The Education-Innovation Gap*

Barbara Biasi† Song Ma‡

May 10, 2023

Abstract

This paper studies the dissemination of frontier knowledge through higher education. Applying natural language processing (NLP) techniques to the text of 1.7M university course syllabi and 20M academic articles, we construct the "education-innovation gap," a measure of a syllabus's distance from frontier knowledge. Using this measure, we document four new facts. First, courses differ greatly in their education-innovation gap, even after controlling for field, course-level, and time. Second, instructors play an important role in shaping course content. Research-active instructors teach more frontier knowledge, particularly when their research is close to the course topic. Third, access to frontier knowledge is unequal: Schools enrolling more socio-economically advantaged students offer courses with a lower gap. Lastly, students from lower-gap schools are more likely to complete a doctoral degree, produce more patents, and earn more after graduation.

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

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

*We thank Jaime Arellano-Bover, Pierre Azoulay, Nicola Bianchi, Kirill Borusyak, David Deming, Richard Freeman, Rob Jensen, David Robinson, Fabiano Schivardi, Kevin Stange, Carolyn Stein, Scott Stern, Sarah Turner and Seth Zimmerman; seminar participants at Yale, Erasmus, Maastricht, Harvard (HBS; HGSE), Ohio State, HKU, Stanford (GSB; Hoover), UCL, Queens, Stockholm School of Economics, Duke (Economics; Fuqua), IIES Stockholm, HEC (Paris), MIT (Sloan), Boston University (Wheelock), UConn, Purdue, Sciences Po, Baruch, Fed Board, Tinbergen, Queen Mary, CREST, Dartmouth, Boston College, Columbia Business School, Virginia, Michigan, USC, Boston College, LSE, Warwick, CSEF Naples, Emory; and conference participants at 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

Creating and disseminating frontier knowledge is essential for innovation and economic growth. Frontier knowledge is a key input for the production of new ideas (Jones, 2009; Moser and Voena, 2012; Williams, 2013; Iaria et al., 2018), and access to it has been linked to human capital accumulation, skill formation, and economic growth (Goldin and Katz, 2010; Acemoglu and Autor, 2011; Deming and Noray, 2020b; Bloom et al., 2021). Therefore, understanding how knowledge diffuses has crucial policy implications. Previous studies point to a fundamental role for education, particularly higher education (i.e., universities; HE in short) in this process (Valero and Van Reenen, 2019; Akcigit et al., 2020). This implies that investigating how HE content is shaped is an important first step to study the dissemination of frontier knowledge.

This line of inquiry, though, faces a major empirical obstacle: We typically do not observe course content in standard administrative data. As a result, our understanding of the diffusion of frontier knowledge in HE courses remains limited: how the presence of frontier knowledge in courses varies across HE institutions, whether instructors—who often perform both knowledge production and dissemination—play a big role in this process, and whether frontier knowledge is equally accessible to students from different socio-economic backgrounds.

In this paper, we overcome this challenge by making use of a novel source of information on educational content: the text of millions of syllabi from US colleges and universities. To quantify the diffusion of frontier knowledge through higher education, we propose a new metric: the *education-innovation gap*. This metric is designed to capture the distance between the content of a specific course, as described in each syllabus, and frontier knowledge. To measure frontier knowledge—the most advanced understanding of a particular topic or field—we rely on cutting-edge research, captured by the content of academic articles that recently appeared in high-impact factor journals. For STEM disciplines, we additionally use the text of patents to capture frontier knowledge. We then construct the gap by applying state-of-the-art text-analysis techniques that compare the content of syllabi and articles.

The first part of the paper details our methodology to quantify the presence of frontier knowledge in course content. We begin by collecting and processing the full text of 1.7 million syllabi, covering about 540,000 courses taught at 800 four-year US institutions between 1998 and 2018.[1] This sample represents about 5% of all courses taught in this time window and spans nearly all

[1] Previous studies have used syllabi to quantify the content of higher education courses. For example, Angrist and Pischke (2017) use undergraduate econometrics course syllabi from 38 universities to study the evolution of econometrics instruction.

fields. While the sample over-represents courses from very selective schools, it is representative of the population in terms of fields, course levels (basic, advanced undergraduate, and graduate), and a broad set of school characteristics. Our baseline analysis makes use of courses of all fields and course levels, to provide the broadest possible picture of the presence of frontier knowledge in HE. The granularity of our data also allows us to examine differences across fields and levels. We also collect the text of over 20 million academic articles published in top academic journals since the journals' creation.[2]

Combining information on syllabi and articles, we calculate pair-wise textual similarities between these documents, following three steps. First, we represent each document (a syllabus or an article) as a term frequency vector, projecting the text of the document on a comprehensive list of terms that refer to knowledge items. Each vector element is the frequency of a given term in the document, divided by the length of the document. Second, we use the "term-frequency-backward-inverse-document-frequency" (*TFBIDF*) approach (Kelly et al., 2021) to increase the importance of terms that most meaningfully capture a document's content. This approach gives higher weights to more informative terms and de-emphasizes terms more commonly used across all documents at a certain point in time. Third, we use these reweighted term frequency vectors to compute the cosine similarity between each syllabus and each article.

Armed with these cosine similarities, we construct the education-innovation gap of a given syllabus as the *ratio* of its average similarities with (a) "older" knowledge vintages, i.e., all articles published τ years prior to the syllabus's date and (b) "frontier" knowledge, i.e., all articles published $\tau' < \tau$ years prior to the syllabus's date. We account for cross-field differences in the speed of knowledge expansion (the rate at which new knowledge gets produced, which may be higher in some fields than others) by choosing τ and τ' to match the temporal patterns of citations to academic articles in each field. Specifically, defining citation lags as the differences between the year of publication of a citing article and that of a cited article, we select τ and τ' to be the 90th and 5th percentiles of citation lags across all the articles in the syllabus's field. This allows us to account for cross-field differences in the speed and trajectory of frontier knowledge creation.[3]

Naturally, the gap is higher for syllabi that cover more knowledge that is older, rather than newer. For example, a Computer Science course that teaches *Visual Basic* (a relatively obsolete programming language) in 2020 would have a larger gap than a course that teaches *Julia* (a more recent

[2]Studies that have used recent academic publications to capture the research frontier include Iaria et al. (2018) and Angrist et al. (2017).

[3]Our results are robust to small variations in the timing definition of old and new knowledge vintages within fields, as well as defining τ and τ' to be constant across fields.

programming language), because *Visual Basic* is mostly covered by old articles and *Julia* is mostly covered by recent articles.[4] By virtue of being constructed as a ratio of similarities, the gap is not affected by idiosyncratic attributes of a syllabus (such as length, structure, or writing style), which could introduce noise in cosine similarities as measures of content but would cancel out in a ratio measure.

Implementing the *TFBIDF*-adjustment in the construction of our measure helps to properly capture a syllabus's distinctive knowledge content. This adjustment implies that the gap does not penalize syllabi for covering "classic" or "fundamental" knowledge. Although they belong to older knowledge vintages, terms related to classic topics are still widely taught. Therefore, they appear across many documents and receive a low weight. Similarly, the *TFBIDF*-adjustment reduces the impact of "buzzwords"—terms that become popular within a time window (and therefore receive a low weight) but may not necessarily represent new breakthrough knowledge.

