement: An experimental evaluation, Technical report, National Bureau of Economic Research.

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.

Romer, Paul M, 1994, The origins of endogenous growth, *Journal of Economic perspectives* 8, 3–22.

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.

Table 1: Summary Statistics

Panel a): Counts of Syllabi

	All years	1992	2000	2010	2018
syllabi per year	91,407	944	37,106	98,003	271,955
	(92,929)				
syllabi per student per year	0.0109	0.0005	0.0051	0.0123	0.0273
	(0.0562)	(0.0016)	(0.0217)	(0.0576)	(0.1071)
syllabi per FTE instructor per year	0.416	0.0140	0.1320	0.4889	0.9896
	(2.815)	(0.0489)	(1.0533)	(3.8518)	(5.0690)
fields per year	60.92	56	62	62	62
	(2.25)				

Panel b): Gap, means and standard deviations, by year

Year	Mean	St. dev.	Min	Max	N
1993	0.983	0.0535	0.6794	1.318	996
2001	0.9438	0.0566	0.4656	3.1387	37,106
2010	0.9428	0.0487	0.5571	1.4882	98,003
2017	0.9723	0.0456	0.2231	1.5499	271,955

Panel c): Gap, means and standard deviations, by field (most frequent fields)

Field	Mean	St. dev.	Min	Max	N
Mathematics	0.9797	0.0456	0.2968	3.1387	214,912
English Literature	0.9518	0.0485	0.5105	1.4916	186,809
Business	0.9494	0.0425	0.4568	1.3184	158,420
Computer Science	0.9408	0.0461	0.2231	1.7888	138,639
Education	0.9314	0.0529	0.2187	1.4162	122,967

Note: Means and standard deviations (in parentheses) of key summary statistics.

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	0.0059	(0.0022)	-0.0010	(0.0046)
Other elite	0.0154	(0.0070)	0.0128	(0.0057)
Highly selective-selective public	0.0442	(0.0186)	0.0384	(0.0177)
Highly selective-selective private	0.0320	(0.0237)	0.0116	(0.0065)
Non-selective	0.0003	(0.0003)	0.0003	(0.0003)
Two-year	0.0170	(0.0140)	0.0078	(0.0149)

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)	-0.0021	(0.0058)	-0.0021	(0.0057)
ln Endowment per capita (2000)	0.0061	(0.0068)	0.0053	(0.0068)
ln Sticker price (2013)	-0.0004	(0.0090)	-0.0032	(0.0076)
ln Avg faculty salary (2013)	0.0338	(0.0202)	0.0398	(0.0191)
ln Enrollment (2013)	0.0068	(0.0058)	0.0102	(0.0051)
Share Black students (2000)	-0.0327	(0.0274)	-0.0532	(0.0277)
Share Hispanic students (2000)	0.0670	(0.0684)	0.0807	(0.0677)
Share alien students (2000)	0.1803	(0.2202)	0.2339	(0.1945)
Share grad in Arts & Humanities (2000)	0.0003	(0.0006)	0.0002	(0.0005)
Share grad in STEM (2000)	-0.0002	(0.0005)	0.0001	(0.0004)
Share grad in Social Sciences (2000)	0.0002	(0.0005)	0.0001	(0.0004)

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	Share of explained variation
Baseline (Year FE)	0.05	.
Institution and Year	0.10	0.11
Field and Year	0.13	0.14
Institution, Field and Year	0.26	0.28
Course and Year	0.63	0.66
Instructor and Year	0.59	0.62
Instructor, Course, and Year	0.82	0.86

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 "baseline" regression. Each observation corresponds to a course, instructor, and year.

Table 4: Gap and Instructor Changes. OLS Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	(1)	(2)	(3)
share of new instructors	-0.0006***	-0.0007***	-0.0006***
	(0.0002)	(0.0002)	(0.0002)
Course FE	Yes	Yes	Yes
Year FE	Yes	No	No
Inst. x year FE	No	Yes	No
Field x year FE	No	No	Yes
N (Course x year)	409682	404857	408881
# Courses	112366	110804	112273

Note: OLS estimates of an empirical model of the education-innovation gap as a function of the share of instructors who are new for each course in a given year, controlling for course and year fixed effects. In column 2 we also control for institution-specific year fixed effects, and in column 3 we control for field-specific year effects. Each observation corresponds to a course in a given year. Standard errors are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 5: Gap and Instructor Changes. OLS Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	CS	Eng	Math	Business	Econ	Psych	English Lit	History
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
share of new instructors	-0.0010	-0.0004	-0.0012**	-0.0008	-0.0028***	-0.0015**	0.0020**	-0.0032***
	(0.0008)	(0.0009)	(0.0006)	(0.0006)	(0.0010)	(0.0007)	(0.0008)	(0.0012)
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inst. x year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (Course x year)	14959	9525	26357	22059	9428	16391	17899	9250
# Courses	4457	3046	7140	6527	2437	4454	5236	2841

Note: OLS estimates of an empirical model of the education-innovation gap as a function of the share of instructors who are new for each course in a given year, controlling for course and year fixed effects. Each column corresponds to a field. Each observation corresponds to a course in a given year. Standard errors are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 6: Gap and School Tiers. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	All fields		STEM	Humanities	Social Sciences	Business	Vocational
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School tier							
Ivy Plus/Elite	-0.012***	-0.010***	-0.008***	-0.008***	-0.015***	-0.008**	-0.004
	(0.002)	(0.001)	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)
Highly selective/selective	-0.007***	-0.005***	-0.004**	-0.004*	-0.008***	-0.002	-0.007*
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)
Syll. year FE	Yes	No	No	No	No	No	No
Field*year FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Mean of dep. var. for non-selective	0.97	0.97	0.97	0.97	0.97	0.97	0.97
N	1713501	1713476	574461	425531	448347	169644	93602
# Inst-by-field	34413	34409	10138	11009	7990	2303	2828

Note: The table shows OLS estimates of the coefficients for indicators for school tiers (*Ivy Plus/Elite, Highly selective/selective*) on the gap between syllabi and publications. The omitted category is *Non-selective* institutions. 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 1 to 3 years prior. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, while columns 2-7 controls for field-by-year fixed effects. Columns 3-7 show estimates by macro-field class. Standard errors in parentheses are clustered at the school-by-field level. Information on school tiers is taken from Chetty et al. (2019). $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 7: Gap and Schools' Median Parental Income. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	All fields		STEM	Humanities	Social Sciences	Business	Vocational
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.1-1%	-0.005***	-0.004**	-0.004*	0.001	-0.009**	-0.003	-0.005
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)	(0.005)
1-5%	-0.009***	-0.005***	-0.009***	0.006	-0.013***	-0.005	-0.002
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)	(0.006)
5-15%	-0.012***	-0.010***	-0.012***	-0.003	-0.016***	-0.009**	-0.003
	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.008)
$>15\%$	-0.012***	-0.010***	-0.014***	0.002	-0.021***	-0.010**	0.003
	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.006)
Syll. year FE	Yes	No	No	No	No	No	No
Field*year FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Dep. var. mean, share in top1 $<0.1\%$	0.97	0.97	0.98	0.96	0.96	0.96	0.96
N	2410858	2410842	832290	594569	554739	215474	209221
# Inst-by-field	46202	46200	13892	14246	10472	3188	4179

