The Education-Innovation Gap*

Barbara Biasi[†] Song Ma[‡]

February 7, 2022

Please click here for the most updated version

Abstract

This paper documents differences across higher-education courses in the coverage of frontier knowledge. Comparing the text of 1.7M syllabi and 20M academic articles, we construct the "education-innovation gap," a syllabus's relative proximity to old and new knowledge. We show that courses differ greatly in the extent to which they cover frontier knowledge. More selective and better funded schools, and those enrolling socio-economically advantaged students, teach more frontier knowledge. Third, instructors play a big role in shaping course content; research-active instructors teach more frontier knowledge. Lastly, the presence of frontier knowledge is strongly related to students' ability to innovate and their labor market outcomes.

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

Keywords: Education, Innovation, Syllabi, Instructors, Text Analysis, Inequality

*We thank Jaime Arellano-Bover, David Deming, David Robinson, Kevin Stange, Sarah Turner, and seminar and conference participants at Yale, Harvard (HBS and HGSE), Stanford (Hoover), UCL, Ohio State, Duke, Erasmus, Maastricht, Queens, Stockholm School of Economics, IIES Stockholm, NBER (Education; Entrepreneurship; Innovation), AEA, CEPR/Bank of Italy, Junior Entrepreneurial Finance and Innovation Workshop, SOLE, IZA TOM and Economics of Education Conferences, and CESifo Economics of Education Conference for helpful comments. Xugan Chen provided outstanding research assistance. We thank the Yale Tobin Center for Economic Policy, Yale Center for Research Computing, Yale University Library, and Yale International Center for Finance for research support. All errors are our own.

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

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

1 Introduction

The dissemination of up-to-date knowledge is key for innovation, which drives economic growth (Goldin and Katz, 2010; Jones, 2009). Higher education (HE) plays a central role in this process. Through the teaching of their curricula, HE programs facilitate human capital accumulation and nurture future innovators (Biasi et al., 2020). Programs, though, might differ in the curricula they offer and in their ability to equip students with up-to-date knowledge.[1] These differences may have implications for labor market outcomes, education choices, and technological progress. To date, these differences have remained unexplored; very little is known on how the content of HE varies across and within schools.

This paper brings together new data and a novel methodology to quantify this variation and measure the extent to which HE covers "frontier," i.e., recently produced, knowledge. Using a novel metric applied to textual information on course syllabi (the content of HE courses) and academic publications (the frontier of knowledge), we document five main facts. First, HE courses differ greatly in the extent to which they cover frontier knowledge. Second, more selective and better funded schools, as well as those enrolling socio-economically advantaged students, teach more frontier knowledge. Third, instructors play a big role in shaping the content of their courses, and research-active instructors teach more frontier knowledge. Lastly, the presence of frontier knowledge is strongly related to students' ability to innovate and their labor market outcomes.

Our empirical analysis uses a novel source of information: the text of course syllabi, documents that describe the content of a HE course. Using this corpus of text, we construct the "education-innovation gap," aimed at capturing the distance between the content of a course and frontier knowledge. We define the gap as a ratio of similarities of a course's content with old knowledge (covered by articles published decades ago) and new, frontier knowledge (covered by the most recent articles). For example, a Computer Science course that teaches *Visual Basic* (a relatively obsolete programming language) in 2020 would have a larger gap compared with a course that teaches *Julia* (a more recent programming language), because *Visual Basic* is mostly covered by old articles and *Julia* is mostly covered by recent articles.[2]

[1]Differences in programs attended have been associated with differences in earnings (Hoxby, 2020; Mountjoy and Hickman, 2020) and rates of invention (Bell et al., 2019).

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

To construct this measure, we start from the raw text of a sample of 1.7 million college and university syllabi, including about 540,000 courses taught at 800 four-year US institutions between 1998 and 2018 and covering nearly all fields. This sample represents about 5% of all courses taught in this time window. While the sample over-represents courses from very selective schools, it is representative of the population in terms of fields, course levels (basic and advanced undergraduate and graduate), and school characteristics.

For each syllabus in the sample, we compare its text with the title, abstract, and keywords of over 20 million academic articles published in top journals since the journal's creation. This process involves three steps. First, we represent each document (a syllabus or an article) as a word vector. We project the text of the document on a dictionary, a comprehensive list of words and expressions that refer to topics and concepts; each element of the vector equals one if the document contains the corresponding dictionary word. Second, we follow Kelly et al. (2018) and weigh each vector element by the word's frequency in the document relative to its frequency in all documents published in previous years. This allows us to account for the fact that words might differ in the frequency of their usage in the English language, their popularity in research articles at a given point in time, and their importance within a given document. Third, we use these weighted word vectors to compute the cosine similarity between each syllabus and each article.

Using these cosine similarities, we construct the education-innovation gap of a given syllabus as the *ratio* of average similarities with (a) all articles published 13-15 years prior to the syllabus's date and (b) all articles published 1-3 years prior. By construction, the gap is higher for syllabi that are more similar to older (rather than newer) knowledge. By virtue of being constructed as a ratio of cosine similarities, the gap is not affected by idiosyncratic attributes of a syllabus such as length, structure, or writing style. Importantly, the use of a weighting scheme that downplays the role of terms used in many documents implies that a syllabus is not necessarily penalized for covering "classics," relatively old articles still commonly used as the fundaments of a discipline or topic.

A few empirical regularities confirm the ability of the education-innovation gap to capture a course's distance from the knowledge frontier. First, the gap is lower for syllabi that reference more recent articles and books as required or recommended readings. Second, the gap varies reasonably across course levels: It is largest for basic undergraduate courses (taught in the first two years of a bachelor's degree and more likely to cover the fundaments of a discipline) and smallest for graduate-level courses (master's and PhD). Third, gradually replacing the "older" knowledge

and a 87 percent increase in a single year (https://juliacomputing.com/blog/2021/08/newsletter-august/, retrieved September 30, 2021).

in a syllabus with "newer" knowledge (i.e., words most frequently used in old and new articles, respectively), as we do in a simulation exercise, progressively reduces the gap.

Significant differences in the gap exist across syllabi. To move a syllabus from the 25th to the 75th percentile of the gap distribution, approximately half of its content would have to be replaced with newer knowledge. Between 4 and 6 percent of this variation can be explained by fields, likely due to differences in teaching styles or in the publication process. Less than one percent of the variation is due to differences across course levels (basic undergraduate, advanced undergraduate, or graduate). Slightly less than 20 percent of the variation occurs over time. To account for these differences, the rest of our analysis compares syllabi within each field, course level, and year. The average syllabus in our data is more similar to newer than to older knowledge: Multiplying the gap by 100 for simplicity, its average equals 95.

To better understand how the content of higher education is shaped, we begin by looking at differences in the education-innovation gap across schools. The gap is smaller in schools with a stronger focus on research (ranked as R1 in the Carnegie classification) and with more resources (higher endowment and spending on instruction and research). The gap is also smaller in more selective schools (for example the "Ivy-Plus," including the eight Ivy League colleges plus Stanford, MIT, Duke, and the University of Chicago) compared to non-selective schools. The magnitude of this difference is such that, in order to make the average syllabus in non-selective schools comparable to the average syllabus in an Ivy-Plus school, 8 percent of its content would have to be replaced with newer knowledge.

Wealthier and more selective schools enroll more socio-economically advantaged students (Chetty et al., 2019). This translates into disparities in access to up-to-date knowledge across students with different backgrounds. In particular, the education-innovation gap is significantly higher in schools enrolling students with lower median parental income. It is also higher in schools with a higher share of Black or Hispanic students.

In principle, part of these differences could be due to a "vertical differentiation" of educational content across schools. If students with greater ability are more able to absorb more up-to-date content, cross-school differences in the gap might simply reflect schools' efforts to provide students with better tailored educational content. We do not find evidence supporting this hypothesis: The negative correlation between the gap and parental income remains when we control for student ability, using the SAT and ACS scores of admitted students.

While differences across schools are significant, most of the variation in the gap occurs within schools across courses taught by different people; instructors explain a quarter of the total variation in the gap. In addition, while the content (and thus the gap) of the typical course remains stable over time, it declines significantly when the instructor of a course changes. Course updating is thus not a gradual process over time, but rather a discontinuous one associated with instructor turnover.

Most higher-education instructors have two jobs –teaching and research– often seen as competing, whose relative importance depends on the instructor's seniority and job title (Courant and Turner, 2020). Our findings, though, point to some complementarities between these two tasks. The education-innovation gap is significantly lower for courses taught by instructors who are more active in producing research (i.e., they publish more, are cited more, and receive more and larger grants), who are more likely to be tenure-track or tenured. This relationship is stronger for graduate courses. The gap is also lower when the instructor's own research is closer to the topics of the course; this could occur because, in these instances, instructors are better updated about the frontier of research and more likely to cover it in their courses. These findings highlight that a proper deployment of faculty across courses can have important impacts on the content of education (for example, ladder faculty are likely best employed in graduate courses). They also suggest that investments on faculty research (both public, in the form of government grants, and institution-specific) can generate additional returns in the form of more updated instruction.

Our results indicate significant differences in the coverage of frontier knowledge across HE courses, across and within schools. Do these differences matter for the production of innovation and for students? To answer this question, the ideal experiment would randomly allocate students to courses with different gaps. In the absence of this random variation, we set on the more modest goal of characterizing the empirical relationship between the education-innovation gap and graduation rates, incomes, and measures of innovation of the students at each school. In an attempt to account for endogenous differences across schools, we control for a large set of school observables such as institutional characteristics, various types of expenditure, instructional characteristics, enrollment by demographic groups and by major, selectivity, and parental background. We find that schools with courses with a lower gap have higher graduation rates and a higher share of undergraduate students who attend graduate school. Students at these schools also produce more patents and have higher earnings after graduation.

So far, we have focused on measuring the novelty of a course with respect to its academic content. The way content is delivered, though, may also differ across courses and be important for students, because it may affect which skills students develop the most. Recent works have highlighted the increasing importance, for students' labor market success, of soft skills, defined as non-cognitive attributes that shape the way people interact with others (Deming, 2017; Deming and Kahn, 2018). We thus measure the novelty of a course with respect to its teaching style and methods by calculating the "soft-skills intensity" of each course, i.e., the extent to which evaluations are based on activities such as group projects, presentations, and surveys, that train soft skills. Courses that are novel with respect to academic content also tend to be novel with respect to teaching styles: The soft-skills intensity of a syllabus is significantly correlated with its education-innovation gap. Richer and more selective schools, as well as those serving fewer socio-economically disadvantaged students, teach more soft-skills intensive courses. Within schools, research-active instructors are most likely to teach soft-skills intensive courses. Lastly, soft-skills intensity is strongly positively associated with student outcomes.