A few empirical regularities confirm the ability of our measure to capture the distance between course content and the knowledge frontier. For example, the gap is lower for syllabi that reference more recent articles and books in their lists of recommended readings. Moreover, the gap varies reasonably across course levels: It is the largest for basic undergraduate courses (taught in the first two years of a bachelor's degree and more likely to cover the fundamentals of a discipline) and smallest for graduate-level courses (master's and PhD). Using a simulation exercise, we also show that gradually replacing "older" knowledge in a syllabus with "newer" knowledge (i.e., words most frequently appearing in old and new articles, respectively) progressively reduces the syllabus's gap. Lastly, we show that the education-innovation gap for courses in STEM fields is very similar when we use patents (instead of academic publications) to capture the knowledge frontier.

In the second half of the paper, we use the education-innovation gap to learn how frontier knowledge is disseminated through HE. We uncover significant differences across courses in their education-innovation gap. To move a syllabus from the 25th to the 75th percentile of the gap distribution, approximately 70% of its content would have to be replaced with newer knowledge. About 25% of the overall variation in the gap occurs within the same school and course, across different instructors. The impact of instructors can also be seen from the fact that the gap of a typical course remains stable over time (i.e., instructors update syllabi very infrequently), but it declines signifi-

[4] First released in 1991, *Visual Basic* is still supported by Microsoft in recent software frameworks, but it was discontinued in 2020 (https://visualstudiomagazine.com/articles/2020/03/12/vb-in-net-5.aspx, retrieved 9/30/2020). *Julia* is a general-purpose language initially developed in 2009. Constantly updated, it is among the best for numerical analyses and computational science and is used at over 1,500 universities (https://juliacomputing.com/blog/2021/08/newsletter-august/, retrieved 9/30/2021).

cantly when the instructor of the course changes. In contrast, a much smaller share of the overall variation can be attributed to differences across schools, fields, and course levels. To account for these differences, the rest of our analysis compares syllabi within each field, course level, and year.

Instructors who are more active in producing research (i.e., they publish more, are cited more, and receive more grants) teach more frontier knowledge in their courses. This implies that research and teaching, two major tasks of a faculty member sometimes portrayed as competing with each other (Becker and Kennedy, 2005; Arnold, 2008; Hattie and Marsh, 1996; Courant and Turner, 2020), are actually complements. Instructors teaching courses closer in topic to their research are also more likely to cover frontier knowledge. Importantly, all these patterns are most pronounced for graduate-level courses. These findings are in line with a model in which the cost of keeping a course up-to-date depends on the instructor's familiarity with the research frontier. Our results also imply that proper deployment of faculty across courses can bring the content of education closer to the knowledge frontier. Investments in faculty research (both public, in the form of government grants, and made by each institution) can generate additional returns in the form of more updated instruction.

In principle, the education-innovation gap could also vary across schools, which feature different organizational models, use different educational inputs, and enroll students from different backgrounds. However, our data do not reveal a strong relationship between these factors and the education-innovation gap. For example, although our data indicate that schools that are private, are larger, and spend more teach courses with lower gaps, these relationships become insignificant once we allow these characteristics to jointly explain the gap.

We do, however, find that the education-innovation gap correlates with two sets of school-level characteristics. The first is the average number of faculty publications, which bears a significant, negative relationship with the gap even after controlling for other cross-school differences such as school size, research, and teaching expenses. This result is in line with our previous findings on the role of instructors; furthermore, it confirms that research activities complement teaching and the dissemination of frontier knowledge.

The second set of characteristics refers to students' socioeconomic backgrounds. The education-innovation gap is significantly higher in schools enrolling students with lower median parental income and schools with a higher share of Black or Hispanic students. This relationship remains after accounting for cross-school differences in sector, enrollment, expenditures, faculty composition, and even student preparedness (measured using the SAT and ACT scores of admitted students).

Our data also indicate that students from schools that offer courses with a lower gap are more likely to obtain a PhD and produce more patents. In addition, they are more likely to graduate on time and have higher earnings. One explanation for this finding is that lower-gap courses attract students with a higher propensity to innovate in the future. An alternative explanation is that access to frontier knowledge through college courses promotes the creation of new knowledge (Biasi and Moser, 2021; Iaria et al., 2018), which makes students more likely to engage in innovative activities. While a formal test of the causal link between the education-innovation gap and student outcomes is beyond the scope of this paper, both of these explanations indicate that schools could benefit from expanding the coverage of frontier knowledge in the courses they offer.

In the final part of the paper, we probe the robustness of our results to the use of alternative measures of frontier knowledge coverage. We consider three of them: the share of all "new" knowledge contained in a syllabus, designed to avoid penalizing a syllabus for containing old knowledge in addition to new one; a measure of "tail" knowledge, aimed at capturing the presence of the most recent content; and the education-innovation gap obtained using patent filings as a measure of frontier knowledge. All these alternative measures are strongly correlated with the baseline version of the education-innovation gap, and our main results are qualitatively unchanged when we use them in lieu of the gap.

The main contribution of our paper is a new text-based method to quantify knowledge diffusion through higher education, which we implement on a novel dataset combining syllabi and academic publications. Our metric, the underlying algorithm, and the methodology in general—which will all be made available to researchers—can be used for future analyses of education and innovation. Our analysis provides new evidence about the dissemination of frontier knowledge in HE and highlights key economic forces behind this process. In doing so, this paper sheds new light on some central questions related to innovation and HE and contributes to several strands of literature.

First, we contribute to the literature that has characterized heterogeneity in the production of human capital, focusing on differences in the returns to educational attainment (Hanushek and Woessmann, 2012), majors (Altonji et al., 2012; Deming and Noray, 2020a), college selectivity (Hoxby, 1998; Dale and Krueger, 2014), and the skill content of college majors (Hemelt et al., 2021; Li et al., 2021). In this paper we take a novel approach: We directly examine curricula and educational content, among the most central components of higher education. With this approach, we document significant differences in the knowledge covered by each course, which could have important implications for students.

Second, we provide direct evidence of the importance of instructors in shaping the content of higher education. 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 about why and how instructors impact students (De Vlieger et al., 2020). We study instructors' contributions to the production of educational content and carefully characterize differences across instructor types. Crucially, our findings highlight complementarities between teaching and research activities.

Third, we highlight differences in the ability of higher education programs to equip students with the knowledge necessary to innovate, which originate from heterogeneous course contents. Importantly, these differences confirm a "lack of democratization" in access to valuable knowledge. US inventors have been shown to come from a small set of schools, enrolling very few low-income students (Bell et al., 2019). We find that these schools provide the most up-to-date educational content, which in turn suggests that access to frontier knowledge is highly unequal.

Lastly, our results are related to the literature on the production of ideas, knowledge, and innovation. Existing studies have shown how access to existing frontier knowledge can foster the creation of new knowledge. For example, Moser and Voena (2012), Williams (2013), and Galasso and Schankerman (2015) show how, in various settings, easier access to pre-existing patents fosters the creation of new patents. Similarly, Iaria et al. (2018) show that reduced scientific cooperation due to World War II leads to a slow-down in the production of new science, and Biasi and Moser (2021) show that a decline in the cost of accessing frontier knowledge in books leads to an increase in the diffusion of those books. Education systems have been pointed as central for the dissemination of frontier knowledge, particularly in fields such as STEM (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019; Akcigit et al., 2020).[5] We contribute to these works by documenting where frontier knowledge can be accessed within the HE system.

2 Data

Our empirical analysis combines data from multiple sources. These include the text of course syllabi; the abstract of academic publications; job titles, publications, and grants of each instructor; characteristics of US higher education institutions; and labor market outcomes and innovation activities of the students at these institutions. More detail on the construction of our final data set can

[5]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. Educational institutions also play a crucial role in fostering entrepreneurship (Tartari and Stern, 2021).