Note: The table shows OLS estimates of the coefficients for the share of parents in the top percentile of the income distribution on the gap between syllabi and publications. The omitted category is $=0$. 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 1 to 3 years prior. Percentiles of parental incomes are calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, while columns 2-7 controls for field-by-year fixed effects. Columns 3-7 show estimates by macro-field class. Standard errors in parentheses are clustered at the school-by-field level. Information on parental income is taken from Chetty et al. (2019). $^{*}\leq 0.1$, $^{**}\leq 0.05$, $^{***}\leq 0.01$.

Table 8: Gap and Students' Race/Ethnicity. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	All fields		STEM	Humanities	Social Sciences	Business	Vocational
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share minority students							
5-10%	0.002	0.003***	-0.001	0.005***	0.003	0.001	0.018**
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.008)
10-20%	0.003	0.004***	0.002	0.005**	0.003	0.006**	0.014*
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.008)
20-40%	0.006***	0.007***	0.006***	0.008***	0.004*	0.006**	0.020***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.002)	(0.008)
$>40\%$	0.011***	0.010***	0.006***	0.013***	0.011***	0.001	0.020***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.004)	(0.008)
Syll. year FE	Yes	No	No	No	No	No	No
Field*year FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Mean of dep. var., share $<5\%$	0.95	0.95	0.96	0.95	0.94	0.95	0.94
N	2434201	2434186	847869	590692	561614	218611	210874
# Inst-by-field	45606	45604	13763	14033	10319	3131	4134

Note: The table shows OLS estimates of the coefficients for indicators for intervals of the share of students who are either Black or Hispanic in each school. The omitted category is schools with $<5\%$ Black or Hispanic students. 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 1 to 3 years prior. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, while columns 2-7 controls for field-by-year fixed effects. Columns 3-7 show estimates by macro-field class. Standard errors in parentheses are clustered at the school-by-field level. Information on students' race and ethnicity at the school level is taken from Chetty et al. (2019). $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 9: Graduation Rate and the Education-Innovation Gap

	Dep. Var: Graduation rate		
	(1)	(2)	(3)
Gap	-0.658***		
	(0.188)		
Gap * Ivy Plus		-0.293*	
		(0.168)	
Gap * Other elite		-0.526**	
		(0.229)	
Gap * (Highly) Selective Private		-0.334**	
		(0.138)	
Gap * (Highly) Selective Public		-0.549**	
		(0.243)	
Gap * Non-selective		-0.975***	
		(0.368)	
Gap * $<$ 0.1%			0.178
			(0.608)
Gap * 0.1-1%			0.110
			(0.164)
Gap * 1-5%			-0.348*
			(0.198)
Gap * 5-15%			-0.887**
			(0.402)
Gap * $>$15%			-1.014**
			(0.481)
Syll. year FE	Yes	Yes	Yes
Mean of dep. var.	0.56	0.56	0.56
N	15162	15162	15162

Note: The table shows OLS estimates of specifications where the dependent variable is a school's graduation rate. 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 1 to 3 years prior. The variables *Ivy Plus, Other elite, (Highly) selective public, (Highly) selective private*, and *Non-selective* denote school tiers. The variables $<$ *0.1%*, *0.1-1%*, *1-5%*, *5-15%* , and $>$ *15%* denote the shares of students with parental income in the top one percent of the distribution in each school. Estimates are obtained controlling for year fixed effects and either school tier fixed effects (column 2) or indicators for the share of students with parental income in the top percentile (column 3). Standard errors are clustered at the school-by-field level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 10: Student Income and the Education-Innovation Gap

	Dep. Var: ln(Income)		
	(1)	(2)	(3)
Gap	-1.271***		
	(0.375)		
Gap * Ivy Plus		-0.635	
		(1.647)	
Gap * Other elite		-2.155	
		(1.390)	
Gap * (Highly) Selective Private		-0.515**	
		(0.241)	
Gap * (Highly) Selective Public		-0.705	
		(0.491)	
Gap * Non-selective		-0.894	
		(0.549)	
Gap * $< 0.1\%$			-0.996
			(1.034)
Gap * 0.1-1%			0.184
			(0.337)
Gap * 1-5%			-1.015***
			(0.351)
Gap * 5-15%			-1.625
			(1.456)
Gap * >15%			-4.218*
			(2.534)
Syll. year FE	Yes	Yes	Yes
Mean of dep. var.	4004	4004	4004

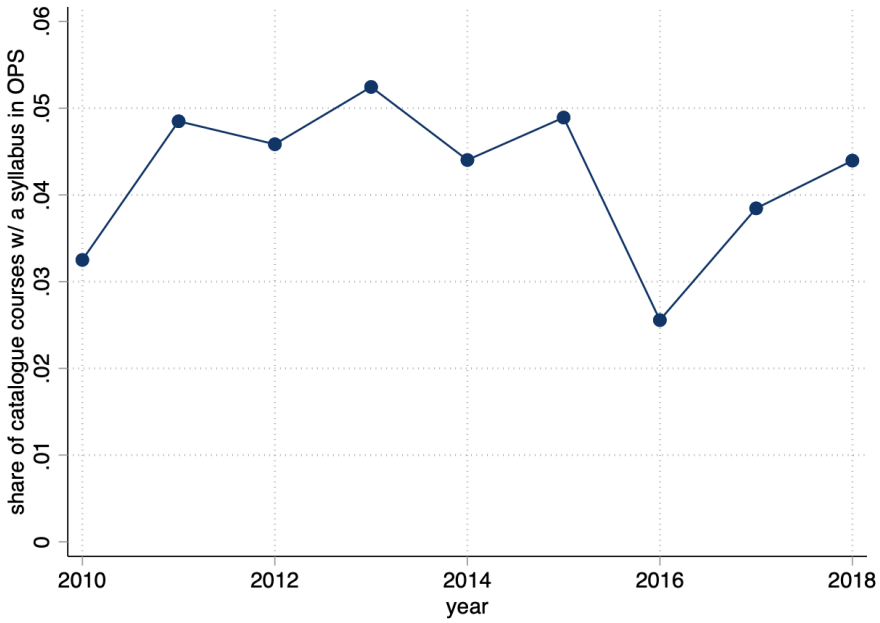
Note: The table shows OLS estimates of specifications where the dependent variable is a school's student earnings (measured ten years after entering college). The *Gap* is defined as the ratio between syllabi's cosine similarity with publications 13 to 15 years prior to the syllabus date and the cosine similarity with publications 1 to 3 years prior. The variables *Ivy Plus, Other elite, (Highly) selective public, (Highly) selective private*, and *Non-selective* denote school tiers. The variables $< 0.1\%$, *0.1-1%*, *1-5%*, *5-15%* , and $> 15\%$ denote the shares of students with parental income in the top one percent of the distribution in each school. Estimates are obtained controlling for year fixed effects and either school tier fixed effects (column 2) or indicators for the share of students with parental income in the top percentile (column 3). Standard errors are clustered at the school-by-field level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 11: Intergenerational Mobility and the Education-Innovation Gap