In the final part of the paper, we probe the robustness of our results to the use of additional measures of novelty of a course' academic content. We consider three measures: the share of all "new" knowledge contained in a syllabus (designed not to penalize a syllabus that contains old and new knowledge compared with one that only contains new knowledge); a measure of "tail" knowledge, aimed at capturing the presence of the most recent content; and a version of the education-innovation gap obtained using patents, rather than academic articles, as a measure of frontier knowledge. All these alternative measures are significantly correlated with the education-innovation gap, and our main results are qualitatively unchanged when we use them in lieu of the gap.

Our results document a new and important dimension of heterogeneity in the curriculum of higher-education courses. Our findings can be useful to better understand how educational content is shaped across and within schools and within fields, how access to it varies across students from different backgrounds, and the role of instructors in the design of it. Our results can also be used to rationalize part of the cross- and within-school differences in productivity (including innovation productivity, Hoxby, 2020) outlined in previous research. Lastly, our results are helpful to shed new light on the origin of differences across instructors in their impact on educational content and on students. The latter is important to understand how to best select and deploy faculty across courses (Courant and Turner, 2020).

This paper relates to several strands of the literature. Several studies have characterized heterogeneity in the production of human capital focusing on differences in the returns to educational attainment (Hanushek and Woessmann, 2012), majors and curricula (Altonji et al., 2012), college selectivity (Hoxby, 1998; Dale and Krueger, 2011), and the skill content of college majors (Hemelt et al., 2021; Li et al., 2021). Using new data and a novel empirical approach, we instead document differences in the specific concepts and topics covered in each higher education course.

Second, we contribute to the study of the production of frontier knowledge and innovation. Earlier works (Nelson and Phelps, 1966; Benhabib and Spiegel, 2005) have highlighted an important role for human capital and education in the diffusion of ideas and technological advancements. More recent ones have emphasized the importance of specific fields, such as STEM (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019).[3] Here, we focus on differences in the provision of frontier knowledge, key for innovation and growth, across and within schools.

Third, we document an important role of instructors in shaping the content of higher education, something that has been so far hard to measure (De Vlieger et al., 2020). While some studies have found important effects on student outcomes (Hoffman and Oreopoulos, 2009; Carrell and West, 2010; Braga et al., 2016; Feld et al., 2020), much less is known on why and how instructors impact students. We contribute to filling this gap by studying instructors' contribution to the production of educational content and carefully characterizing differences across instructor types.

Our findings also relate to recent studies on the "democratization" (or lack thereof) of access to valuable knowledge. Bell et al. (2019) have shown that US inventors come from a small set of top US schools, enrolling 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 highly unequal.

Lastly, by using the text of course syllabi as information on content of higher-education instruction, our exercise relates to Kelly et al. (2018), who use the text of patent documents to measure patent quality, and Gentzkow and Shapiro (2010), who characterize the language of newspaper articles to measure media slant. Our approach is most similar to Angrist and Pischke (2017), who use hand-coded syllabi information to study the evolution of undergraduate econometrics classes.

2 Data

Our empirical analysis combines data from multiple sources. These include the text of course syllabi; the abstracts of academic publications; salaries, job titles, publications, and grants of each instructor; information on US higher education institutions; and labor market outcomes for the stu-

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

dents at these institutions. More detail on the construction of our final data set can be found in Appendix B.

2.1 College and University Course Syllabi

We obtained the raw text of a large sample of college and university syllabi from Open Syllabus (OSP), a non-profit organization which collects these data by crawling publicly-accessible university and faculty websites.[4] The initial sample contains more than seven million English-language syllabi of courses taught in over 80 countries, dating back to the 1990s until 2019.

Most syllabi share a standard structure. Basic details of the course (such as title, code, and the name of the instructor) are followed by a description of the content and a list of required and recommended readings for each class session. Syllabi also contain information on evaluation criteria (such as assignments and exams) and general policies regarding grading, absences, lateness, and misconduct. After parsing each syllabus into sections, we extract four pieces of information: (i) basic course details, (ii) the course's content, (iii) the list of required and recommended readings, and (iv) a description of evaluation methods.

Basic course details These include the name of the institution, the title and code of the course, the name of the instructor, as well as the quarter or semester and the academic year in which the course is taught (e.g., Fall 2020). Course titles and codes allow us to classify each syllabus into one of three course levels: basic undergraduate, advanced undergraduate, or graduate. OS assigns each syllabus to one of 69 detailed fields.[5] We use this classification throughout the paper. For some tests, we further aggregate fields into four macro-fields: STEM, Humanities, Social Sciences, and Business.[6]

Course content We identify the portion of a syllabus that contains a description of the course's content by searching for section titles such as "Summary," "Description," and "Content."[7] Typically, this portion describes the basic structure of the course, the key concepts that are covered, and (in many cases) a timeline of the content and the materials for each lecture.

[4]The Open Syllabus Project was founded at the American Assembly of Columbia University but has been independent since 2019. The main purpose of the Project is to support educational research and novel teaching and learning applications.

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

[6]Appendix Table BVII lists all 69 fields and shows the correspondence between fields and macro-fields.

[7]The full list of section titles used to identify each section is shown in Appendix Table BVI.

List of readings We compile a list of bibliographic information for the required and recommended readings of each course by combining the list provided to us by OSP with all other in-text citations that we could find, such as "Biasi and Ma (2021)." We were able to compile a list of references for 71 percent of all syllabi. We then collect bibliographic information on each reference from Elsevier's SCOPUS database (described in more detail in Section 2.2); this includes title, abstract, journal, keywords (where available), and textbook edition (for textbooks).

Methods of evaluation To gather information on the methods used to evaluate students and the set of skills trained by the course, we use information on exams and other assignments. We identify and extract the relevant portion of each syllabus by searching for section titles such as "Exam," "Assignment," "Homework," "Evaluation," and "Group." Using the text of these sections, we distinguish between hard skills (assessed through exams, homework, assignments, and problem sets) and soft skills (assessed through presentations, group projects, and teamwork). We were able to identify this information for 99.9 percent of all syllabi.

Sample restrictions and description To maximize consistency over time, we focus our attention on syllabi taught between 1998 and 2018 in four-year US institutions with at least one hundred syllabi in our sample.[8] We excluded 35,917 syllabi (1.9 percent) with less than 20 words or more than 10,000 words (the top and bottom 1 percent of the length distribution).

Our final sample, described in panel (a) of Table 1, contains about 1.7 million syllabi of 542,251 courses at 767 institutions. Thirty-one percent of all syllabi cover STEM courses, 11 percent cover Business, 30 percent cover Humanities, and 26 percent cover Social Science. Basic courses represent 39 percent of all syllabi and graduate courses represent 33 percent. A syllabus contains an average of 2,226 words in total, with a median of 1,068. Our textual analysis focuses on "knowledge" words, i.e., words that belong to a dictionary (see Section 3 for details). The average syllabus contains 420 unique knowledge words.

2.2 Academic Publications

To construct the research frontier in each field and year, we use information from Elsevier's SCOPUS database and compile the list of all peer-reviewed articles that appeared in the top academic journals of each field since the journal's foundation.[9] We define top journals as those ranked among the

[8]For consistency and comparability, we removed 129,429 syllabi from one online-only university, the University of Maryland Global Campus.

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

top 10 by Impact Factor (IF) in each field at least once since 1975 (or the journal's creation if it happened after 1975).[10] Our final list of publications includes 20 million articles in the same fields as our syllabi, corresponding to approximately 100,000 articles per year.[11] We capture the knowledge content of each article with its abstract.

Alternative measure of knowledge: Patents An alternative way to measure the knowledge frontier is to use the text of patents, rather than academic publications. To this purpose, we collected data on the text of more than six million patents issued since 1976 from the US Patents and Trading Office (USPTO) website.[12] We capture the content of each patent with its abstract.

2.3 Instructors: Research Productivity, Funding, Job Title, and Salary,

Nearly all course syllabi report the name of the course instructor. Using this information, we collected data on instructors' research productivity (publications and citations) and the receipt of public research funding. For a subset of instructors, we also collected information on job titles and annual salary.

Research Productivity Publications and citations data come from Microsoft Academic (MA), a search engine that lists publications, working papers, other manuscripts, and patents for each listed researcher, together with the counts of citations to these documents. We linked MA records to syllabi instructors via fuzzy matching based on name and institution (details on the procedure are in Appendix B). We are able to successfully find 33 percent of all instructors, and we assume that the instructors we could not find never published any article (Table 1, panel (b)).

Using data from MA, we measure each instructor's research quantity and quality with the number of publications and the number of citations received in the previous five years.[13] On average, instructors published 5.5 articles in the previous five years, with a total of 125 citations (Table 1, panel (b)). The distributions of citation and publication counts are highly skewed: The median

[10] Even if a journal appeared only once in the top 10, we collect all articles published since its foundation.

[11] SCOPUS classifies articles into 191 fields. To map each of these to the 69 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. Details on the mapping of fields between the syllabi and SCOPUS articles are contained in Appendix B.

[12] Our web crawler collected the text content of all patents (in HTML format) from `http://patft.uspto.gov/netahtml/PTO/srchnum.htm`, with patent numbers ranging from 3850000 to 10279999).

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

instructor in our sample only published one article in the previous five years and received no citations.

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

Job Title and Salary In many US states, salaries of public college and university employees are made public to comply with state regulations on transparency and accountability. These records are usually disclosed online, together with each employee's name and job title. We were able to collect information on salaries and job titles for 35,178 instructors in our syllabi sample (10.6 percent of all instructors and 14.3 percent of public-sector instructors), employed in 490 public institutions in 16 states. We observe an average of two years' worth of salary for each employee (the modal year is 2017). We detail the coverage of the salary data in the Online Appendix.

Among all syllabi instructors for which we have job title information, 42 percent are ladder faculty (including 11 percent of assistant professors, 13 percent of associate professors, and 18 percent of full professors; Appendix Figure AI, panel (a)). Instructors earn $80,388 on average, although large variation in pay exists between job titles. Conditional on field, course level, and year, adjunct professors and lecturers earn $63,396; clinical and practice professors earn $119,685; assistant, associate, and full professors earn $85,261, $97,766, and $128,589 respectively (Appendix Figure AI, panel (b)).[15]

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

[15] Panel (b) of Appendix Figure AI displays point estimates and confidence intervals of indicators for job titles in an OLS regression of salaries (expressed in $1,000), controlling for field-by-course level-by-year fixed effects. We cluster standard errors at the instructor level).

2.4 Information on US Higher Education Institutions

The last component of our dataset includes information on all US colleges and universities of the syllabi in our data. Our primary source is the the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES).[16] For each school, IPEDS reports a set of institutional characteristics (such as name and address, control, affiliation, and Carnegie classification); the types of degrees and programs offered; expenditure and endowment; characteristics of the student population, such as the distribution of SAT and ACT scores of all admitted students, enrollment figures for different demographic groups, completion rates, and graduation rates; and faculty composition (ladder and non-ladder). We link each syllabus to the corresponding IPEDS record via a fuzzy matching algorithm based on school names. We are able to successfully link all syllabi in our sample.