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 the Open Syllabus Project (OSP), a non-profit organization that collects these data by crawling publicly accessible university and faculty websites to support educational research and its applications. The initial sample contains nearly seven million syllabi of courses taught in over 80 countries between 1990 and 2018.

Most syllabi share a standard structure. The standard syllabus begins with basic details of the course (such as title, code, and the name of the instructor). It proceeds with a short description of its content, followed by a more detailed list of topics and required or recommended readings for each class session. Most syllabi contain information on evaluation criteria, such as assignments and exams; some also include general policies regarding grading, absences, lateness, and misconduct. Following this general structure, we parse each syllabus and extract four sets of information, which we describe in detail below: (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, the quarter or semester, and the academic year in which the course is taught. Course titles and codes allow us to classify each syllabus into one of three course levels: basic undergraduate, advanced undergraduate, or graduate. OSP assigns each syllabus to one of 69 detailed fields. 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 most cases) a timeline of the content and the materials for each lecture.

Reference list 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 (2023)." We are able to compile a list of references for 71 percent of all syllabi. We then collect bibliographic information on each reference from Elsevier's

[6]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. Appendix Table BV 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 BIV.

SCOPUS database (described in more detail in Section 2.2); this includes title, abstract, journal, keywords (where available), and textbook edition (for textbooks).

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 100 syllabi in our sample. We remove universities that exclusively or primarily focus on online instruction. We also exclude 35,917 syllabi (1.9 percent) with fewer than 20 words or more than 10,000 words (the bottom and top 1 percent of the length distribution).

One of the advantages of our data is its breadth, which allows us to examine courses in all fields and course levels. We believe studying all of them is important. While basic courses should in principle be more focused on core discipline concepts, they may also cover frontier knowledge. Similarly, STEM and more technical fields are naturally associated with innovation and frontier knowledge. However, many scholars have noted how also the arts and humanities play an important role in the knowledge economy (Bullen and Robb, 2004). For these reasons, in our analysis, we retain courses of all fields and course levels and present all our tests separately by macro-field and course level.

Our final sample, described in panel (a) of Table 1, contains about 1.7 million syllabi of 542,251 courses at 767 institutions. Thirty-three percent of all syllabi cover STEM courses, ten percent cover Business, 30 percent cover Humanities, and 24 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 1,778. Our textual analysis focuses on "knowledge" words, i.e., words that belong to a dictionary, a list of words compiled to capture a document's academic content (defined in greater detail in Section 3). The average syllabus contains 420 unique knowledge words, with a median of 330.

2.2 Academic Publications

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. Top journals are defined as those ranked among the top 10 by Impact Factor (IF) in any of SCOPUS's 191 fields at least once since 1975 (or the journal's creation, if it occurred after 1975).[8] Our final list of publications includes 20 million articles, corresponding to approximately 100,000 articles per year. For each article, we extract information on its title, abstract, keywords, authors, and authors' affiliations.

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

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

Panel (a): Syllabi Characteristics	count	mean	std	25%	50%	75%
Education-innovation gap	1,706,319	93.9	6.9	89.8	93.6	97.7
# 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
STEM	1,706,319	0.326	0.469	0	0	1
Business	1,706,319	0.103	0.304	0	0	0
Humanities	1,706,319	0.299	0.457	0	0	1
Social science	1,706,319	0.240	0.427	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): Instructors' Research Productivity	count	mean	std	25%	50%	75%
Ever Published?	332,064	0.41	0.49	0	0	1.00
# Publications per year	135,364	1.51	1.94	1.00	1.00	1.38
# Publications, last 5 years	111,404	6.01	14.89	0	1.00	5.42
# Citations per year	135,364	29.22	105.92	0	1.85	17.92
# Citations, last 5 years	111,404	172.46	887.99	0	0	54.32
Ever Grant?	332,064	0.18	0.38	0	0	0
# Grants	58,136	10.14	19.96	2.00	4.00	10.00
Grant amount ($1,000)	54,462	4,023	19,501	236	912	3,201
Panel (c): Students' Characteristics and Outcomes at the School 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.

2.3 Alternative Data Source to Capture Frontier 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 collect the text of more than six million patents issued since 1976 from the US Patent and Trademark Office (USPTO) website. We capture the content of

each patent with its title and abstract.

2.4 Instructors: Research Productivity, Funding, and Job Titles

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

Research productivity Individual-level publications and citations data are from Microsoft Academic (MA). As one of the world's top academic search engines, MA listed publications, working papers, other manuscripts, and patents for each researcher, together with citation counts for these documents, until its discontinuation in December 2021. We link MA records to syllabi via fuzzy matching based on instructor name and institution (details on this procedure are in Appendix B). We are able to successfully find 41 percent of all instructors, and we assume that instructors without a MA profile never published any article in a top academic journal tracked by MA (Table 1, panel (b)), an assumption that is supported by manual random searches.

Using data from MA, we measure the quality and quantity of each instructor's research output with the number of articles published and citations received in the previous five years.[9] On average, instructors published 6 articles in the previous five years, with a total of 172 citations (Table 1, panel (b)). The distributions of citation and publication counts are highly skewed: The median instructor in our sample only published one article in the previous five years and received no citations.

Funding We also collect information on US government grants received by each instructor, which allows us to measure public investment in academic research. We focus on two of the main funding agencies of the US government: the National Science Foundation (NSF) and the National Institute of Health (NIH).[10] Our grant data include 480,633 NSF grants active between 1960 and 2021 (with an average size of $582K in 2019 dollars) and 2,566,358 NIH grants active between 1978 and 2021 (with an average size of $504K). We link grants to instructors via fuzzy matching between the name and institution of the investigator and those of the instructor (more details can be found in Appendix B). Eighteen percent of all syllabi instructors are linked to at least one grant. Among these, the median instructor receives four grants, with a combined size of $912K (Table 1, panel (b)).

Job titles In many US states, information on public college and university employees is disclosed online, to comply with state regulations on transparency and accountability. These records usually

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

[10] 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.

contain each employee's name and job title. We are able to collect information on job titles for 32,090 instructors in our syllabi sample (9.7 percent of all instructors and 13 percent of public-sector instructors), employed in 278 public institutions in 13 states. We are able to observe instructors' titles for the most recent years (the modal year is 2017; we detail the coverage of these data in Appendix B). Among all syllabi instructors for which we have job title information, 42 percent are ladder faculty (including 11 percent who are assistant professors, 13 percent who are associate professors, and 18 percent who are full professors; Appendix Figure AI).

2.5 Information on US Higher Education Institutions

The last component of our data set includes information on all US colleges and universities covered in our syllabi sample. Our primary source is the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES).[11] For each school, IPEDS reports a set of institutional characteristics (such as name and address, public or private sector, 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 using 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 data set assembled and used by Chetty et al. (2020), 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 a measure of intergenerational mobility (the share of students with parental income in the bottom quintile who reach the top income 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 source is the Survey of Earned Doctorates, conducted by the NSF, which reports the characteristics of all doctoral degree recipients in US institutions each year. We use information on students' graduating cohorts and bachelor's institutions to construct the share of undergraduate students in each school and graduation year who eventually complete a doctoral degree for the years 1998-2018.[12] The third is the College Scorecard

[11] 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.

[12] The Survey of Earned Doctorates has been conducted since 1957. To assign a doctoral degree recipient to their bachelor's degree cohort, we subtract six from their year of doctoral degree completion.