	Dep.Var: Mobility Rate		
	(1)	(2)	(3)
Gap	-0.310***		
	(0.041)		
Sorted by School Tiers			
Gap * Ivy and Other Elite		-0.526***	
		(0.153)	
Gap * Highly selective		-1.145***	
		(0.142)	
Gap * Selective Private		-0.140**	
		(0.059)	
Gap * Selective Public		-0.200**	
		(0.079)	
Gap * Non-selective		-0.470***	
		(0.091)	
Sorted by Share of Parents in Top 1 Percent			
Gap * $< 0.1\%$			-1.618***
			(0.350)
Gap * 0.1%,1%			-0.569***
			(0.061)
Gap * 1%,5%			-0.298***
			(0.069)
Gap * 5%,15%			-0.050
			(0.087)
Gap * $> 15\%$			0.111
			(0.135)
N	15,279	15,279	15,279

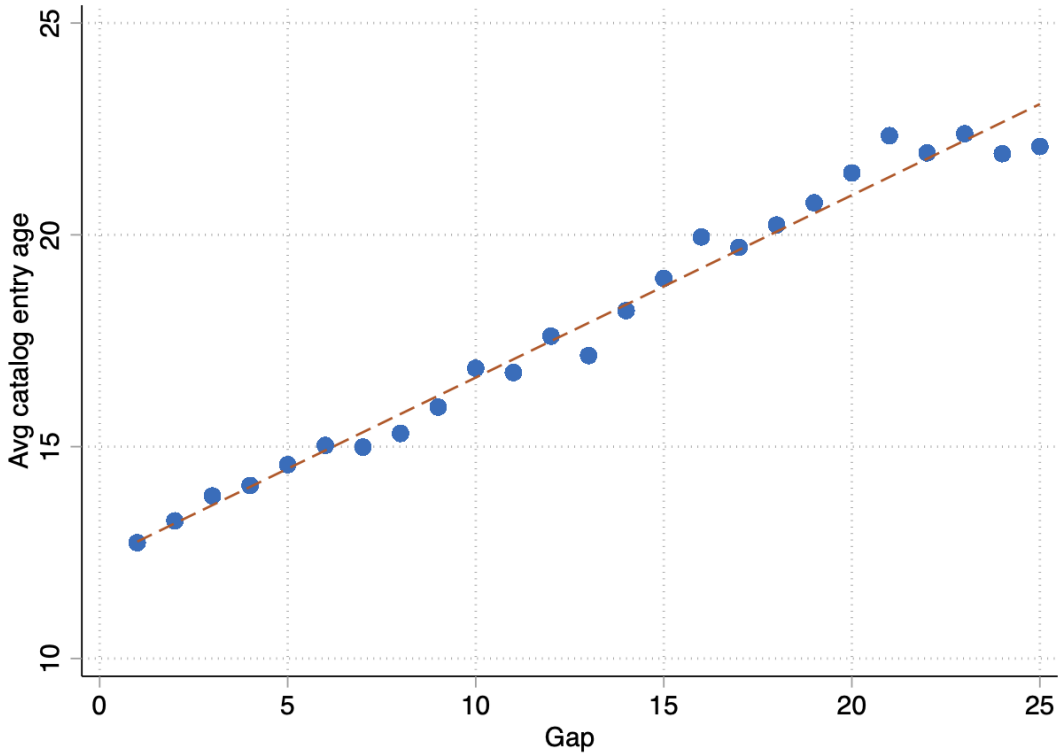
Note: The table shows OLS estimates of specifications where the dependent variable is a school's intergenerational mobility, defined as the probability that a student with parental income in the bottom quartile of the distribution reaches the top income quartile during adulthood. The *Gap* is defined 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 1 to 3 years prior. The variables *Ivy Plus, Other elite, (Highly) selective public, (Highly) selective private*, and *Non-selective* denote school tiers. The variables $<$ *0.1%, 0.1-1%, 1-5%, 5-15%* , and $>$ *15%* denote the shares of students with parental income in the top one percent of the distribution in each school. Estimates are obtained controlling for year fixed effects and either school tier fixed effects (column 2) or indicators for the share of students with parental income in the top percentile (column 3). Standard errors are clustered at the school-by-field level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Figure 1: Selection Into the OSP: Share of Covered Syllabi, Catalogue Data



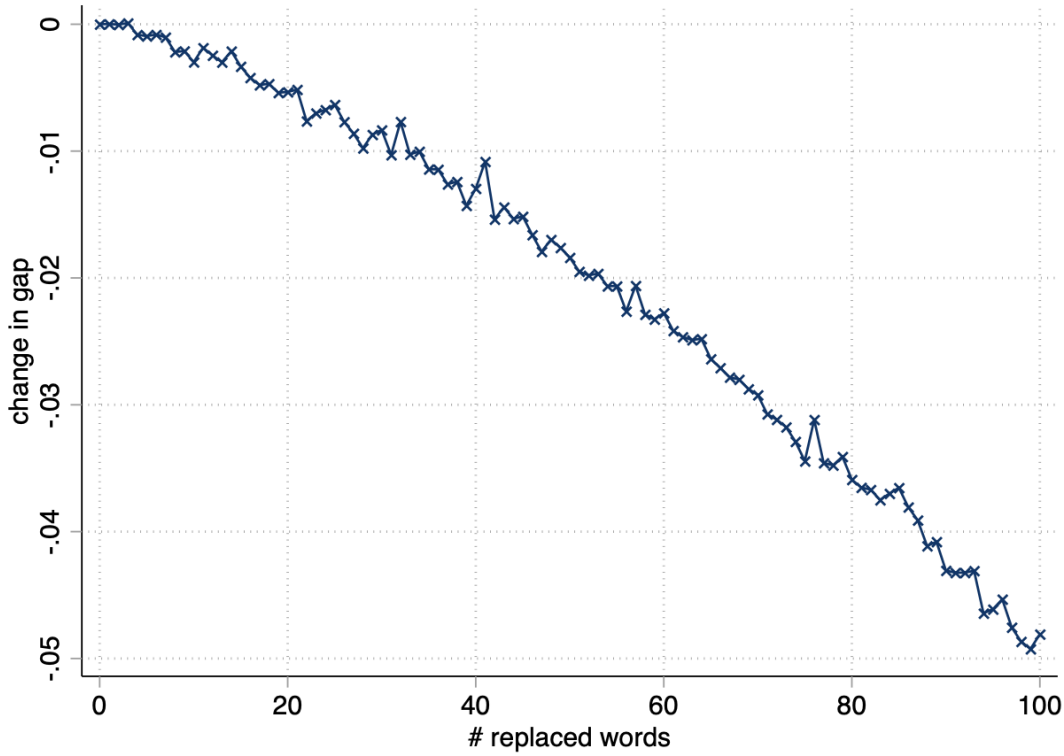
Note: Share of syllabi from the full catalogue of 161 selected institutions that are included in our sample. Catalogue data are collected from university archives.