We complement data from IPEDS with information on schools and students from three additional sources. The first one is the school-level dataset assembled and used by Chetty et al. (2019), which includes a school's selectivity tier (defined using Barron's scale), the incomes of students and parents, the number of patents obtained by all students, and measures of intergenerational mobility (we use the share of students with parental income in the bottom quintile who have incomes in the top quintile as adults). These data are calculated using data on US tax records for a cross-section of cohorts who graduated between 2002 and 2004. The second is the Survey of Earned Doctorates, conducted by the NSF, which reports characteristics of all PhD receivers in US institutions each year. We use information on students' graduating cohort and bachelor's institution to construct the share of undergraduate students in each school and year who eventually complete a PhD, for the years 1998-2018.[17] The third is the College Scorecard Database of the US Department of Education, an online tool designed to help users compare costs and returns of attending various colleges and universities in the US. This database reports the incomes of graduates ten years after the start of the program. We use these variables, available for the academic years 1997-98 to 2007-08, to measure student outcomes for each school.

Panel (c) of Table 1 summarizes the sample of colleges and universities for which we have syllabi data. The median parental income at these schools is $97,917 on average. Across all schools, 3 percent of all students have parents with incomes in the top percentile. The share of minority

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

[17] The Survey of Earned Doctorates has been conducted since 1957. To assign a PhD recipient to their undergraduate institution, we use information on the institution where they obtained their bachelor's degree; to assign the recipient to a bachelor's degree cohort, we subtract 6 to their year of PhD receipt.

students is equal to 0.22, with a standard deviation of 0.17. Graduation rates average 61.4 percent in 2018, whereas students' incomes ten years after school entry, for the 2003–04 and 2004–05 cohorts, are equal to $45,035. Students' intergenerational mobility is equal to 0.29 on average.

2.5 Data Coverage and Sample Selection

Our syllabi sample only covers a small fraction of all courses taught in US schools between 1998 and 2018. The number of syllabi increases over time, from 17,479 in 2000 to 68,792 in 2010 and 190,874 in 2018 (Appendix Figure AII).

To more accurately interpret our empirical results, it is crucial to establish whether and how our sample is selected. To do so, we compiled the full list of courses offered between 2010 and 2019 in a subsample of 161 US institutions (representative of all institutions included in IPEDS) using the course catalogs in the archives of each school.[18] This allows us to compare our sample to the population of all courses for these schools and years.

This exercise does not reveal stark patterns of selection based on observables. The share of catalog courses covered by the syllabi sample remained stable over time, at 5 percent (Appendix Figure AV). This suggest that, at least among the schools with catalog information, the increase in the number of syllabi over time is driven by an increase in the number of courses that are offered, rather than an increase in sample coverage. Our syllabi sample is also similar to the population in terms of field and course level composition. In 2018, STEM courses represent 32 percent of syllabi in our sample and 24 percent of courses in the catalog; Humanities represent 25 and 31 percent, and Social Sciences represent 21 and 19 percent, respectively (Appendix Figure AIII). Similarly, basic undergraduate courses represent 40 percent of syllabi in our sample and 31 percent of courses in the catalog; advanced undergraduate courses represent 26 and 30 percent, and graduate courses represent 33 and 38 percent (Appendix Figure AIV). These shares are fairly stable over time.

In addition, a school's portion of the catalog that is included in our sample and the change in this portion over time are unrelated to school observables. We show this in panel (a) of Table 2 (column 1), where we regress a school's share of courses included in our sample in 2018 on the following variables, one at the time and also measured in 2018: financial attributes (such as expenditure on instruction, endowment per capita, sticker price, and average salary of all faculty), enrollment, the share of students in different demographic categories (Black, Hispanic, alien), and the share of

[18] We begin by randomly selecting 200 schools among all 4-year IPEDS institutions. Among these, we were able to compile course catalogs for 161 institutions, listed in Appendix Table AII. These look very similar in terms of observables to all schools in our sample (Appendix Table AIII). We focus our attention on years from 2010 to maximize our coverage. For an example of a course catalogue, see https://registrar.yale.edu/course-catalogs.

students graduating in Arts and Humanities, STEM, and the Social Sciences. We also estimate the joint significance of all these variables together. We find that these variables are individually and jointly uncorrelated with the share of courses in the syllabi sample, with an F-statistic smaller than one. In column 2 we repeat the same exercise, using the 2015-2018 change in the share of courses included in the syllabi as the dependent variable. Our conclusions are unchanged.

The only dimension in which our syllabi sample appears selected is school selectivity. Relative to non-selective institutions (for whom the share of courses in the sample is less than 0.1 percent), Ivy-Plus and Elite schools have a 0.9 percentage point higher share of courses included in the syllabi sample, and selective public schools have a staggering 4.5 percent higher share. Taken together, these tests indicate that our syllabi sample does not appear to be selected on the basis of observable characteristics of schools and fields, although it does over-represent Ivy-Plus and Elite and selective public schools. By construction, though, we cannot test for selection based on unobservables. Our results should therefore be interpreted with this caveat in mind.

3 Measuring the Education-Innovation Gap

To construct the education-innovation gap we combine information on the content of each course, captured by its syllabus, with information on frontier knowledge, captured by academic publications. We now describe the various steps for the construction of this measure, provide the intuition behind it, and perform validation checks.

Step 1: Measuring Similarities in Text

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

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

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

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

The *TFIDF* weight of a term w in document d is:

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

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

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

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

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

Accounting for changes in term relevance over time The weighting approach described so far calculates the frequency of each term by pooling together documents published in different years. This is not ideal for our analysis, because the resulting measures of similarity between syllabi and publications would ignore the temporal ordering of these documents. Instead, we are interested in the novelty of the content of a syllabus d relative to research published in the years prior to d, without taking into account the content of future research. To see this consider, for example, course CS229 at Stanford University, taught by Andrew Ng in the early 2000 and one of the first entirely focused on *Machine Learning*. Pooling together documents from different years would result in a very low $TFIDF_{wd}$ for the term "machine learning" in the course's syllabus: Since the term has been used very widely in the last years, its frequency across all documents would be very high and its IDF very low. Not accounting for changes in the frequency of this term over time would then lead us to misleadingly underestimate the course's path-breaking content.

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

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

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

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

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

$$\rho_{dk} = \frac{V_d}{\|V_d\|} \cdot \frac{V_k}{\|V_k\|} \tag{1}$$

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

3.1 Calculating the Education-Innovation Gap

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

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

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

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

where ρ_{dk} is the cosine similarity between syllabus d and a article k (defined in equation (1)), $\Omega_\tau(d)$ is the set of all articles published in the three-year time interval $[t(d) - \tau - 2, t(d) - \tau]$, and $|\Omega_\tau(d)|$ is the count of all such articles.[19]

Step 3: We construct the education-innovation gap as the ratio between the average similarity of a syllabus with older technologies (published in τ) and the similarity with more recent ones ($\tau' < \tau$):

$$Gap_d \equiv \left(\frac{S_d^\tau}{S_d^{\tau'}} \right) \tag{2}$$

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

Our measure features two attractive properties. First, being constructed as a ratio, the gap is not affected by syllabus-specific attributes such as style or language, which instead could impact a simple measure of similarity (for example S_d^1).[20] Second, the use of $TFBIDF$ weights in the construction of word vectors reduces the impact on the gap of terms frequently used across documents; topics which are present in many syllabi (regardless of how new they are) will receive a smaller weight. This implies that our measure does not heavily penalize syllabi for covering "classic" topics in the literature, as long as these are widespread across courses.

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

[20] We demonstrate this in a simulation exercise where we manually create a sample of 1.7 million syllabi as sets of dictionary words, for which we know ex ante the ratio between "old" knowledge terms (more popular among old publications) and "new" knowledge terms (most popular among recent publications). The education-innovation gap performs much better as a measure of this ratio than a simple measure of similarity with new terms (Appendix C).

3.2 Validating The Measure and Interpreting Its Magnitude

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

Second, we show that the gap varies reasonably across course levels: More advanced and graduate courses have a lower gap compared with basic undergraduate courses. Controlling for field-by-year effects, basic undergraduate courses have a gap of 95.7; advanced undergraduate courses have a gap of 95.3, and graduate courses have a gap of 94.7 (Figure 1, panel (b)). This suggests that more advanced courses cover content closer to frontier research.

Third, we simulate how changing the content of a course affects the education-innovation gap. Specifically, we progressively replace "old" words with "new" words in a randomly selected subsample of 100,000 syllabi and re-calculate the gap for each syllabus as we replace more words. New words as those in the top 5 percent in terms of frequency in the new publication corpus between t-3 and t-1 or in the new publication corpus between t-3 and t-1 but not in the old publication corpus between t-15 and t-13; old words are those in the top 5 percent in terms of frequency in the old publication corpus between t-15 and t-13 or in the old publication corpus between t-15 and t-12 but not in the new publication corpus between t-3 and t-1. This exercise shows that the gap monotonically decreases as we replace more old words with new ones (Figure 1, panel (c)). This simulation is also useful to gauge the economic magnitude of changes in the gap. In particular, a unit change in the gap requires replacing 10 percent of a syllabus's old words (or 34 old words, compared with 331 words for the median syllabus).

3.3 The Education-Innovation Gap: Variation and Variance Decomposition

The average course has a gap of 95.3, with a standard deviation of 5.8, a 25th percentile of 91.6, and a 75th percentile of 98.8 (Appendix Figure AVI). To give an economic meaning to this variation, we use of the relationship illustrated in panel (c) of Figure 1. In order to move a syllabus from the 75th to the 25th percentile of the distribution (a 7.2 change in the gap) we would have to replace approximately 200 of its words, or 60 percent of the content of the median syllabus.

To better understand what drives the variation in the gap, we perform a Shapley-Owen decomposition of its variance into five sets of factors: year, field, school, course, and instructor. For each factor j, we calculate the partial R^2 as

$$R_j^2 = \sum_{k \neq j} \frac{R^2 - R^2(-j)}{K!/j!(K-j-1)!}$$

where $R^2(-j)$ is the adjusted R^2 of a regression that excludes fixed effects for all factors except j; this quantity captures the share of the variation captured by factor j.[21]

This exercise indicates that differences across fields explain 4 percent of the total variation in the gap, while differences across schools explain 2 percent. Courses explain a large 33 percent, indicating a great deal of persistence in the content of a course over time. Importantly, differences across instructors explain a large 25 percent. In column 2, we obtain similar exercises when substituting courses with course levels; the latter explain less than 1 percent of the total variation.

4 The Education-Innovation Gap Across Schools

The decomposition exercise indicates that differences across schools explain a small share of the total variation in the gap. Yet, cross-school differences can be useful to understand how the content of higher education is shaped and how access to it relates to students' socio-economic background. We explore these differences in this section.