Database of the US Department of Education, an online tool designed to help users compare the costs and returns of attending various colleges and universities in the US. This database reports the earnings 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. On average, the median parental income of all students at each school is $97,917. Across all schools, three percent of all students have parents with incomes in the top percentile. The share of minority students equals 0.22. 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' average intergenerational mobility is equal to 0.29; this implies that, in the colleges and universities in our sample, over one-fourth of graduates with parental income in the bottom quintile reach the top quintile of personal income in adulthood.

2.6 Data Coverage and Sample Selection

Our syllabi sample only covers a fraction of all courses taught in US schools between 1998 and 2018.[13] To more accurately interpret our empirical results, it is crucial to clarify patterns of selection into the sample. To do so, we compile 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 course catalogs in the archives of each school.[14] 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 remains stable over time, at 5 percent (Appendix Figure AIII). This suggests that, among these randomly selected schools, 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. Between 2010 and 2018, STEM courses represent 33 percent of syllabi in our sample and 24 percent of courses in the catalog; Humanities represent 30 and 32 percent, and Social Sciences represent 24 and 20 percent, respectively (Appendix Figure AIV). Similarly, basic undergraduate courses represent 39 percent of syllabi in our sample and 31 percent of courses in

[13] 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).

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

the catalog; advanced undergraduate courses represent 28 and 30 percent, and graduate courses represent 33 and 38 percent (Appendix Figure AV). These shares are fairly stable over time.

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

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

	Share in OSP, 2018		ΔShare in OSP, 2008-18	
	(1)	(2)	(3)	(4)
	Corr.	SE	Corr.	SE
ln Expenditure on instruction	0.002	(0.005)	0.015	(0.010)
ln Endowment per capita	-0.001	(0.002)	-0.001	(0.002)
ln Sticker price	0.003	(0.007)	0.007	(0.010)
ln Avg faculty salary	0.016	(0.020)	0.049	(0.024)
ln Enrollment	0.018	(0.009)	0.019	(0.011)
Share Black students	-0.030	(0.038)	0.035	(0.060)
Share Hispanic students	0.171	(0.145)	0.161	(0.115)
Share Asian students	0.186	(0.214)	0.324	(0.239)
Share grad in Arts & Humanities	0.159	(0.168)	0.189	(0.179)
Share grad in STEM	-0.001	(0.028)	0.064	(0.056)
Share grad in Social Sciences	0.014	(0.024)	0.104	(0.056)
Share grad in Business	0.037	(0.065)	0.116	(0.065)
F-stat	1.015		1.376	
F-stat p-value	0.442		0.194	

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

	Share in OSP, 2018		ΔShare in OSP, 2008-18	
	(1)	(2)	(3)	(4)
	Mean	SE	Mean	SE
Ivy Plus/Elite	0.024	(0.008)	0.022	(0.009)
Highly Selective	0.003	(0.003)	0.006	(0.004)
Selective Private	0.029	(0.018)	0.001	(0.029)
Selective Public	0.040	(0.023)	0.009	(0.029)
F-stat	3.677		1.806	
F-stat p-value	0.008		0.136	

Note: The top panel shows OLS coefficients ("Corr.") and robust standard errors ("SE") of univariate regressions of each listed dependent variable on the corresponding independent variable. The bottom panel shows OLS coefficients ("Mean") 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 (Open Syllabus Project) 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.

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 a 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, Asian), and the share of students graduating in Arts and Humanities, STEM, Social Sciences, and Business. We also test for the joint significance of all these variables. We find that these variables are individually and jointly uncorrelated with the share of courses in the syllabi sample, with an F-statistic close to one. In column 2 we repeat the same exercise, using the 2008-2018 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 2.4 percentage point higher share of courses included in the syllabi sample, and selective public schools have a 4.0 percentage point 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, 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

This section describes the construction of the education-innovation gap. We first explain how we measure textual similarities between course syllabi and academic publications. Then, we define and construct the gap using measures of similarity, implementing a series of adjustments to better capture each syllabus's content. Lastly, we validate our measure and describe its variation. Appendix C provides additional details on the construction of the measure.

3.1 Measuring The Similarity Between Syllabi and Academic Publications

3.1.1 Constructing Term Frequency Vectors

We start by representing each document d (a syllabus or an article) as a term-frequency vector $\mathbf{TF_d}$. Each element TF_{dw} of $\mathbf{TF_d}$ represents the frequency of term w in d:

$$TF_{dw} \equiv \frac{c_{dw}}{\sum_{k \in W} c_{dk}},$$

where, in the numerator, c_{dw} counts the number of times term w appears in d and the denominator is the total number of terms in d. To maximize our ability to capture the knowledge content of each

document, we construct **TF** vectors focusing exclusively on terms related to knowledge concepts and skills, belonging to a dictionary W with $|W|$ terms (as a result, each term vector contains $|W|$ elements). 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.[15]

3.1.2 Adjusting for Term Relevance

When constructing similarity metrics, it is crucial to ensure that each term receives a weight proportional to its importance in capturing a document's content. Unadjusted **TF** vectors give more weight to terms with a higher document frequency. However, terms that are very common across *all* documents receive more weight regardless of their ability to capture the content of a given document. For example, terms such as "Programming" or "Animals"—very common among Computer Science and Biology syllabi, respectively—are usually less informative of content than terms such as "Natural Language Processing" or "CRISPR."[16]

To this purpose, we use a leading approach in the text analysis literature called "term-frequency-inverse-document-frequency" (TFIDF, Kelly et al., 2021). This approach assigns each term a weight inversely proportional to the frequency of the term across all documents, underweighting terms that are very common and thus not diagnostic of a document's content. We implement this approach by constructing an inverse-document frequency vector **IDF** (of length $|W|$) with elements defined as

$$IDF_w \equiv \ln\left(\frac{|D|}{\sum_{n\in D} \mathbb{1}(c_{nw} > 0)}\right),$$

where D is the set of all documents (syllabi *and* articles). The denominator in parentheses is the total number of documents that contain word w. IDF_w is thus the inverse of the share of all documents containing word w. Using **IDF**, we can then transform $\mathbf{TF_d}$ into a term-frequency-inverse-document-frequency vector $\mathbf{TFIDF_d}$, with elements equal to

$$TFIDF_{dw} = TF_{dw} \times IDF_w. \tag{1}$$

Accounting for changes in term relevance over time The generic TFIDF approach discussed above calculates the relative importance of each term for a given document pooling together doc-

[15] 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.

[16] Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) is a family of DNA sequences found in the genomes of prokaryotic organisms such as bacteria and archaea. The term also refers to a recent technology that can be used to edit genes.

uments published in different years. This is not ideal for our analysis, because we are interested in the novelty of the content of a syllabus d relative to research published in the years *prior* to d. Consider, for example, course CS229 at Stanford University, taught by Andrew Ng in the early 2000s and one of the first that entirely focused on *Machine Learning*. The term "machine learning" has become very popular in later years, so its frequency across all documents is very high and its IDF_w very low. Pooling together documents from different years would thus result in a very low $TFIDF_{dw}$ for the term "machine learning" in the course's syllabus, failing to recognize the course's novelty as of early 2000s. Generally, not accounting for changes in term frequency over time would lead us to severely mischaracterize a course's path-breaking content.