Figure 2: Validating The Gap With Syllabi References



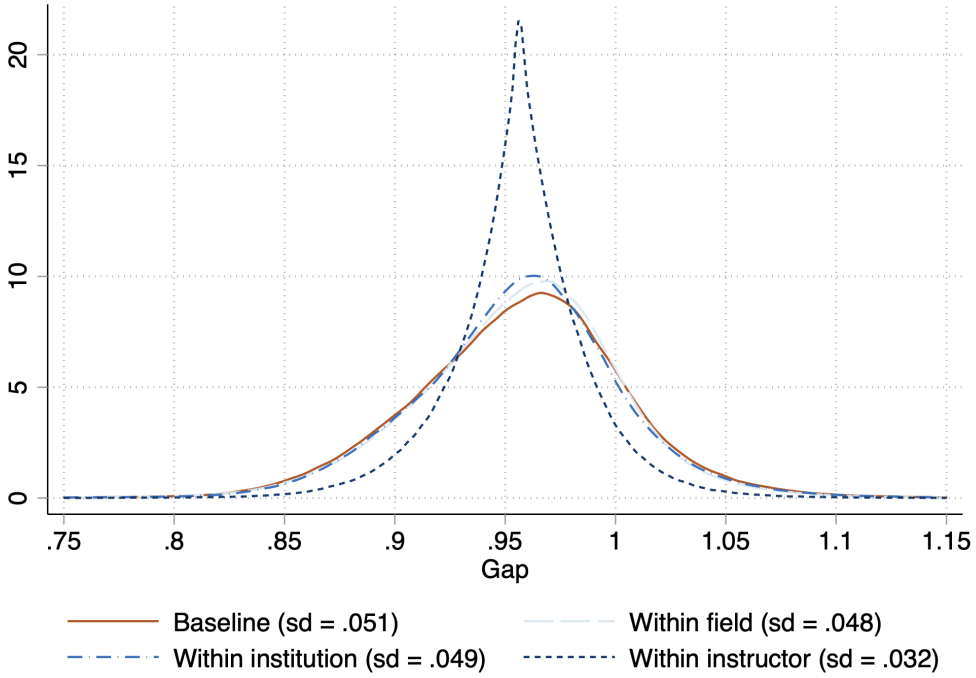
Note: This figure shows the correlation between the gap and the reference age of each syllabus. The reference age is defined as the average difference between the year of the syllabus and the year of each reference listed in the syllabus as a required or recommended reading. We divide syllabi in 25 equally-sized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

Figure 3: Economic Magnitude of Changes in The Education-Innovation Gap



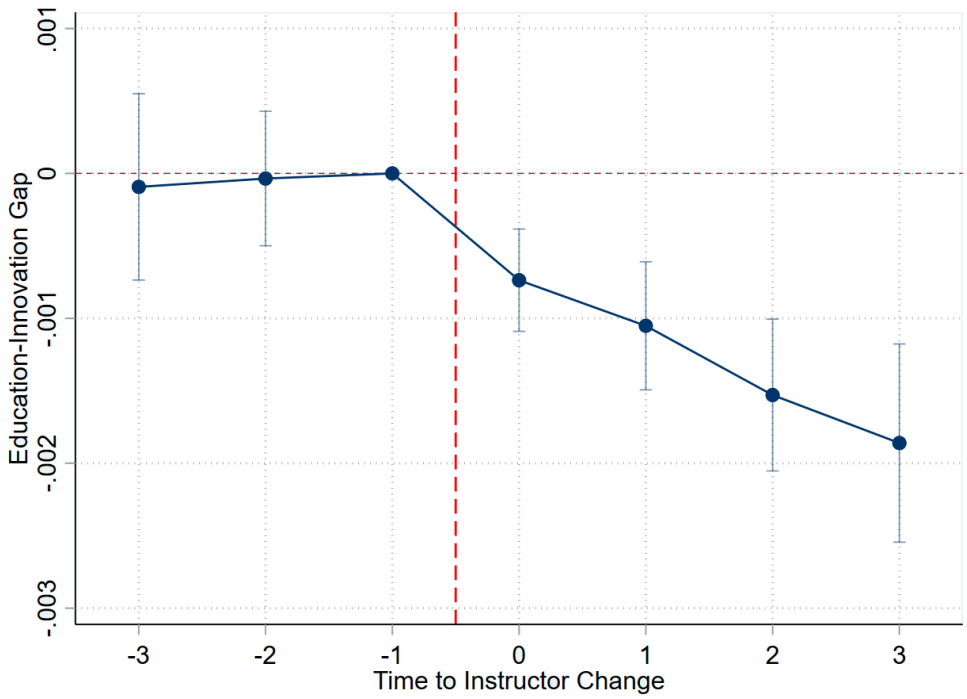
Note: This figure links the differences in education-innovation gap with the associated number of "knowledge words" that must be replaced with newer words in each syllabus. We obtain this relationship by a) randomly choosing 100,000 syllabi from the sample, b) replacing a varying number of "old" knowledge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

Figure 4: Education-Innovation Gap: Variation



Notes: The figure shows the distribution of the gap, overall (dashed line) and within attributes: field, institution, course, and instructor.