4.1 School Characteristics

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

We estimate pairwise correlations between the gap and these attributes controlling for field, course level, and year of the syllabus, with the following specification:

$$\text{Gap}_i = \beta X_i + \phi_{f(i)l(i)t(i)} + \varepsilon_i \tag{3}$$

[21] We use adjusted R^2 throughout to account for the large number of fixed effects in the model.

where Gap_i measures the education-innovation gap of syllabus i, taught in school $s(i)$ in year $t(i)$; the variable X_i is the institutional characteristic of interest; and field-by-level-by-year fixed effects ϕ_{flt} control for systematic, time-variant differences in the gap that are common to all syllabi in the same field and course level. We cluster standard errors at the institution level.

Institutional and financial attributes Estimates of β for each school characteristic are shown in Figure 2. Public schools have a slightly larger gap compared with non-public schools, but this difference is indistinguishable from zero. No differences emerge between LACs and other schools. R1 schools have a 0.2 smaller gap compared with schools with a lower research intensity.

In order to quantify the economic magnitude of these differences, we can use the simulation results in Figure 1 (panel c). In order to close the difference in the gap between R1 and other schools, we would have to replace approximately 2 percent of the knowledge content of the median syllabus (7 terms). The difference between R1 and other schools is therefore quite small.

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

Selectivity Next, we test whether the gap differs across schools with different selectivity. Following Chetty et al. (2019), we bin schools in four "tiers" according to their selectivity in admissions, as measured by Barron's 2009 ranking. "Ivy Plus" include Ivy League universities and the University of Chicago, Stanford, MIT, and Duke. "Elite" schools are all the other schools classified as tier 1 in Barron's ranking. "Highly selective" schools include those in tiers 2 and 3, while "Selective" schools are those in tiers 4 and 5. Lastly, "Non-selective" schools include those in Barron's tier 9 and all four-year institutions not included in Barron's classification.

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

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

where the vector $\mathbf{S}_i'$ contains indicators for selectivity tiers (we omit non-selective schools), and everything is as before.

Point estimates of the coefficients vector $\boldsymbol{\beta}$ in equation (4), shown as diamonds in Figure 2, indicate that more selective schools offer content closer to the research frontier. Ivy Plus and Elite schools have the smallest gap, -0.84 smaller than non-selective schools (corresponding to an 8 percent difference in the median syllabus). Highly selective schools have a -0.67 smaller gap and selective schools have a -0.51 percent smaller gap (5 percent). A possible interpretation for these differences is that more selective schools offer higher-quality education. However, if higher-ability students are better able to absorb newer content, another possibility is that schools tailor instruction to the abilities of their students. We attempt to test this hypothesis in the next section and in Section 6, where we relate the education-innovation gap to student outcomes.

4.2 Students' Characteristics

Schools with different characteristics serve different populations of students; for example, Ivy-Plus and Elite schools are disproportionately more likely to enroll students from wealthier backgrounds (Chetty et al., 2019). Cross-school differences might therefore translate into significant disparities in access to up-to-date knowledge among students with different backgrounds. Here, we focus on two dimensions of socio-economic backgrounds: parental income and race and ethnicity.

Parental income To establish a relationship between the education-innovation gap and parental income of students enrolled at each school, we re-estimate equation (3) using two measures of income as the explanatory variable: Median parental income and the share of parents with incomes in the top percentile of the national distribution, constructed using tax returns for the years 1996 to 2004 (Chetty et al., 2019). These estimates, shown as the full triangles in the bottom panel of Figure 2, indicate that schools serving more economically disadvantaged students offer courses with a lower gap. Specifically, a one-percent increase in parental median income is associated with a 0.56 lower gap, which corresponds to a 5 percent difference in the median syllabus. Similarly, an increase in the share of students with parental income in the top percentile from 0.01 to 0.10 is associated with a 0.42 lower gap (4 percent).

In principle, part of these differences could be due to a "vertical differentiation" of educational content across schools. If students with greater ability are better able to absorb more up-to-date content, cross-school differences in the gap might reflect schools' efforts to provide students with appropriate educational content. Our data, however, do not support this hypothesis. Controlling for the average SAT score of students admitted at each school as a proxy for their ability yields estimates only slightly smaller than the baseline (hollow triangles in the bottom panel of Figure

2). This rules out vertical differentiation as an explanation for differences in the gap across schools serving students with different backgrounds.

Students' race and ethnicity Schools that enroll a higher share of minority students (defined as those who are either Black or Hispanic) also tend to have a higher gap. Using the share of minority students as the explanatory variable in equation (3) reveals that a one-percentage point increase in the share of minority students at each school is associated with a 0.58 higher gap, equivalent to a 6 percent change in the average syllabus. As before, this relationship holds (although it becomes less precise) if we control for average student ability.

In line with existing evidence on disparities in access to selective schools among more and less advantaged students, our results document a new dimension of inequality: That in access to educational content that is close to the research frontier. Importantly, this inequality cannot be explained by differences in ability.

5 The Role of Instructors

Most of the variation in the gap occurs within schools, especially across courses taught by different people. We now investigate in depth the role of instructors and their characteristics in shaping the content of higher education.

5.1 Persistency In A Course's Content Over Time and Changes in Instructors

Our decomposition exercise indicates that 33 percent of the total variation in the gap occurs across courses. This suggests that the content of a given course tends to stay relatively stable over time. To better understand the process of updating of a course's content, we study how content evolves when the course instructor changes. We estimate an event study of the gap in a 8-years window around the time of an instructor change:

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

where i, c, f, and t denote a syllabus, course, field, and year respectively, and the variable T_c represents the first year in our sample in which the instructor of course c changed.[22] To more precisely

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

capture the impact of an instructor change, we restrict our attention to courses taught by a maximum of two instructors in each year and we set the indicator function equal to zero for all courses without an instructor change, which serve as the comparison group. We cluster our standard errors at the course level. In this equation, the parameters δ_k capture the differences between the gap k years after an instructor change relative to the year preceding the change.

OLS estimates of δ_k, shown in Figure 3, indicate that a change in a course's instructor is associated with a sudden decline in the education-innovation gap. Estimates are indistinguishable from zero and on a flat trend in the years leading to an instructor change; the year of the change, the gap declines by 0.1. This decline is equivalent to replacing 2 percent of the content of a syllabus, or 8 knowledge words.

In Table 4 (panel a) we re-estimate equation (4) for different subsamples of syllabi, pooling together years preceding and following an instructor change. After a change, the gap declines for all fields and course levels by about 0.1 on average (8 additional words or 2 percent of a course's content, column 1, significant at 1 percent). The decline is largest for Humanities and STEM courses (-0.12 and -0.10, columns 3 and 4, respectively), as well as for graduate courses (-0.11, column 8).

These results indicate that course updating is not a gradual process over time. Instructors who teach the same course for many years tend to leave the content unchanged, while those who take over a course from someone else significantly update content, bringing it closer to the knowledge frontier. Our findings also confirm that instructors play a crucial role in shaping the content of the courses they teach, particularly for advanced courses.

5.2 The Education-Innovation Gap and Instructors' Characteristics

The decline in the gap that follows an instructor change, though, could mask substantial differences across instructors. For example, the decline could differ for instructors who are more research-active, who spend less time teaching but are better informed on the frontier of knowledge. Similarly, the decline could depend on whether the new instructor is an expert on the topics covered by the course. We now explore these possibilities in depth.

Ladder vs non-ladder faculty Ladder (i.e., tenure-track or tenured) faculty are generally more focused on research compared with non-ladder faculty, whose primary job is to teach. Ex ante, whether one type of faculty or the other would be better at teaching up-to-date content is ambiguous. On the one hand, research and teaching are often seen as two competing tasks; if non-ladder faculty are more specialized on teaching, they might be better at keeping educational content updated. On the other hand, ladder faculty might be better informed on frontier knowledge, and thus be more likely to include it in the courses they teach.

Comparing the education-innovation gap across job titles, controlling by field-by-course level-by-year effects, indicates that non-ladder faculty (adjunct professors) have the largest gap, at 95.8 (Figure 4). Tenure-track assistant professors, on the other hand, have the lowest gap at 95. The difference between assistant and adjunct professors is equivalent to 30 words, or 7 percent of a syllabus's content.

Notably, the gap is almost as high for full (tenured) professors as it is for adjuncts, at 95.6. Associate professors have a slightly smaller gap than full at 95.5, but still significantly higher than assistant professors. Junior faculty on the tenure track thus appear to teach the courses with the most updated content.

Research productivity and fit One possible explanation for these results is that assistant professors are more active in research, and thus more informed on the knowledge frontier. We test this hypothesis directly by exploring the relationship between a course's gap and the research productivity of the instructor, measured using individual counts of citations and publications in the previous five years. We estimate the following equation:

$$\text{Gap}_i = \sum_{n=1}^{4} \beta^n q^n_{k(it(i))} + \gamma_{c(i)} + \psi_{f(i)t(i)} + \varepsilon_i \tag{5}$$

where q^n_k equals one if instructor k's measure of research productivity (publications or citations) is in the nth quartile of the distribution (the omitted category are courses whose instructors do not have any publications or citations). Course fixed effects $\gamma_{c(i)}$ and field-by-year fixed effects $\psi_{f(i)t(i)}$ control for unobserved determinants of the gap that are specific to a course in a given field and year. Estimates of β^n, which capture the difference in the gap between courses taught by faculty with productivity in the nth quartile and those taught by faculty with no citations or publications, are thus obtained off of changes in instructors for the same course over time.

Estimates of β^n, shown in Table 5, indicate that the gap progressively declines as the research productivity of the instructor grows. In particular, a switch from an instructor without publications and one with a number of publications in the top quartile of the field distribution is associated with a 0.11 decline in gap (equivalent to changing 8 words or 2 percent of a course's syllabus; Table 5, panel (a), column 1, significant at 1 percent). Similarly, a switch from an instructor without citations to one with citations in the top quartile is associated with a 0.06 lower gap (panel b, column 1, significant at 5 percent). These relationships are stronger for Social Science courses (column 5) and for courses at the graduate level (column 8).[23]

A possible explanation for this finding is that research-active instructors are better informed about the research frontier. If this is the case, we should expect the relationship between productivity and the gap be stronger for courses that are closer in terms of topics to the instructor's own research. To test for this possibility, we construct a measure of "fit" between the course and the instructor's research, defined as the cosine similarity between the set of all syllabi from the same course across *all* schools (to avoid issues of self-reflection) and the instructor's research in the previous 5 years.[24] We then correlate this measure with the education-innovation gap, controlling for course and field-by-year fixed effects like we do in equation 5. Estimates of this relationship indicate that a one-standard deviation increase in instructor-course fit is associated with a 0.09 decline in the gap (Table 6, significant at 5 percent). This relationship is particularly strong for STEM and Social Science (column 4) and for courses at the advanced undergraduate level (column 6).