To overcome this issue, we modify the traditional **TFIDF** and construct a retrospective or "point-in-time" version of **IDF**, meant to capture the inverse frequency of a term among all documents published *prior to* d. We call this vector "backward-**IDF**," or $\mathbf{BIDF_t}$. It is indexed by t because it varies over time. We define the set of documents published prior to t as D_t; the elements of $\mathbf{BIDF_t}$ can be defined as

$$BIDF_{tw} \equiv \ln\left(\frac{|D_t|}{\sum_{n \in D_t} \mathbb{1}(c_{nw} > 0)}\right).$$

The use of this weighting approach allows us to give a temporally appropriate weight to each term in a document. Using $\mathbf{BIDF_t}$, we can then calculate a "backward" version of $\mathbf{TFIDF_d}$—called $\mathbf{TFBIDF_d}$—whose elements are

$$TFBIDF_{dw} = TF_{dw} \times BIDF_{t(d)w}, \tag{2}$$

where $t(d)$ is the publication year of document d.

3.1.3 Building Textual Similarities Between Syllabi and Articles

Armed with weighted term vectors, we can now construct measures of textual similarities between syllabi and articles. For simplicity, we denote $\mathbf{TFBIDF_d}$ as $\mathbf{V_d}$ for each d. The measure of similarity we use is the cosine similarity, defined for two documents d and d' as

$$\rho_{d,d'} = \frac{\mathbf{V_d}}{\|\mathbf{V_d}\|} \cdot \frac{\mathbf{V_{d'}}}{\|\mathbf{V_{d'}}\|} \tag{3}$$

where $\|\mathbf{V_d}\|$ is the Euclidean norm of $\mathbf{V_d}$. Since each element of $\mathbf{V_d}$ is non-negative, ρ lies in the interval $[0, 1]$. If d and d' use the exact same set of terms with the same frequency, $\rho_{d,d'} = 1$; if they

have no terms in common, $\rho_{d,d'} = 0$.

3.2 Calculating the Education-Innovation Gap

We capture the similarity between each syllabus d and different vintages of knowledge using the average similarity of d with all the articles published in a three-year time period ending τ years before the syllabus year $t(d)$:

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

where ρ_{dk} is the cosine similarity between syllabus d and an article k, $\Omega_\tau(d)$ is the set of all articles published in the three-year time interval $[t(d) - \tau - 1, t(d) - \tau + 1]$, and $|\Omega_\tau(d)|$ is the total number of these articles.[17]

We construct the education-innovation gap as the ratio between the average similarity of a syllabus with older technologies (published in the interval $[t(d) - \tau - 1, t(d) - \tau + 1]$) and the similarity with more recent ones (published in $[t(d) - \tau' - 1, t(d) - \tau' + 1]$, where $\tau' < \tau$):

$$Gap_d \equiv 100 * \left(\frac{S_d^\tau}{S_d^{\tau'}} \right) \tag{4}$$

Given this definition, the syllabus of a course taught in t has a lower education-innovation gap if its text is more similar to more recent research (published in $[t(d) - \tau' - 1, t(d) - \tau' + 1]$) than to older research (published in $[t(d) - \tau' - 1, t(d) - \tau' + 1]$). We multiply the ratio by 100 for readability.

In order to implement the measure in (4), we need to determine the appropriate τ and τ', the parameters that define the timing of the old and new knowledge vintages. We choose them to be field-specific, to accommodate the fact that knowledge production may be a faster process in some fields (in which case τ and τ' are both small) than in others (in which case τ and τ' will both be large). To choose the appropriate τ and τ', we look at the distribution of the age of citations in each field's publications. Specifically, we select each field's τ and τ' to be the 90th and 5th percentiles of citation lags across all the articles in that same field. The median 5th percentile of the citation lag, meant to capture recent knowledge vintages, is 2 years. The 90th percentile, which captures older vintages, has a median of 16 and ranges between 12 years for Medicine and Nursing and 36 years for Religion. Field-specific lags are shown in Appendix Figure AVI.

Our measure features two attractive properties. First, being constructed as a ratio, the gap is not affected by syllabus-specific attributes such as style, format, or length, which could introduce noise when measuring a syllabus's similarity to knowledge. For example, two courses covering the same

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

materials could have different similarities to research publications if one syllabus is more detailed or uses more academic terms. Since these stylistic differences would affect both the numerator and the denominator, the ratio would net them out. We illustrate this point with a simulation exercise in Appendix C.[18]

Second, our measure does not heavily penalize syllabi for covering "classic" topics in the literature, as long as these are widespread across courses. This is guaranteed by the use of a $TFBIDF$ approach, which reduces the impact on the gap of terms—such as those pertaining to classics—frequently used across all documents. For example, the term "Ordinary Least Squares" ("OLS") refers to a relatively old but very common concept taught in most econometrics and statistics courses. As such, it will receive a low weight, and syllabi will not be penalized much by covering it.

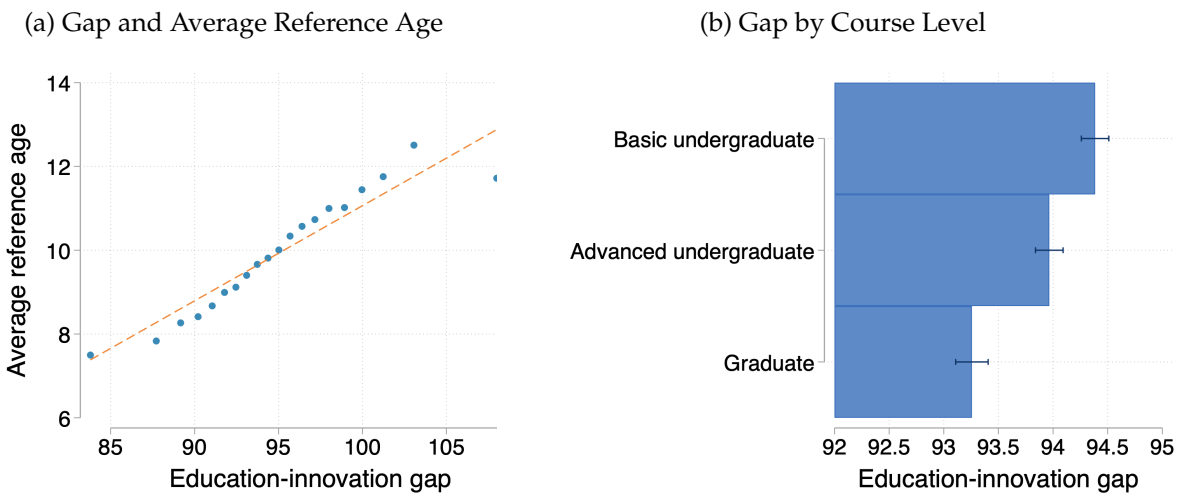
3.3 Validating The Measure and Interpreting Its Magnitude

We perform a series of tests to validate our measure's ability to capture the distance between the content of a course and the research frontier. First, we show a positive relationship between the gap and the average age of its reference list (defined as the average difference between the year of the syllabus and the publication year of each reference (Figure 1, panel (a)). This correlation is small in magnitude and equal to 0.11. This is not surprising: Syllabi often only reference a textbook, whose date of publication is not necessarily informative of content. Notably, the correlation is larger for graduate-level courses (particularly those in the top decile of the length distribution, equal to 0.17), for which where the reference list is more likely to reflect actual content. In spite of the fact that the average reference age is easy to calculate, our text-based measure is available for all syllabi (including those for which the reference list is unavailable) and is more accurate in capturing the content of courses that only rely on very few bibliographic sources (for example, a textbook).