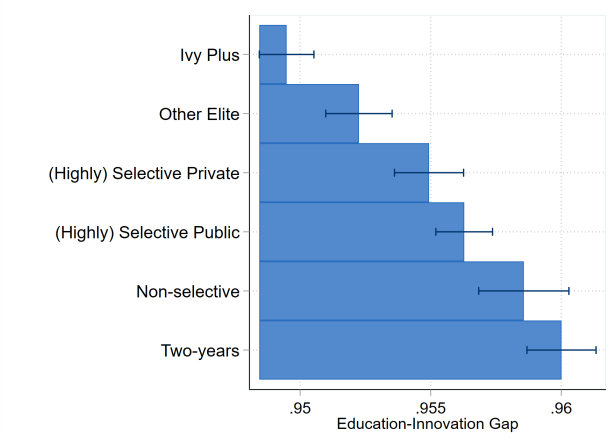
Figure 5: Education-Innovation Gap Around The Time of An Instructor Change



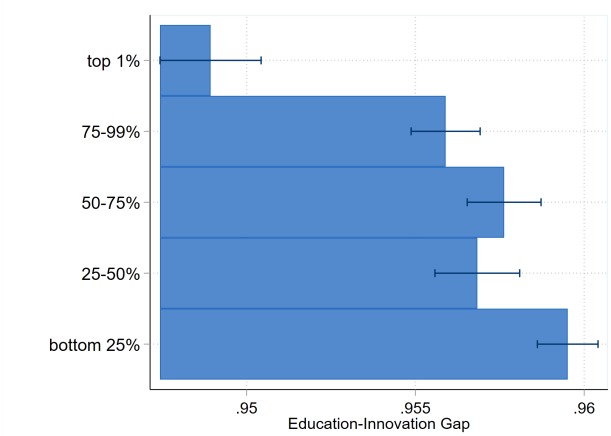
Notes: OLS estimates of the coefficients of education-innovation gap changes around instructor changes, as specified in equation (9) of the paper.

Figure 6: Gap with Publications, By School Characteristics

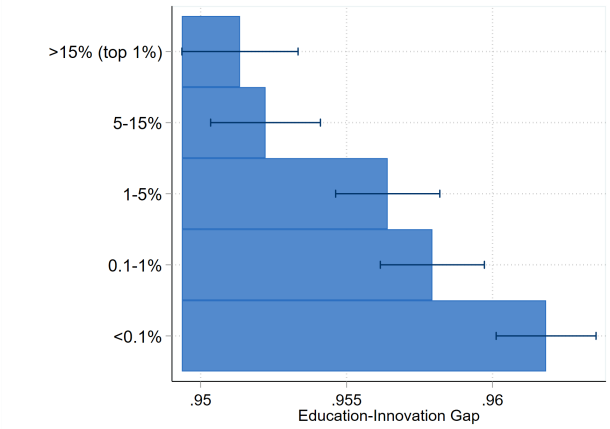
(a) By school tier



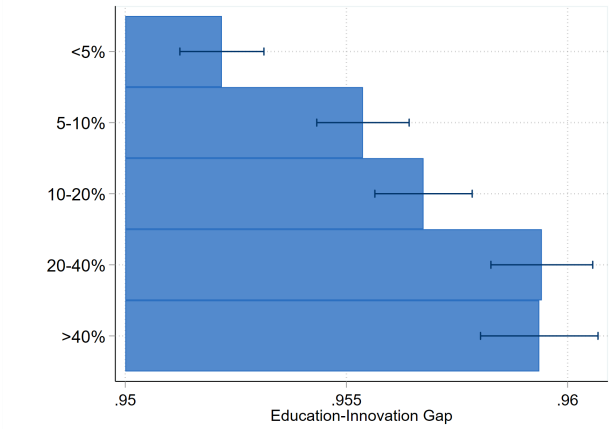
(b) By percentile of parental income



(c) By % of parents in top income percentile



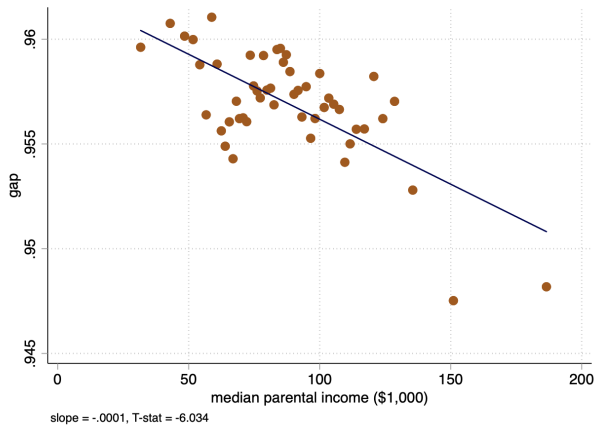
(d) By % of minority students (Black/Hispanic)



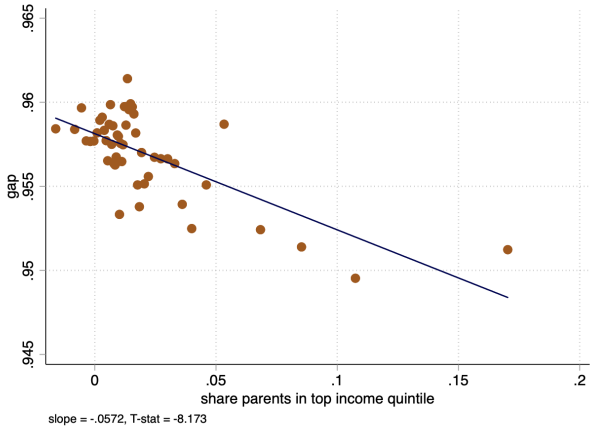
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by school tier (panel a), percentile of median parental income in the school (panel b), share students with parents in the top income percentile in the school (panel c), and share of students who are either Black or Hispanic (panel d). The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two 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 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on school tiers, parental income, and students' race and ethnicity is taken from Chetty et al. (2019).

Figure 7: Gap with Publications And Schools' Characteristics

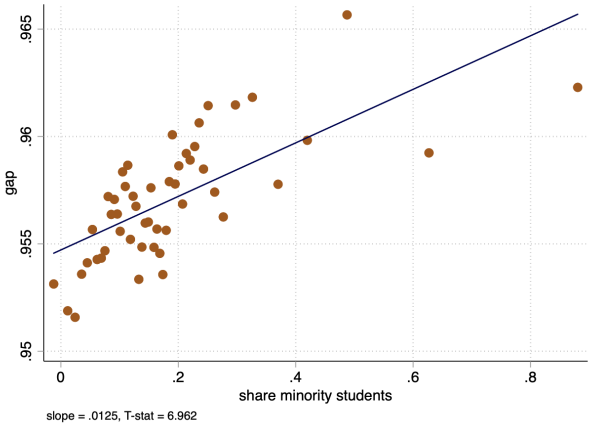
(a) Percentile of parental income



(b) % of parents in top income percentile



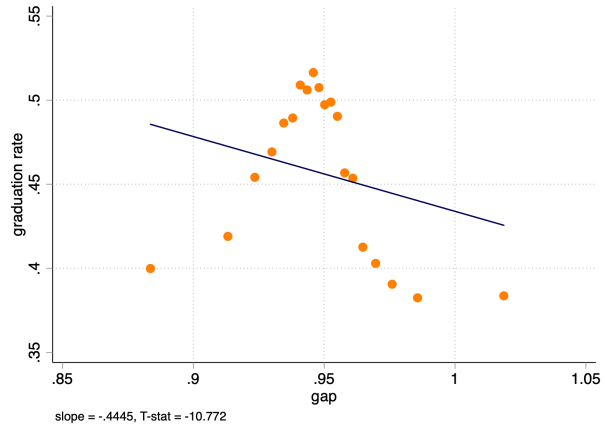
(c) By % of minority students (Black/Hispanic)