Research funding Our results so far indicate a positive relationship between research output and the education-innovation gap. In Table 7, we use data on the number of NSF and NIH grants received by each instructor and test whether the same relationship holds for research inputs, such as government grants; as before, we control for course and field-by-year effects. A switch from an instructor who never received a grant to one with at least one grant is associated with a 0.05 reduction in the gap (column 1, significant at 5 percent). This suggest that public investments in academic research can yield additional private and social returns in the form of more up-to-date instruction.[25]

Taken together, these findings reveal a crucial role for instructors in shaping the content of the course they teach, with some complementarities between research and teaching: Research-active instructors are more likely to cover frontier knowledge in their courses, especially when teaching advanced courses and courses closest in topic to their own research agenda. Our findings suggest

[23] Panels (a) and (b) of Appendix Figure AVII show a binned scatterplot of the gap and either citations (panel a) or publications (panel c) in the prior 5 years, controlling for field effects. n this figuIre, the horizontal axis corresponds to quantiles of each productivity measures; the vertical axis shows the average gap in each quantiles.

[24] One attractive property of this measure is that it is does not uniquely reflect the content of the syllabus itself, which is of course directly shaped by the instructor; rather, it aims at capturing the content of all courses on the same topic. Constructing this measure requires obtaining a unique identifier for courses on the same field or topic (e.g. Machine Learning) across schools. We describe the procedure we use to do this in Appendix B.

[25] A binned scatterplot reveals a negative relationship between the gap and the number of NSF and NIH grants (Appendix Figure AVII, panel d).

that a proper deployment of faculty across courses can have important impacts on the content of education, and that investments on faculty research (both public, in the form of government grants, and institution-specific) can generate additional returns in the form of more updated instruction.

6 The Education-Innovation Gap and Students' Outcomes

Significant differences in access to up-to-date knowledge exist both across and within schools, and across courses taught by different people. A natural question is whether these differences matter for students' outcomes and for the production of innovation. To this end, we now explore the relationship between the gap and a) "innovation measures," such as a school's share of undergraduate students who attend graduate school and the share of students with at least one patent; and b) "labor-market measures," such as graduation rates, income, and intergenerational mobility.

All these outcomes are measured at the school level or at the school-by-cohort level (with the exception of the share of students who attend graduate school, as we explain later).The education-innovation gap is instead measured at the syllabus level. To construct a school-level measure we follow the school value-added literature (see Deming, 2014, for example) and estimate the school component of the gap using the following model:

$$\text{Gap}_i = \theta_{s(i)} + \phi_{f(i)l(i)t(i)} + \varepsilon_i. \tag{6}$$

In this equation, the quantity θ_s captures the average education-innovation gap of school s, accounting flexible time trends that are specific to the level l and the field f of the course. Because outcome measures refer to students who complete undergraduate programs at each school, we construct θ_s using only undergraduate syllabi; our results are robust to the use of all syllabi. Appendix Figure AX shows the distribution of θ_s; the standard deviation is 0.85, corresponding to a 5 percent change in the average syllabus.

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

$$Y_{st} = \delta\hat{\theta}_s^z + X_{st}\gamma + \tau_t + \varepsilon_{st} \tag{7}$$

where Y_{st} is the outcome for students who graduated from school s in year t, $\hat{\theta}_s^z$ is the estimated school fixed effect in equation (6) standardized to have mean zero and variance one, X_{st} is a vector of school observables, and τ_t are year fixed effects. We calculate bootstrapped standard errors, clustered at the level of the school, to account for the fact that $\hat{\theta}_s^z$ is an estimated quantity.

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

6.1 Innovation Measures

Graduate school attendance We begin by studying the relationship between the gap and the share of students who obtain a PhD. We construct this variable using data from the NSF Survey of Earned Doctorates, separately for five macro-fields: STEM, Health, Business, Social Science, and Humanities. To map the level of aggregation of this variable, we aggregate the education-innovation gap at the school-by-macro field level (rather than just at the school level) and modify equation (7) slightly so that one observation in our data is a school-by-macro field in a year. In column 1 of Table 8 (panel a) we pool data across macro-fields. The unconditional correlation between the gap and the share of students who attend graduate school is negative and statistically significant: A one-standard deviation lower gap is associated with a 0.4 percentage point higher share, or 17 percent compared with an average of 0.0265 percent. The correlation is particularly strong for Social Science (-0.0124) and Health (-0.0074), while it is small and indistinguishable from zero for STEM, Business, and Humanities. These correlations remain remarkably robust when we control for school characteristics (Table 8, panel b).

Students' patents Next, we test whether students at schools which offer courses with a lower gap produce more innovation, in the form of patents. We do so by substituting the total number of patents received by students at each school in equation (7). Unconditionally, a one-standard deviation decline in the gap is associated with 27 additional patents at a given school, or 20 percent compared with an average of 131 patents (Table 8, panel b, column 8, p-value equal to 0.11). The relationship remains robust and even becomes more precise controlling for school observables (Table 8, panel b).

6.2 Labor Market Outcomes

Graduation rates Next, we examine the relationship between the education-innovation gap and labor market outcomes. We begin with an outcome which immediately precedes entry in the labor market: graduation. Column 1 of Table 9 shows the relationship between the gap (measured in standard deviations) and graduation rates. An estimate of -0.05 in panel a, significant at 1 percent, indicates that a one-standard deviation decline in the gap (or a 10 percent change in the content of a syllabus) is associated with a 5 percentage point higher graduation rates. Compared with an average of 0.61, this corresponds to a 8 percent increase in graduation rates.

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

Students' income and intergenerational mobility Graduation rates are a strictly academic measure of student success; however, they are also likely to affect students' long-run economic trajectories. To directly examine the relationship between the education-innovation gap and students' economic success after they leave college, in columns 2-8 of Table 9 we study the relationship between the gap and various income statistics.

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

Information on mean student incomes at the school level is also reported by Chetty et al. (2019), calculated using tax records for a cross section of students. Unconditional estimates (which omit year effects due to the cross-sectional structure of the data) indicate that a one-standard deviation in the gap is associated with a 7 percent increase in students' mean income (panel a, column 5, significant at 1 percent). This estimate is smaller, at 1.4 percent, when controlling for institutional characteristics (panel b, column 5, significant at 1 percent).

Lastly, in columns 6 through 8 of Table 9 we investigate the relationship between the gap and the probability that students' incomes reach the top echelons of the income distribution. Estimates with the full set of controls indicate that a one-standard deviation decline in the gap is associated with a 0.84 percentage-point increase in the probability of reaching the top 20 percent (2.2 percent, panel b, column 6, significant at 1 percent), a 0.53 percentage-point increase in the probability of reaching the top 10 percent (2.5 percent, column 7, significant at 5 percent), and a 0.31 percentage-point increase in the probability of reaching the top 5 percent (2.7 percent, column 8, significant at 10 percent). Taken together, these results indicate a positive relationship between the school-level education-innovation gap and students' average and top incomes.

Lastly, in column 9 of Table 9 we study the association between the gap and intergenerational mobility. The unconditional correlation between these two variables is equal to -0.0293, indicating that a one-standard deviation lower gap is associated with a 2.9 percentage-points increase in intergenerational mobility (9.9 percent, panel a, column 9, significant at 1 percent). This correlation, however, becomes smaller and indistinguishable from zero when we control for school observables, reaching -0.0047 when we include the full set of controls (column 9, panel b, p-value equal to 0.15).

Summary Our findings indicate that a lower education-innovation gap at the school level is associated with more innovation and improved academic and economic outcomes of the students at each school. The lack of experimental variation in the gap across schools prevents us from pinning down a causal relationship with certainty; yet, our results are robust to the inclusion of controls for a large set of school and student characteristics, indicating that these correlations are unlikely to be driven by cross-school differences in spending, selectivity, major composition, or parental background. Thee findings point to a potentially important role for up-to-date instruction on innovation and the outcomes of students as they exit college and enter the labor market.

7 Novelty in Teaching Styles: Soft Skills Intensity

By definition, the education-innovation gap focuses on the novelty of a syllabus with respect to its academic *content*, and it largely abstracts from the way this content is taught. It is possible, however, that courses with a similar gap might feature very different teaching styles; some might be taught in a way that emphasizes abstract content and assesses students with midterms and exams, whereas others might place more focus on teamwork. To examine heterogeneity across syllabi in teaching styles, we focus here on an alternative dimension of "novelty:" soft skills, defined as non-cognitive abilities that define how a person interacts with their colleagues and peers, and identified by recent literature as increasingly demanded in the labor market (Deming, 2017).

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

The measure of soft-skills intensity is negatively correlated with the education-innovation gap (with a correlation of -0.14, Figure 5, panel a). Cross-school differences in the skill intensity of the courses display the same patterns we found for the education-innovation gap: The prevalence of soft skills is higher in schools with higher expenditure on instruction and salaries, increases with school selectivity, and it is larger for schools where the median parental income is in the top portion of the distribution and those enrolling a higher share of minority students (Figure AVIII, panel a). Soft skills are also more prevalent among courses taught by the most research-productive instructors (Figure AIX, panel a).

In closing, we examine the relationship between courses' soft-skills intensity and student outcomes. Controlling for the full set of school observables used in Tables 8 and 9, a one-standard deviation increase in the soft-skills intensity of a school's courses is associated to a 1.2 percentage-point increase in graduation rates (2 percent, Table AIV, panel h, column 1, significant at 1 percent); a 1.7 percent higher mean income (column 2, significant at 1 percent); and a 1.2 percent higher chances of reaching the top income quintile for students with parental income in the bottom quintile (18 percent, column 9, significant at 1 percent).

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

8 Alternative Measures of Course Novelty

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

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

Presence of Old Content The education-innovation gap measures the presence, in a syllabus, of new content relative to older one. Consider two syllabi which both cover the same frontier research in a given field; the first syllabus is shorter and only contains this new content, while the second one is longer also contains older one. Our measure would assign a lower gap to the first syllabus compared to the second, even if both do an equal job in terms of covering frontier knowledge.

To address this limitation of the education-innovation gap, we construct an alternative metric which measures the *share of old knowledge* of each syllabus, defined as one minus the ratio between the number of "new words" in each syllabus (defined as knowledge words that are (a) in the top 5 percent of the word frequency among articles published between $t-3$ and $t-1$, or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$) and the number of all new words. The correlation between the share of old knowledge and the education-innovation gap is 0.22 (Figure 5, panel b), and our main results carry through if we use the former as an alternative measure of novelty of a syllabus's content (see panel b of Figure AVIII for the correlation with school-level characteristics; panel b of Figure AIX for the correlation with instructors' research productivity; and panels a and b of Table AIV for the relationship with student outcomes).