Second, we show that the gap varies reasonably across course levels. Graduate-level courses and advanced undergraduate courses have lower gaps than basic undergraduate courses. Controlling for field-by-year effects, basic undergraduate courses have a gap of 94.4, advanced undergraduate courses have a gap of 94.0, and graduate courses have a gap of 93.2 (Figure 1, panel (b)). This confirms the intuition that more advanced courses cover content that is closer to the knowledge frontier.

[18] We manually create a sample of 1.7 million simulated syllabi, for which we know ex ante the ratio between "old" knowledge terms (more popular among old publications) and "new" knowledge terms (more popular among recent publications). In the presence of syllabi idiosyncracies, the education-innovation gap performs significantly better at recovering a syllabus's knowledge content (the ex ante ratio between old and new knowledge terms) than a simple measure of similarity with new terms (Appendix C).

Figure 1: Validating the Education-Innovation Gap

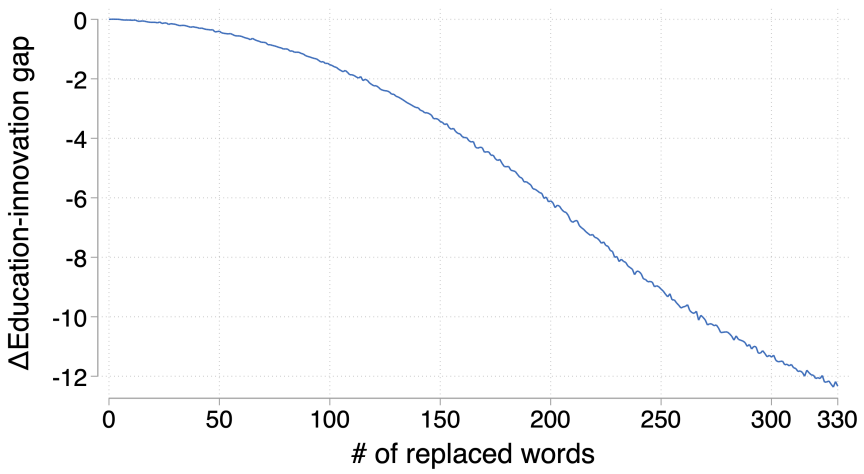


Note: Panel (a) shows a binned scatterplot of the education-innovation gap and the average age of a syllabus's references (required or recommended readings), in which reference age is calculated as the difference between the year of the syllabus and the year of publication of each reference. Panel (b) shows the mean and 95-percent confidence intervals of the gap by course level, controlling for field-by-year effects.

Third, we use a simulation exercise to confirm that our measure is able to capture even small changes in a syllabus's coverage of different knowledge vintages. We first randomly draw a subsample of 100,000 syllabi. In these syllabi, we progressively replace terms that are more frequent in older knowledge vintages ("old words") with terms more frequent in newer vintages ("new words"), and we re-calculate the gap as we replace more words. Old words are either (a) in the top 5 percent in terms of frequency in the old publication corpus (between $t - \tau - 1$ and $t - \tau + 1$), or (b) in the old publication corpus but not in the new publication corpus (between $t - \tau' - 1$ and $t - \tau' + 1$), where t is the year of the syllabus and τ and τ' are field-specific and constructed as in Section 3.2). New words are defined in a symmetric way, as either (a) in the top 5 percent in terms of frequency in the new publication corpus, or (b) in the new publication corpus but not in the old publication corpus.

The gap monotonically decreases as we replace more old words with new ones. We show this in Figure 2, which plots the median change in the education-innovation gap when we replace a given number of words in the syllabi subsample. This simulation is also useful for gauging the economic magnitude of changes in the gap. In particular, a unit change in the gap is equivalent to the replacement of 26 percent of a syllabus's old words (or 85 old words out of 330 words for the median syllabus).

Figure 2: Change in Gap as Old Knowledge Is Replaced With New Knowledge



Note: Median change in the gap in a simulation exercise based on a random subsample of 100,000 syllabi, in which we progressively replace "old" knowledge words with "new" knowledge words.

3.4 The Education-Innovation Gap: Variation and Variance Decomposition

The average course has a gap of 93.9, with a standard deviation of 6.9, a 25th percentile of 89.8, and a 75th percentile of 97.7 (Table 1, panel (a) and Appendix Figure AVII). To give an economic meaning to this variation, we use the relationship illustrated in Figure 2. In order to move a syllabus from the 75th to the 25th percentile of the distribution (a 7.9 change in the gap), we would have to replace about 70% of the knowledge content of the median syllabus (330 words).

To better understand what drives variations in the gap, we calculate the contribution to the total variance of each of the following five attributes: year, field, school, course, and instructor. We perform this decomposition by means of a Shapley-Owen decomposition (Israeli, 2007; Huettner et al., 2012).

The method proceeds in three steps. We first estimate OLS regressions of the gap on fixed effects for all possible combinations of attributes.[19] Second, for each of these regressions, we compute how much the adjusted R^2 declines if we exclude the fixed effects for a specific attribute j. Lastly, we calculate the average decline for each j across all these regressions, which we denote as the *partial-R^2* of attribute j, or R^2_j. This statistic, which is analogous to the Shapley value used in game theory, represents the portion of the total variance in the education-innovation gap that can be attributed

[19]Since school effects are subsumed by course effects (each course is taught only at one school), school effects are not separately identified in a regression that also contains course fixed effects. Our method, however, still allows us to quantify the contribution of school effects to the total variation in the education-innovation gap out of the regressions of those combinations of the five attributes that do not include course effects.

to j. Analytically, it is equal to

$$R_j^2 = \sum_{T \subseteq V \setminus \{j\}} \frac{|T|!(K - |T| - 1)!}{K!} [R^2(T \cup \{j\}) - R^2(T)]$$

where $R^2(S)$ is the adjusted R^2 of a regression of the gap on fixed effects for a set of factors S, V is the set of all attributes considered, $|T|$ is the number of attributes in set T, and $K \equiv |V| = 5$ is the total number of attributes considered. The use of adjusted R^2 accounts for the fact that the various sets of fixed effects have different numbers of categories (using the standard R^2, larger categories would mechanically explain a larger portion of the variance).[20]

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

	Partial R^2					
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Instructor	0.364	–	0.251	0.347	0.248	0.342
Field	0.072	0.121	0.045	0.066	–	–
Year	0.101	0.115	0.094	0.097	–	–
School	–	0.036	0.016	0.021	0.016	0.022
Course	–	–	0.355	–	0.353	–
Course level	–	–	–	0.008	–	0.008
Field-by-year	–	–	–	–	0.161	0.188
All	0.537	0.271	0.760	0.538	0.779	0.560

Note: This table shows a Shapley-Owen decomposition of the adjusted R^2 of a regression of the education-innovation gap into the contribution of each set of fixed effects. The detailed method of the decomposition is described in detail in Section 3.4. *All* reports the adjusted R^2 of a regression with all sets of fixed effects included. We use adjusted R^2 in lieu of R^2 to account for the large number of fixed effects.

The results of this decomposition exercise are shown in Table 3. A large portion of the variation in the education-innovation gap is attributable to instructors. For example, when we use instructor, field, and year effects (column 1), instructors explain 36 percent of the total variance (Table 3, column 1). When we add school and course effects, instructors explain 25 percent of the variance (column 3). This implies that there is significant variation in the gap even within the same course in a school, and this variation is attributable to instructors. Schools, on the other hand, explain a

[20] We perform a placebo test to demonstrate that the large variation explained by courses and instructors is not just an artifact of the large number of categories in these attributes. In this test, we randomly scramble the course codes in the data. In this way, the number of course indicators remain the same, but scrambled course codes do not bear any economic meaning. We replicate the Shapley-Owen decomposition shown in column (1) of Table 3. If the large portion of explained variance were solely driven by the large number of indicators, even scrambled course codes should explain some variance. Instead, we find that they explain less than 1% of the total variation in the education-innovation gap.

much smaller share of the total variance, between 2 and 4 percent. Fields also explain a small share, between 5 and 12 percent. Specifications with course effects show that courses explain about a third of the overall variance. This indicates significant persistence in the content of a course over time.