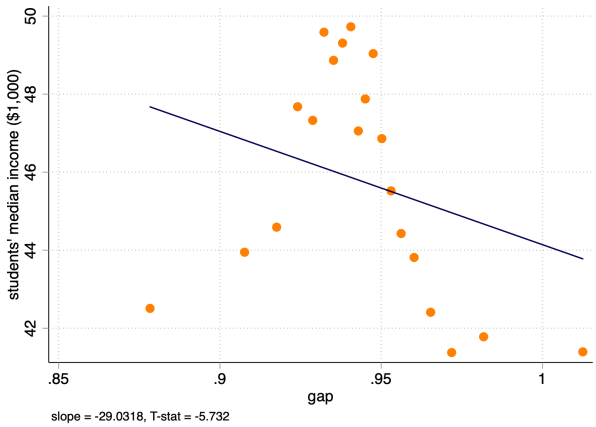
Notes: The figure shows a binned scatterplot of the gap between syllabi and publications (vertical axis) and school characteristics (horizontal axis), including the percentile of median parental income in the school (panel a), the share of students with parents in the top income percentile (panel b), and the hare of students who are either Black or Hispanic (panel c). The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two 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 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on parental income and students' race and ethnicity is taken from Chetty et al. (2019).

Figure 8: Gap with Publications And Students' Outcomes

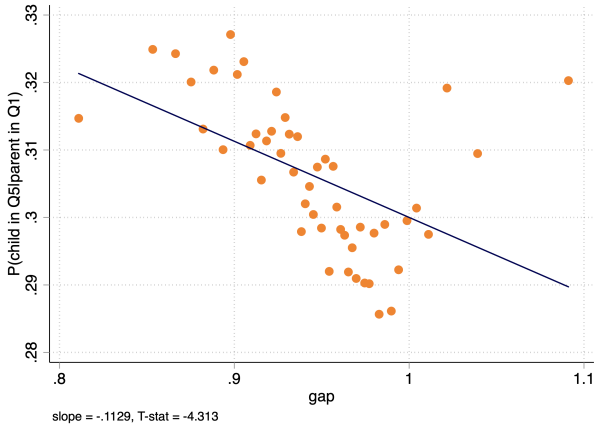
(a) Graduation rates



(b) Income



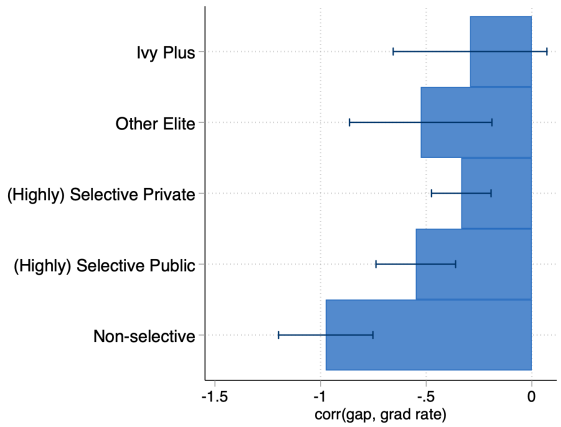
(c) Intergenerational mobility



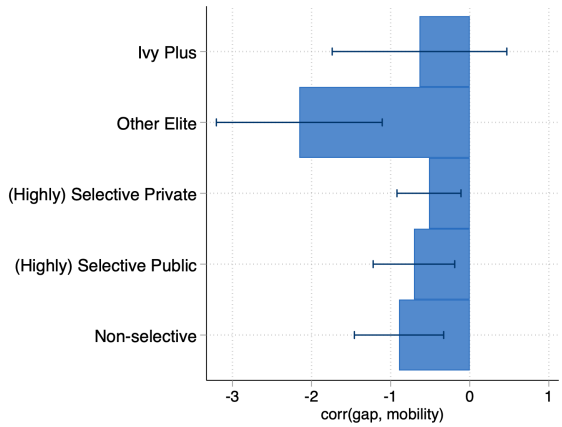
Notes: The figure shows a binned scatterplot of the gap between syllabi and publications (horizontal axis) and either graduation rates (panel (a)), student income ten years after graduation (panel (b)), and intergenerational mobility (panel (c)). The gap is defined as the ratio between syllabi's cosine similarity with publications 13 to 15 years prior to the syllabus date and the cosine similarity with publications 1 to 3 years prior. In panel (a), graduation rates refer to the years 1997 to 2018, and the gap is calculated based on the four years prior to the graduation year. In panel (b), incomes refer to the years 1997 to 2018, and the gap is calculated based on the fourteen to ten years prior to the income year. In panel (c), intergenerational mobility estimates are obtained using information on students who graduated between 2002 and 2004, and the gap is calculated using data from 1999 to 2002.