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

The tail measure is positively correlated with the education-innovation gap, with a correlation of 0.67. All our results hold when using the tail measure as a metric for syllabus novelty (see panel c of Figure AVIII, for the correlation with school-level characteristics; panel c of Figure AIX for the correlation with instructors' research productivity; and panels c and d of Table AIV for the relationship with student outcomes).

Gap with Patents The education-innovation gap is defined using new academic publications as the frontier of knowledge. It is possible for some courses, especially in scientific and technical fields, to rely less on academic content (including new) and more on technological and applied material, including the latest inventions. Our main approach could classify the course as having a large gap, in spite of it including innovative (albeit applied) content. To address this limitation, we construct a version of the education-innovation gap that uses patents in lieu of academic publications. This measure is positively correlated with the gap (Figure 5, panel d). In addition, our main results carry over when using the patent-based gap, indicating that they are not uniquely dependent on defining frontier knowledge using academic publications (see panel d of Figure AVIII, for the correlation with school-level characteristics; panel d of Figure AIX for the correlation with instructors' research productivity; and panels e and f of Table AIV for the relationship with student outcomes).

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

Taken together, the results of this section indicate that our main conclusions on the content of higher-education courses across schools and its relationship with instructors' characteristics and student outcomes are not dependent on the specific way in which we measure up-to-date content.

9 Conclusion

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

Using our measure, we document a set of new findings about the dissemination of new knowledge in US higher-education institutions. First, a significant amount of variation exists in the extent to which this knowledge is offered, both across and within schools. Second, more selective schools, schools serving students from wealthier backgrounds, and schools serving a smaller proportion of minority students offer courses with a smaller gap. Third, instructors play a large role in shaping the content they teach, and more research-active instructors are more likely to teach courses with a lower gap. Fourth, the gap is correlated with students' outcomes such as graduation rates and incomes after graduation. Taken together, our results suggest that the education-innovation gap can be an important measure to study how frontier knowledge is produced and disseminated.

References

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).

Braga, Michela, Marco Paccagnella, and Michele Pellizzari, 2016, The impact of college teaching on students' academic and labor market outcomes, *Journal of Labor Economics* 34, 781–822.

Carrell, Scott E, and James E West, 2010, Does professor quality matter? evidence from random assignment of students to professors, *Journal of Political Economy* 118, 409–432.

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* .

Courant, Paul N, and Sarah Turner, 2020, Faculty deployment in research universities, in *Productivity in Higher Education* (University of Chicago Press).

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

De Vlieger, Pieter, Brian Jacob, and Kevin Stange, 2020, Measuring instructor effectiveness in higher education, in *Productivity in Higher Education* (University of Chicago Press).

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

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

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

Feld, Jan, Nicolás Salamanca, and Ulf Zölitz, 2020, Are professors worth it? the value-added and costs of tutorial instructors, *Journal of Human Resources* 55, 836–863.

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

Goldin, Claudia Dale, and Lawrence F Katz, 2010, *The Race Between Education and Technology* (Harvard University Press).

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.

Hemelt, Steven W, Brad Hershbein, Shawn M Martin, and Kevin M Stange, 2021, College majors and skills: Evidence from the universe of online job ads, Technical report, National Bureau of Economic Research.

Hoffman, F, and P Oreopoulos, 2009, Professor qualities and student performance, *Review of Economics and Statistics* 91, 83–92.

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* .

Hoxby, Caroline M, 2020, The productivity of us postsecondary institutions, in *Productivity in Higher Education*, 31–66 (University of Chicago Press).

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.

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* .

Li, Xiaoxiao, Sebastian Linde, and Hajime Shimao, 2021, Major complexity index and college skill production, *Available at SSRN 3791651* .

Ma, Xuezhe, and Eduard Hovy, 2016, End-to-end sequence labeling via bi-directional lstm-cnns-crf, *arXiv preprint arXiv:1603.01354* .

Mountjoy, Jack, and Brent Hickman, 2020, The returns to college (s): Estimating value-added and match effects in higher education, *University of Chicago, Becker Friedman Institute for Economics Working Paper* .

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

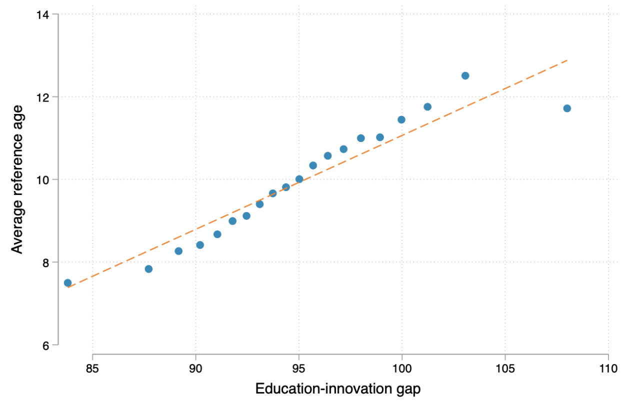
Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, 2019, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* .

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

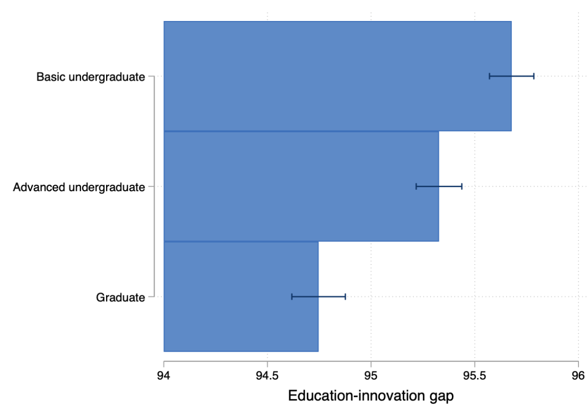
Valero, Anna, and John Van Reenen, 2019, The economic impact of universities: Evidence from across the globe, *Economics of Education Review* 68, 53–67.

Figure 1: Validating The Education-Innovation Gap

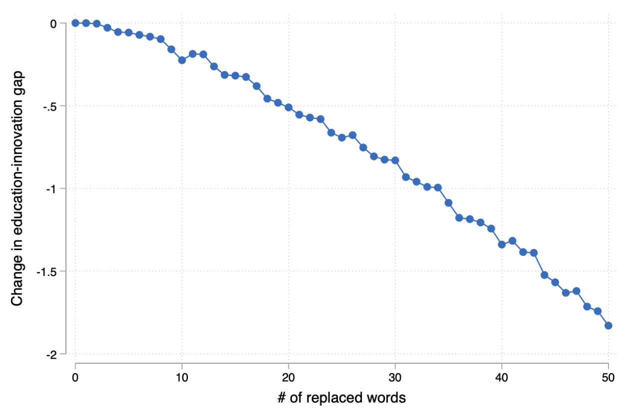
(a) Gap and Age of References Included in The Syllabi



(b) Gap by Course Level

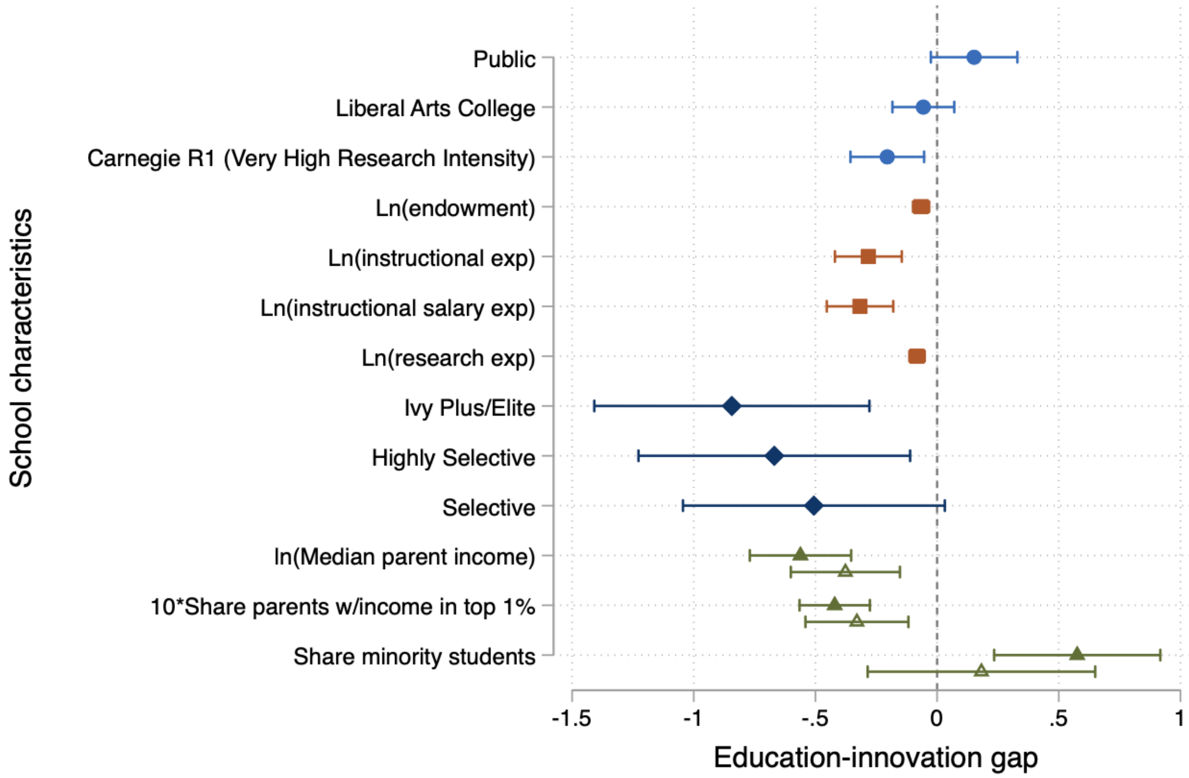


(c) Change in Gap as Old Words Are Replaced with Newer Words



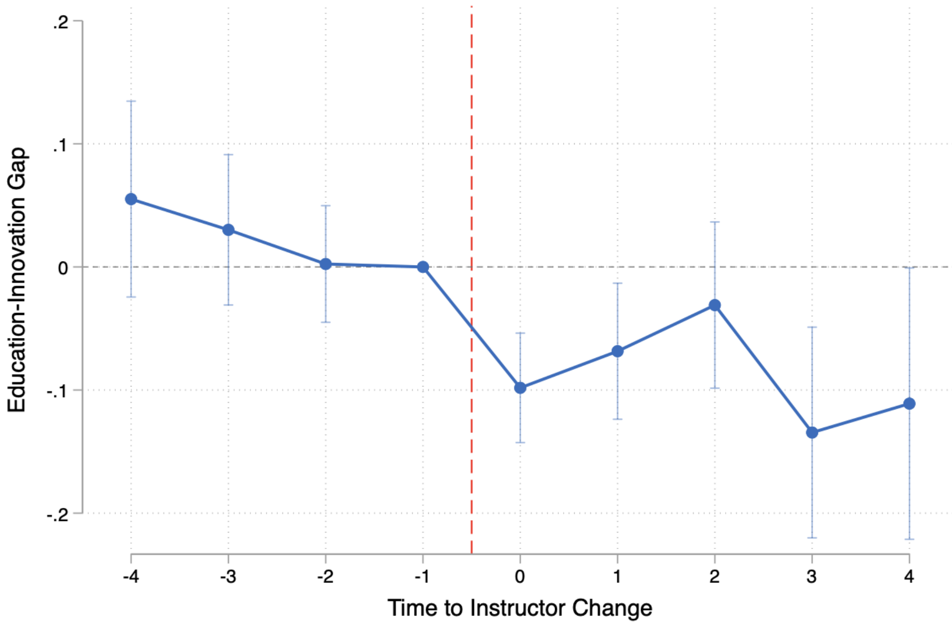
Note: Panel a) shows a binned scatterplot of the education-innovation gap and the average age of a syllabus's references (required or recommended readings), where age is defined as the difference between the year of the syllabus and the year of publication of each reference. Panel b) shows the mean and 95-percent confidence intervals of the gap by course level, controlling for field-by-year effects. Panel c) shows the change in the gap for a subsample of 100,000 syllabi, in which we progressively replace "old" words with "new" words.