In the remainder of the paper, we focus more in-depth on two of these factors: instructors and schools. Specifically, we explore how the gap relates to the characteristics of the person who teaches the course, and we study how it varies across schools with different characteristics and serving different populations of students.

4 Course Instructors and The Education-Innovation Gap

Instructors are considered one of the most important and costly inputs for the production of student learning in schools (De Vlieger, Jacob, and Stange, 2020). In line with this, our variance decomposition shows that a significant part of the overall variation in the education-innovation gap can be attributed to instructors. Motivated by these findings, we now investigate the role of instructors in shaping the education-innovation gap.

4.1 Event Study: The Education-Innovation Gap When Instructors Change

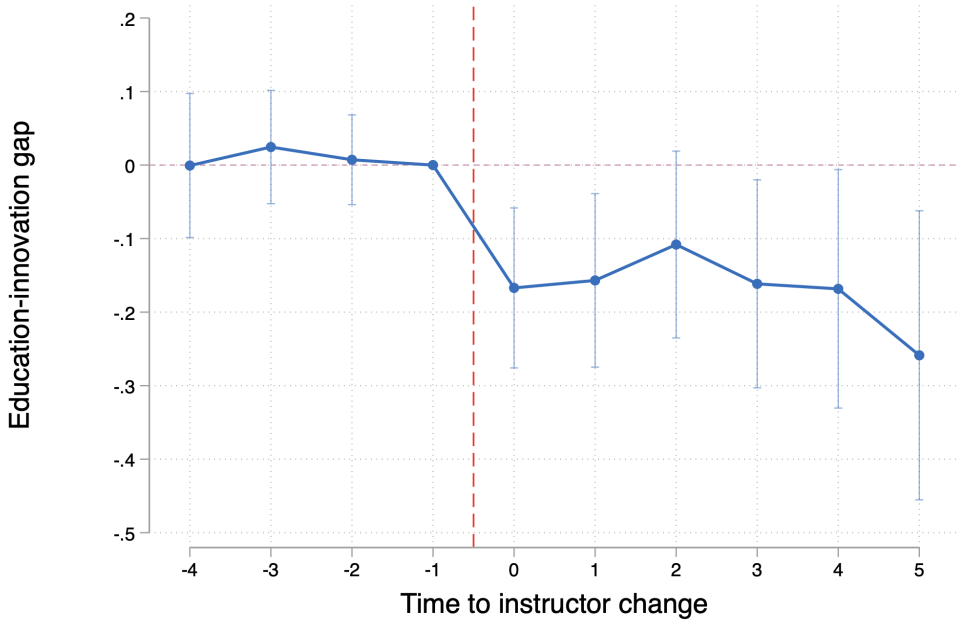
We first study how the education-innovation gap of a course varies when the course instructor changes. This allows us to measure the direct role of instructors in shaping course content. We estimate an event study of the gap in a ten-year window around the time of an instructor change:

$$\text{Gap}_{ct} = \sum_{k=-4}^{5} \delta_k \mathbb{1}(t - T_c = k) + \gamma_c + \eta_t + \varepsilon_{ct}, \tag{5}$$

where c and t denote a course and year, respectively. The subscript c denotes a specific course within each school (for example, Econ 110 at Yale University). Gap_{ct} measures the education-innovation gap of course c in t. T_c represents the first year in our sample in which the instructor of course c changes.[21] We set the indicator function to zero for all courses without an instructor change, which serves as the comparison group. We cluster standard errors at the course level. After normalizing δ_{-1} to be 0, the parameters δ_k capture the differences between the gap k years after an instructor change relative to the year preceding the change. To the extent that changes in instructors are unrelated to other unobservable determinants of course content, these estimates represent the causal effect of a new instructor on the education-innovation gap.

[21] In our data, some courses feature more than one instructor change over time. To better isolate the effect of each change, we restrict our attention to courses taught by only one instructor in each year and with at most three changes over the sample period. Our results are robust to this choice.

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



Notes: Estimates and confidence intervals of the parameters δ_k in equation (5), representing an event study of the education-innovation gap around an instructor change and controlling for course and year fixed effects. Observations are at the course-by-year level; we focus on courses taught by one instructor per year and with at most three episodes of instructor changes. Standard errors clustered at the course level.

Figure 3 shows that OLS estimates of δ_k are indistinguishable from zero and on a flat trend in the years leading to an instructor change. Following the change, however, the education-innovation gap immediately declines by about 0.2. To quantify the economic magnitude of these differences, we use the simulation results in Figure 2. The simulation results indicate that this decline is equivalent to updating 34 words in the median syllabus, or approximately 10 percent of its content. The decline is robust to the presence of plausible deviations from the standard parallel trends assumption of event studies (Rambachan and Roth, 2019) and to the possibility that instructor effects differ depending on the year in which they occurred (Sun and Abraham, 2021) (Appendix Figures AIX and AVIII).[22]

In Table 4 we re-estimate equation (5) for different subsamples of syllabi, pooling together years preceding and following an instructor change. The variable *After change* equals one in years fol-

[22]In Appendix Figure AIX we test the robustness of the statistical significance of δ_0 in equation (5), by implementing the test proposed by (Rambachan and Roth, 2019). Estimates of δ_0 remain distinguishable from zero even under plausible violations of the parallel trends assumption, which indicates that the measured decline in the gap is not due to differential trends.

lowing the change. After an instructor change, the gap declines for all fields and course levels by 0.134 on average (8 percent of a course's content, column 1, significant at 1 percent). The decline is largest for STEM and Social Science courses (-0.20 and -0.16, columns 4 and 5, respectively) and for graduate courses compared with undergraduate (-0.19, column 8).

These results show that instructors play an active role in determining the education-innovation gap. New instructors who take over a course significantly update its content, bringing it closer to the knowledge frontier; this is especially true for instructors of advanced courses. A flat trend prior to the change suggests that course updating is not a gradual process taking place over time. Rather, instructors who teach the same course for many years tend to leave content unchanged.

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

	Field					Course level		
	All	Business	Humanities	STEM	Soc. Sci.	Basic	Advanced	Grad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
After change	-0.134***	-0.107	-0.119*	-0.195***	-0.156***	-0.076	-0.140**	-0.189***
	(0.030)	(0.072)	(0.065)	(0.046)	(0.050)	(0.057)	(0.054)	(0.045)
N (Course * year)	392536	38444	108151	155039	99391	131995	117525	142936
# Courses	129605	12337	36601	46701	32868	45350	36754	47478
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field * year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates. Observations are at the course-by-year level; we focus on courses taught by one instructor per year and with at most three episodes of instructor changes. The dependent variable is the education-innovation gap. The variable *After change* is an indicator for years following an instructor change. All specifications control for course and field-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

4.2 The Education-Innovation Gap and Instructors' Research

Event study estimates suggest an active role for instructors in shaping the education-innovation gap. Instructors, though, differ among themselves in a number of dimensions, including their impact on student outcomes (Hoffman and Oreopoulos, 2009; Carrell and West, 2010; Braga et al., 2016; Feld et al., 2020). We now study whether they also differ in the extent to which they incorporate frontier knowledge in the courses they teach, and whether these differences can be explained by individual characteristics.