Figure 9: Gap with Publications And Students' Outcomes - By selectivity tier

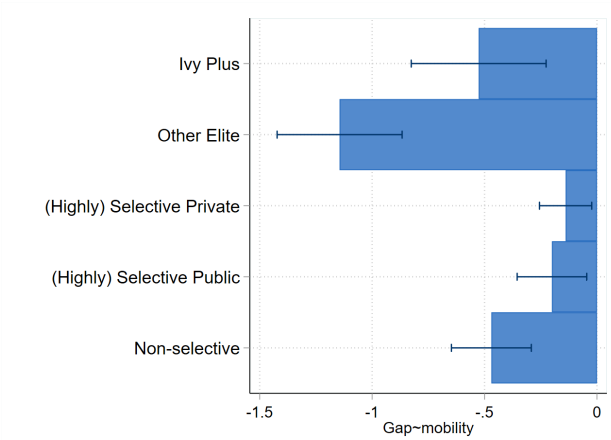
(a) Graduation rates



(b) Income

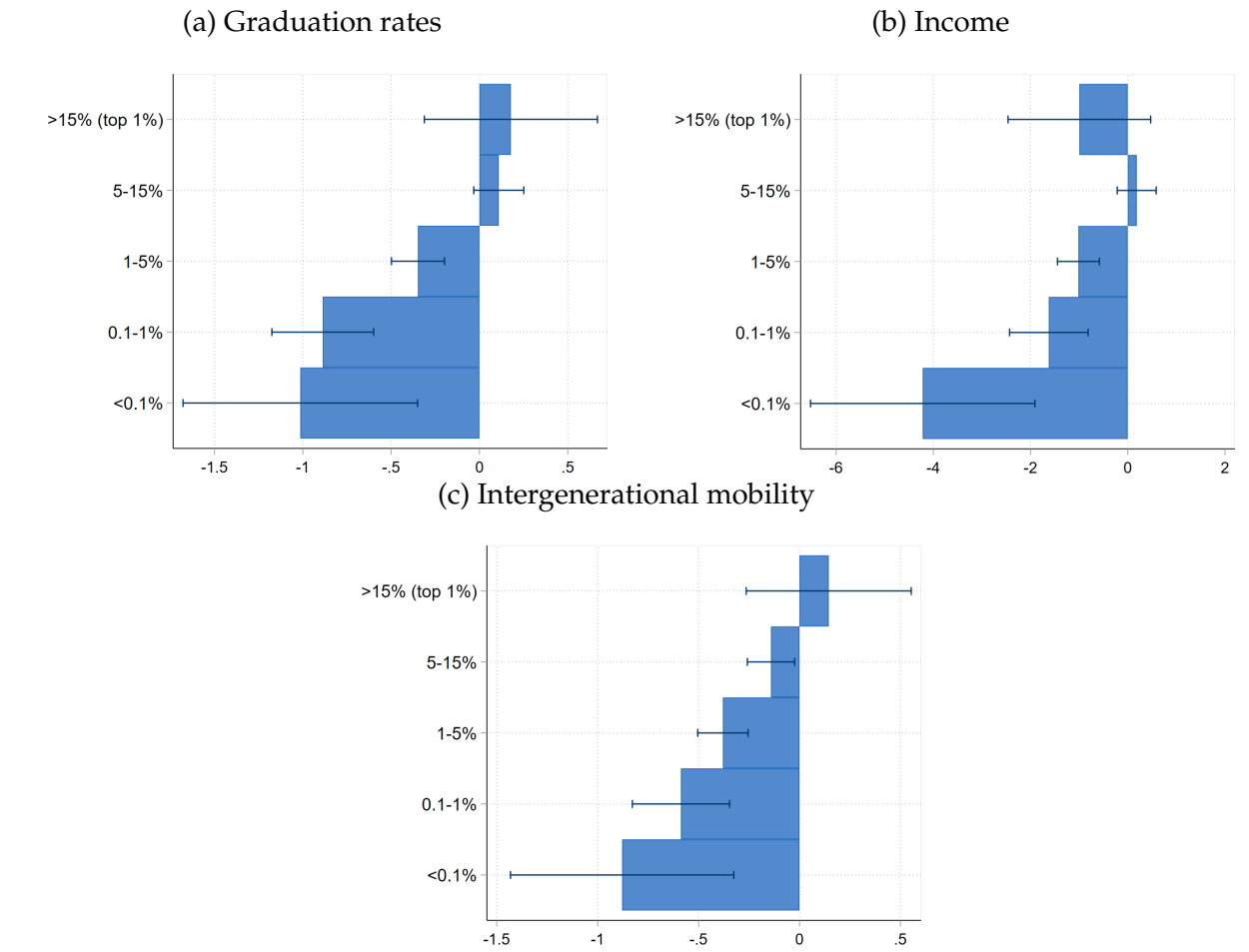


(c) Intergenerational mobility



Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between the education-innovation gap and graduation rates (panel (a)), students' incomes (panel (b)), and intergenerational mobility (panel (c)) separately by school tier. 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 1 to 3 years prior. In panel (a), graduation rates refer to the years 1997 to 2018, and the gap is calculated based on the four years prior to the graduation year. In panel (b), incomes refer to the years 1997 to 2018, and the gap is calculated based on the fourteen to ten years prior to the income year. In panel (c), intergenerational mobility estimates are obtained using information on students who graduated between 2002 and 2004, and the gap is calculated using data from 1999 to 2002.

Figure 10: Gap with Publications And Students' Outcomes - By share of students with parents in top income percentile



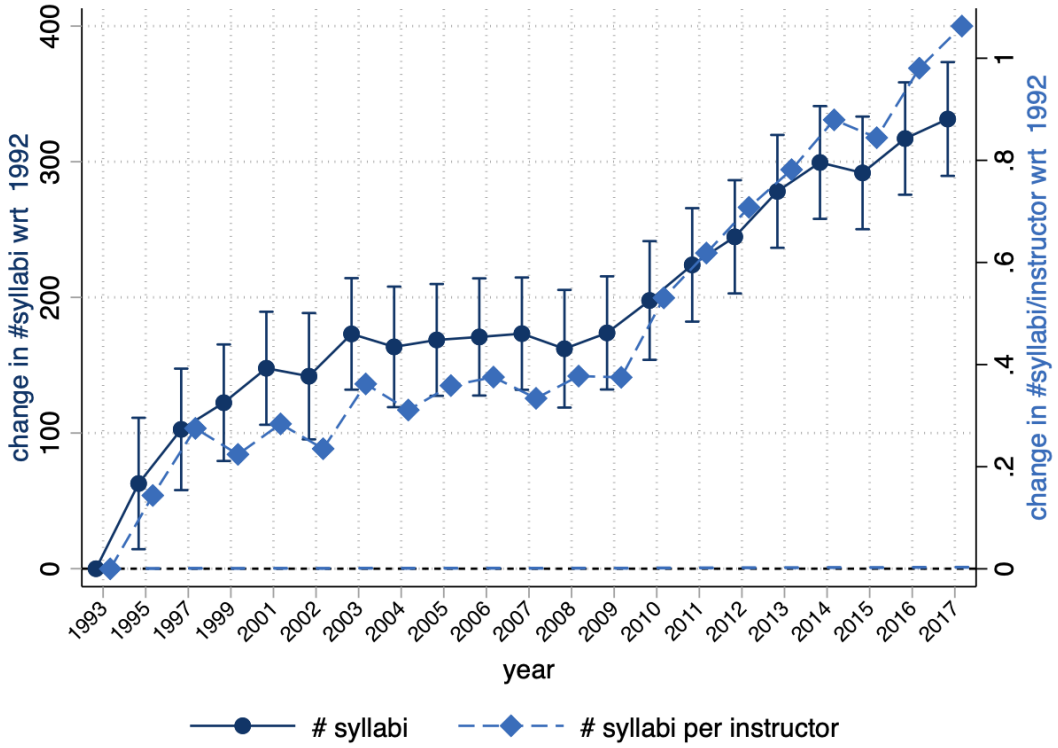
Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between the education-innovation gap and graduation rates (panel (a)), students' incomes (panel (b)), and intergenerational mobility (panel (c)) separately by the share of students with parents in the top income percentile. 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 1 to 3 years prior. In panel (a), graduation rates refer to the years 1997 to 2018, and the gap is calculated based on the four years prior to the graduation year. In panel (b), incomes refer to the years 1997 to 2018, and the gap is calculated based on the fourteen to ten years prior to the income year. In panel (c), intergenerational mobility estimates are obtained using information on students who graduated between 2002 and 2004, and the gap is calculated using data from 1999 to 2002.