Figure 2: The Education-Innovation Gap and School Characteristics



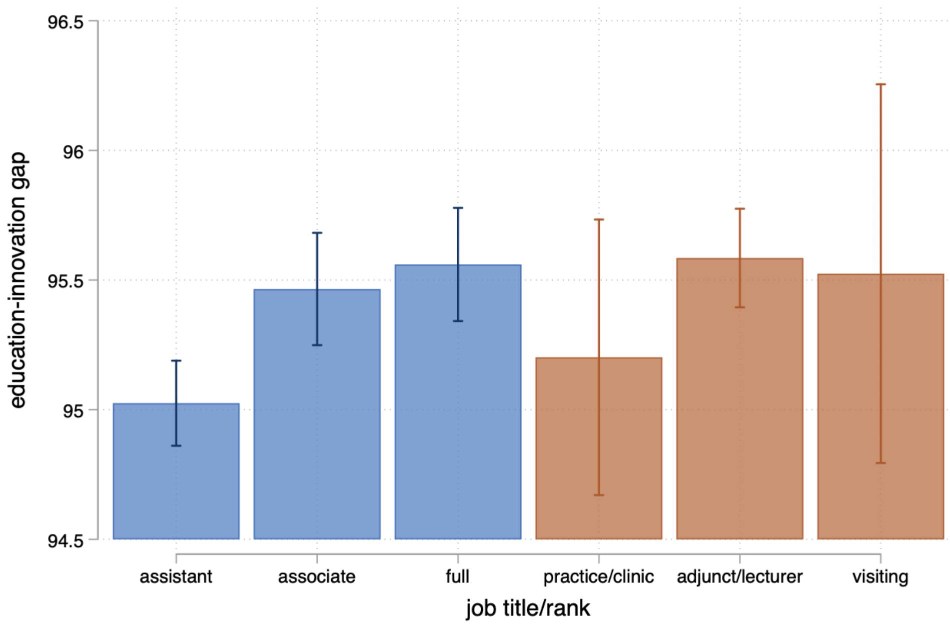
Notes: Point estimates and 95-percent confidence intervals of coefficient β in equation (3), i.e., the slope of the relationship between each reported variable and the education-innovation gap controlling for field-by-course level-by-year fixed effects. Each coefficient is estimated from a separate regression, with the exception of selectivity tiers (Ivy Plus/Elite, Highly Selective, Selective) which are jointly estimated. Endowment, expenditure, and share minority information refers to the year 2018 and is taken from IPEDS. Estimates are obtained pooling syllabi data for the years 1998 to 2018. Standard errors are clustered at the school level.

Figure 3: Event Study: The Education-Innovation Gap Around An Instructor Change



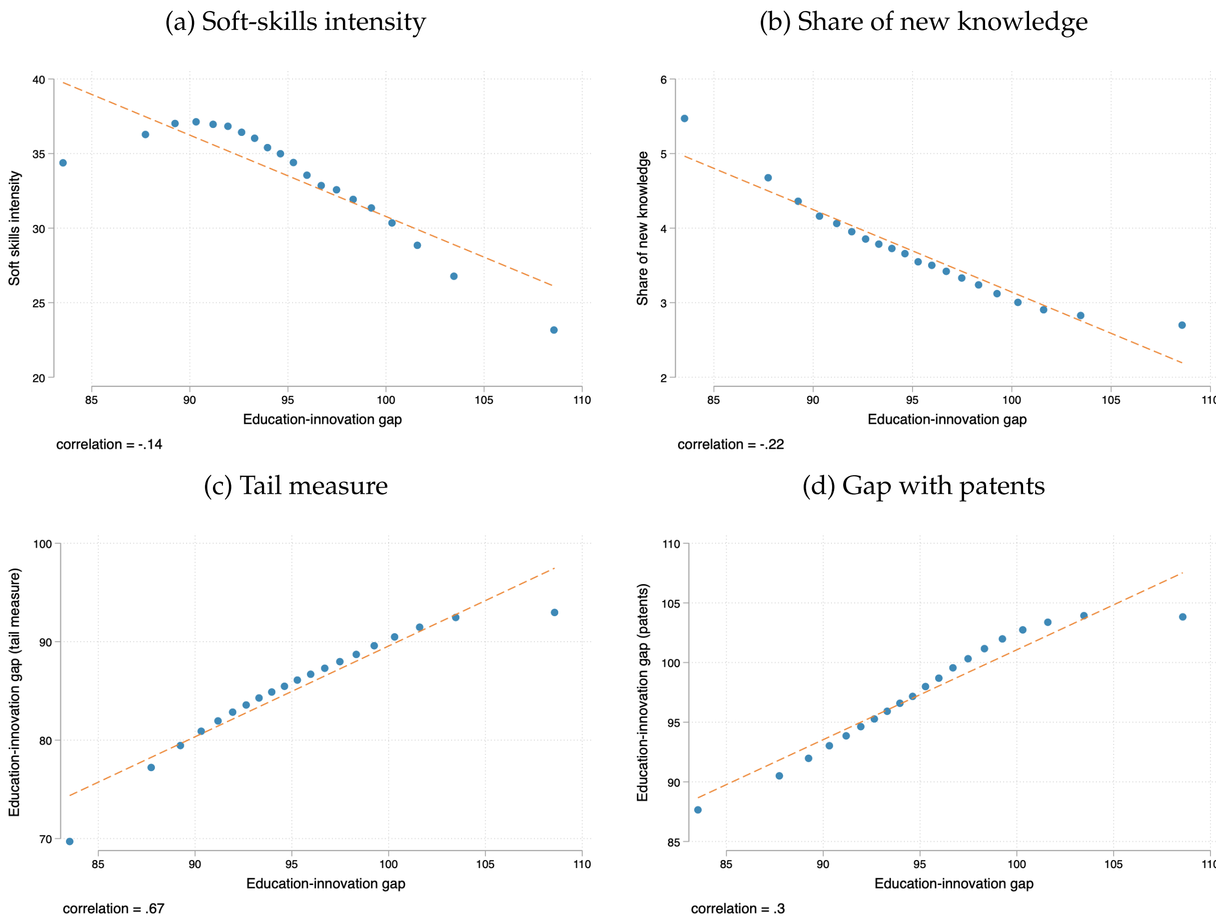
Notes: Estimates and confidence intervals of the parameters δ_k in equation (4), representing an event study of the education-innovation gap around an instructor change episode and controlling for course and field-by-year fixed effects. Observations are at the course-by-year level; we focus on courses with at most two episodes of instructor changes. Standard errors clustered at the course level.

Figure 4: Gap by Job Titles



Notes: Mean education-innovation gap by job title, along with 95-percent confidence intervals. Means are obtained as OLS coefficients from a regression of the gap on indicators for the job title of the instructor, as well as field-by-course level-by-year fixed effects. Estimates are obtained pooling data for multiple years. Standard errors are clustered at the school level.

Figure 5: The Education-Innovation Gap and Alternative Measures of Novelty: Binned Scatterplots



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

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

Panel (a): Syllabus (Course) Characteristics

	count	mean	std	25%	50%	75%
# Words	1,706,319	2226	1987	1068	1778	2796
# Knowledge words	1,706,319	1011	1112	349	656	1236
# Unique knowledge word	1,706,319	420	327	203	330	535
Soft skills	1,703,863	33.4	22.9	14.0	30.5	50.0
STEM	1,706,319	0.307	0.461	0	0	1
Business	1,706,319	0.109	0.312	0	0	0
Humanities	1,706,319	0.296	0.456	0	0	1
Social science	1,706,319	0.257	0.437	0	0	0
Basic	1,706,319	0.393	0.488	0	0	1
Advanced	1,706,319	0.275	0.446	0	0	1
Graduate	1,706,319	0.332	0.471	0	0	1

Panel (b): Instructor (Professor) Research Productivity

	count	mean	std	25%	50%	75%
Ever Published?	727,165	0.36	0.48	0	0	1
# Publications per year	262,344	1.64	2.06	1	1	1.70
# Publications, last 5 years	262,344	6.47	15.22	0	1	6
# Citations per year	262,344	35.34	108.23	0	4	26.94
# Citations, last 5 years	262,344	175.37	840.61	0	0	61
Ever Grant?	727,165	0.13	0.34	0	0	0
# Grants	97,618	13.91	25.37	3	6	14
Grant amount ($1,000)	97.62	5,332	21,100	489.89	1,621	4,672
Salary ($)	63,632	80,388	62,364	34,798	73,027	110,831

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

	count	mean	std	25%	50%	75%
Median parental income ($1,000)	767	97,917	31,054	78,000	93,500	109,900
Share parents w/income in top 1%	767	0.030	0.041	0.006	0.013	0.033
Share minority students	760	0.221	0.166	0.116	0.166	0.267
Graduation rates (2012–13 cohort)	758	0.614	0.188	0.473	0.616	0.765
Income (2003–04, 2004–05 cohorts)	762	45,035	10,235	38,200	43,300	49,800
Intergenerational mobility	767	0.294	0.138	0.182	0.280	0.375
Admission rate	715	0.642	0.218	0.533	0.683	0.800
SAT score	684	1104.4	130.5	1011.5	1079.5	1182.0

Note: Summary statistics of the variables used in the analysis.