Appendix

For online publication only

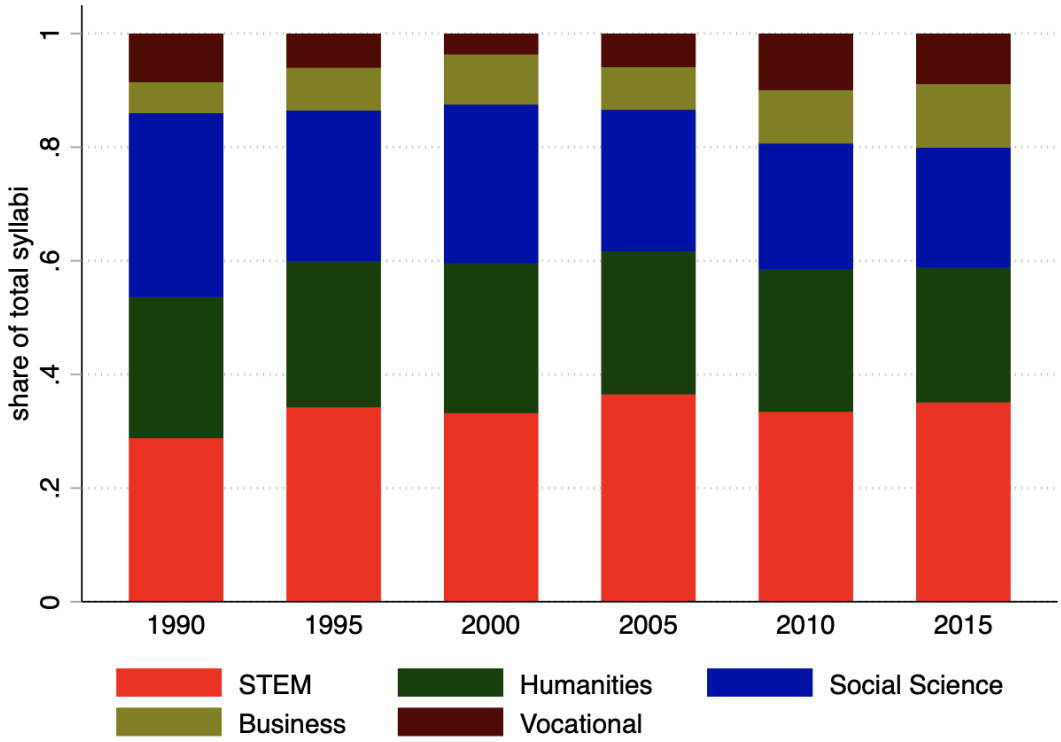
Additional Tables and Figures

Figure AI: 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.

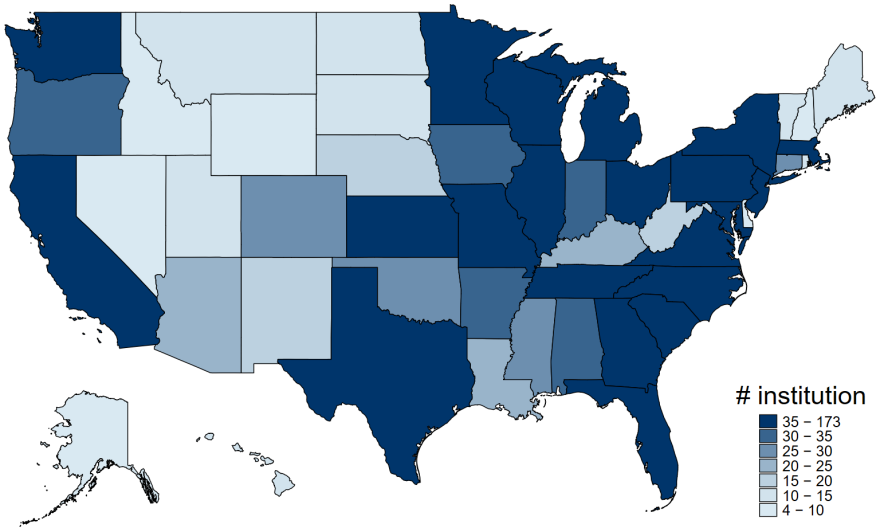
Figure AII: Stable Field Coverage of the Syllabi Data



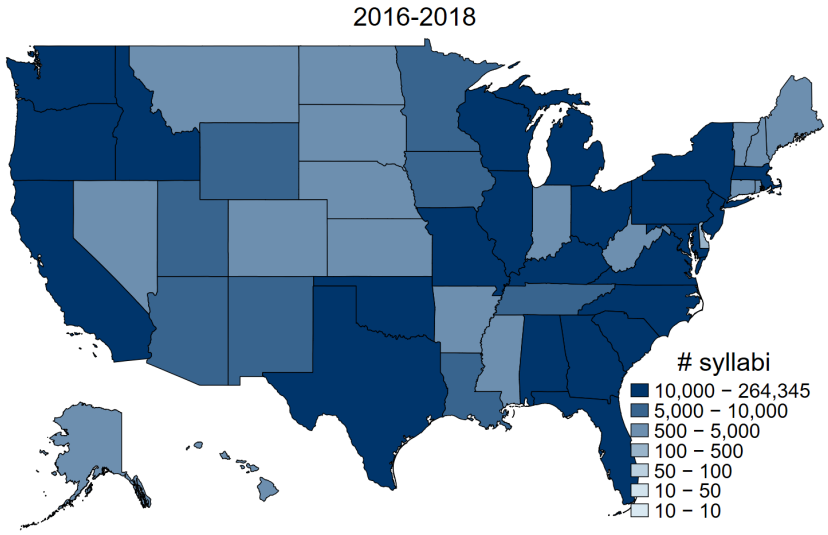
Note: Macro-field composition of the syllabi sample, by five-year periods.

Figure AIII: 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.

Table AI: OSP Fields and Macro-Fields: Mapping

Macro-field	Fields
Business	Business, Accounting, Marketing
Humanities	English Literature, Media / Communications Philosophy, Theology, Japanese Criminal Justice, French, Library Science Classics, Women's Studies, Chinese Journalism, Religion, Sign Language German, Spanish, Hebrew Music, Theatre Arts, Fine Arts Film and Photography, Dance
STEM	Mathematics, Computer Science, Biology Engineering, Chemistry, Physics Architecture, Agriculture, Earth Sciences Basic Computer Skills, Astronomy, Military Science Transportation, Atmospheric Sciences, Medicine Nutrition, Dentistry, Veterinary Medicine Nursing
Social Sciences	Psychology, Political Science, Economics Law, Social Work, Anthropology Geography, Linguistics, Sociology History, 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.