Table 2: Selection Into The Sample: 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	
	(1)	(2)	(3)	(4)
	Corr.	SE	Corr.	SE
ln Expenditure on instruction (2013)	0.002	(0.005)	0.015	(0.010)
ln Endowment per capita (2000)	-0.001	(0.002)	-0.001	(0.002)
ln Sticker price (2013)	0.003	(0.007)	0.007	(0.010)
ln Avg faculty salary (2013)	0.016	(0.020)	0.049	(0.024)
ln Enrollment (2013)	0.018	(0.009)	0.019	(0.011)
Share Black students (2000)	-0.030	(0.038)	0.035	(0.060)
Share Hispanic students (2000)	0.171	(0.145)	0.161	(0.115)
Share Asian students (2000)	0.186	(0.214)	0.324	(0.239)
Share grad in Arts & Humanities (2000)	0.159	(0.168)	0.189	(0.179)
Share grad in STEM (2000)	-0.001	(0.028)	0.064	(0.056)
Share grad in Social Sciences (2000)	0.014	(0.024)	0.104	(0.056)
Share grad in Business (2000)	0.037	(0.065)	0.116	(0.065)
F-stat	1.015	(.)	1.376	(.)

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

	Share in OSP		ΔShare in OSP, 2010-13	
	(1)	(2)	(3)	(4)
	Mean	SE	Mean	SE
Ivy Plus/Elite	0.009	(0.003)	0.022	(0.009)
Highly Selective	0.004	(0.003)	0.006	(0.004)
Selective Private	0.034	(0.026)	0.001	(0.029)
Selective Public	0.045	(0.019)	0.009	(0.029)
F-stat	4.076	(.)	1.806	(.)

Note: The top panel shows OLS coefficients ("means") and robust standard errors ("SE") of univariate regressions of each listed dependent variable on the corresponding independent variable. The bottom panel shows OLS coefficients ("means") and syllabus-clustered standard errors ("SE") of a regression of each dependent variable on indicators for school tiers. The dependent variables are the school-level share of syllabi contained in the OSP sample in 2018 (columns 1-2) and the change in this share between 2008 and 2018 columns (3-4). The F-statistics refer to multivariate regressions that include all the listed independent variables, and test for the joint significance of these variables.

Table 3: Decomposing the Variation In The Gap: Schools, Years, Fields, Courses, and Instructors

Variable	Partial R^2	
Year	0.169	0.180
Field	0.039	0.056
School	0.021	0.028
Course level	.	0.008
Course	0.330	.
Instructor	0.248	0.346
Total	0.161	0.124

Note: The table shows a decomposition of the adjusted R^2 of a regression of the education-innovation gap on all sets of listed fixed effects into the contribution of each set of fixed effects. This is done using a Shapley-Owen decomposition, which calculates the partial R^2 of each set of variables j as $R_j^2 = \sum_{k \neq j} \frac{R^2 - R^2(-j)}{K!/j!(K-j-1)!}$ where $R^2(-j)$ is the R^2 of a regression that excludes variables j. Column 1 includes course fixed effects; column 2 only includes course level fixed effects. We use adjusted R^2 in lieu of R^2 to account for the large number of fixed effects.

Table 4: The Education-Innovation Gap Around An Instructor Change

Instructor change	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
After change	-0.0956***	-0.0993	-0.1211**	-0.0958**	-0.0225	-0.0802*	-0.0743*	-0.1145***
	(0.0244)	(0.0698)	(0.0481)	(0.0438)	(0.0418)	(0.0456)	(0.0450)	(0.0375)
N (Course x year)	379459	35598	97380	134070	94574	125469	112174	137755
# Courses	126352	11558	33209	40129	31589	43533	35386	46226
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates; one observation is a course in a given year. The dependent variable is the education-innovation gap. The variable *After change* is an indicator for years following an instructor change, for courses with only one instructor and at most two instructor changes over the observed time period. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

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

	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
Panel a): #publications	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	-0.0260	0.0527	-0.0834**	0.0432	-0.0692**	-0.0144	-0.0253	-0.0380
	(0.0186)	(0.0479)	(0.0338)	(0.0393)	(0.0293)	(0.0311)	(0.0333)	(0.0318)
2nd quartile	0.0076	0.0300		0.0789	-0.0488	0.0012	0.0306	-0.0023
	(0.0321)	(0.0742)		(0.0547)	(0.0428)	(0.0574)	(0.0590)	(0.0503)
3rd quartile	-0.0102	0.0788	-0.0480	0.0983	-0.1085**	0.0274	0.0005	-0.0513
	(0.0311)	(0.0729)	(0.0707)	(0.0598)	(0.0461)	(0.0588)	(0.0555)	(0.0479)
4th quartile	-0.1086***	0.0475	-0.1065	-0.0637	-0.1773***	-0.0337	-0.0947	-0.1731***
	(0.0386)	(0.0912)	(0.0811)	(0.0735)	(0.0626)	(0.0758)	(0.0715)	(0.0571)
Panel b): #citations	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st quartile	0.0305	0.0026	0.0478	0.1308***	-0.0522	0.0465	0.0669	-0.0092
	(0.0259)	(0.0676)	(0.0650)	(0.0473)	(0.0363)	(0.0440)	(0.0468)	(0.0433)
2nd quartile	0.0195	0.0135	-0.0315	0.1116**	-0.0424	0.0282	0.0128	0.0187
	(0.0292)	(0.0689)	(0.0682)	(0.0539)	(0.0434)	(0.0521)	(0.0523)	(0.0472)
3rd quartile	-0.0802**	-0.0254	-0.0698	-0.0611	-0.1171**	0.0081	-0.1115*	-0.1249**
	(0.0334)	(0.0788)	(0.0809)	(0.0621)	(0.0494)	(0.0638)	(0.0619)	(0.0497)
4th quartile	-0.0624	0.0774	-0.0954	-0.0119	-0.1257*	-0.0144	-0.0345	-0.1345**
	(0.0426)	(0.0963)	(0.1083)	(0.0768)	(0.0662)	(0.0826)	(0.0796)	(0.0625)
N (Course x year)	571449	59768	144945	168866	149578	207360	168992	194829
# Courses	151587	14889	39756	44848	38872	55041	42936	53539
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Table 6: The Education-Innovation Gap and The Fit Between Instructors' Research and Course Content

	All Fields	Business	Humanities	STEM	Social Science	Basic	Advanced	Graduate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fit w/top course (sd)	-0.0877**	0.1480	0.0017	-0.0836	-0.0849	-0.0637	-0.1428*	-0.0611
	(0.0398)	(0.1001)	(0.1723)	(0.0608)	(0.0656)	(0.0832)	(0.0790)	(0.0558)
N (Course x year)	54591	3199	2218	33119	12587	16743	16224	21139
# Courses	17077	1011	761	10267	3909	5208	4833	6883
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates; one observation is a course in a given year. The dependent variable is the education-innovation gap. The variable *Fit w/top course* is a measure of fit between the instructor's research and the content of the course, defined as the cosine similarity between the instructor's research in the previous 5 years and the content of the course with the smallest education-innovation gap among all courses in the same topic across all schools. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 7: The Education-Innovation Gap and Instructors' Research Resources: NSF/NIH Grants

	All Fields (1)	Business (2)	Humanities (3)	STEM (4)	Social Science (5)	Basic (6)	Advanced (7)	Graduate (8)
At least one grant	-0.0453**	-0.0020	-0.0437	-0.0388	-0.0579*	-0.0476	-0.0324	-0.0567*
	(0.0199)	(0.0580)	(0.0411)	(0.0391)	(0.0332)	(0.0327)	(0.0370)	(0.0336)
N (Course x year)	581995	59768	144945	168866	149578	210121	171867	199735
# Courses	153809	14889	39756	44848	38872	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates; one observation is a course in a given year. The dependent variable is the education-innovation gap. The variable *At least one grant* equals one if the course's instructor (or at least one of the course's instructors in case of multiple instructors) has received at least one NSF or NIH grant. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 8: The Education-Innovation Gap and Innovation Measures: Share of Undergraduate Students Who Enroll in Grad School and Total Nr of Patents

	Share of students who attend grad schoo, by fieldl						Nr
	All	STEM	Health	Business	Social Science	Humanities	Patents
Panel (a): no controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gap (sd)	-0.0044**	-0.0010	-0.0074**	-0.0003	-0.0124**	0.0054	-26.8006
	(0.0018)	(0.0022)	(0.0033)	(0.0005)	(0.0051)	(0.0076)	(16.6213)
Mean dep. var.	0.0265	0.0452	0.0249	0.0021	0.0335	0.0228	129.7813
N	65755	14714	9218	12698	14657	14468	1715
Panel (b): w/ controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gap (sd)	-0.0046**	0.0002	-0.0066**	-0.0003	-0.0101**	0.0061	-21.8882*
	(0.0021)	(0.0019)	(0.0030)	(0.0005)	(0.0046)	(0.0074)	(12.5836)
Mean dep. var.	0.0269	0.0461	0.0257	0.0021	0.0342	0.0228	131.0248
N	47723	10673	6656	9243	10645	10506	1610

Note: OLS estimates of the coefficient δ in equation (7). In columns 1-6, the variable *Gap (sd)* is a school-by-macro field-level education-innovation gap (estimated as $\theta_{s(i)}$ in equation (6), separately for each macro-field), standardized to have mean zero and variance one. In column 7, *Gap (sd)* is estimated at the school level pooling data from all fields. In columns 1-6, the dependent variable is the share of undergraduate students at each institution who attend graduate school (from the NSF Survey of Doctorate Recipients, year 2000); in column 7, it is the total number of patents filed by students at each school, from Chetty et al. (2019). All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Columns 1-6 control for year effects. Column 1 also controls for macro field fixed effects. Bootstrapped standard errors in parentheses are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 9: The Education-Innovation Gap and Student Outcomes

	Income (College Scorecard)				Income (Chetty et al., 2019)				
	Grad rate	Mean	$P_y \leq 33$ pctile	Median	Mean	P(top20%)	P(top10%)	P(top5%)	P(top20% $\|P_y$bottom 20%)
Panel (a): no controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap (sd)	-0.0513***	-0.0555***	-0.0645***	-0.0512***	-0.0722***	-0.0333***	-0.0265***	-0.0187***	-0.0293***
	(0.0068)	(0.0104)	(0.0106)	(0.0088)	(0.0124)	(0.0057)	(0.0046)	(0.0036)	(0.0053)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): w/ controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap (sd)	-0.0073**	-0.0067	-0.0083*	-0.0090**	-0.0137***	-0.0084***	-0.0053**	-0.0031**	-0.0047*
	(0.0030)	(0.0041)	(0.0050)	(0.0045)	(0.0048)	(0.0025)	(0.0021)	(0.0015)	(0.0028)
Mean dep. var.	0.5816	10.8281	10.7605	10.7096		0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

Note: OLS estimates of the coefficient δ in equation (7). The variable *Gap (sd)* is a school-level education-innovation gap (estimated as $\theta_{s(i)}$ in equation (6)), standardized to have mean zero and variance one. The dependent variable are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2019), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2019), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panel b control for control (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.