We mainly focus on the relationship between instructors' research activity and the education-innovation gap of the courses they teach. This relationship is not clear ex ante. Most instructors perform both research and teaching activities under time constraints. On the one hand, research

and teaching compete for the instructor's time, and the opportunity cost of keeping a course up to date might be higher for research-active faculty, compared with faculty who do not produce research and whose primary or sole job is to teach. This would lead research-active instructors to teach courses with a higher gap. On the other hand, research-active instructors are more familiar with the knowledge frontier and face a lower cost of keeping a course up to date. This would lead them to teach courses with a lower gap. We now test which of these two forces prevails.

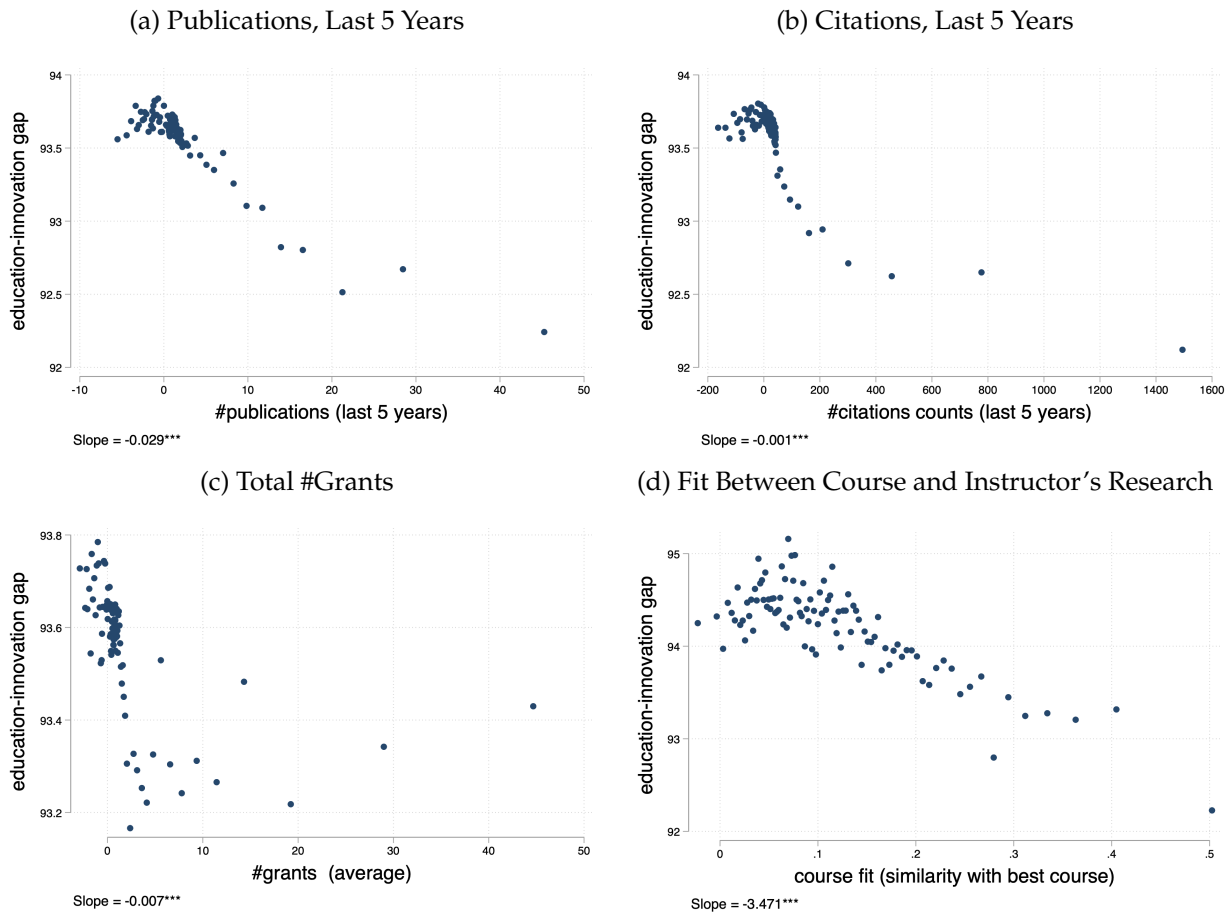
Research productivity and quality We study the relationship between the education-innovation gap and the research productivity and quality of the instructor, measured using the number of publications produced and citations received in the five years prior to the year of the course. Binned scatterplots, controlling for field-by-course level-by-year fixed effects, display negative and statistically significant relationships between the gap and various measures of instructor research productivity (Figure 4). A one-standard deviation (sd) increase in publications (approximately 6.7 additional publications in the previous five years) is associated with a 0.2 lower gap (an update of about 10 percent of the content of the median syllabus, panel (a)). A one-sd increase in citations (equal to 191 additional citations in the previous five years) is also associated with a 0.2 lower gap (panel (b)).

These patterns bear several interpretations. One is that more research-productive instructors are better informed about frontier knowledge, and thus better able to incorporate it into the course they teach. An alternative is that schools assign research-productive instructors to courses that are intended to cover frontier knowledge. To distinguish between these two hypotheses, we estimate the within-course relationship between the gap and instructors' research productivity, using the following equation:

$$\text{Gap}_{ct} = \beta q_{k(ct)t} + \gamma_c + \phi_{f(c)l(c)t} + \varepsilon_{ct} \tag{6}$$

where q_k is a measure of research productivity of instructor k (publications or citations), standardized to have mean zero and variance one within each field across all schools and years. The inclusion of course fixed effects γ_c implies that our estimates are obtained by comparing changes in the gap of the same course across different instructors. Field-by-course level-by-year fixed effects ϕ_{flt} account for any determinants of the gap specific to courses of a given level belonging to a given field and year. In this specification, the parameter β measures the difference in the education-innovation gap associated with a one-sd change in instructor publications or citations. To the extent that course-specific factors, which might drive the allocation of instructors across courses, do not change over time, β should then capture the impact of an instructor's productivity on the education-innovation

Figure 4: Instructors' Research Productivity, Funding, and Fit with the Course and the Education-Innovation Gap



Notes: Binned scatterplots of the gap (vertical axis) and measures of research productivity, quality, funding, and fit between the course topic and the research of the instructor. These measures are the number of publications in the last five years (panel a); the number of citations in the last five years (panel b); the total number of NSF and NIH grants ever received (panel c); and the fit between the instructor's research agenda and the course content, calculated as the cosine similarity between the instructor's publications and the syllabus of the course with the lowest gap among all courses on a given topic across schools in each year (panel d). All graphs control for field-by-course level-by-year effects. Slope coefficients are obtained from linear regressions of the gap on the corresponding variable, controlling for field-by-course level-by-year effects and clustering standard errors at the course level.

gap of the course they teach.

Estimates of β, shown in Table 5, indicate that the gap declines when the research productivity of the course's instructor increases. A one-sd increase in instructor publications over the previous five years is associated with a 0.04 decline in the gap (equivalent to updating 4 percent of a course's syllabus; Table 5, panel (a), column 1, significant at 1 percent). This relationship is particularly pronounced for Social Sciences, where the same increase is associated with a 0.07 decline in the gap

(5 percent of a course's syllabus, panel (a), column 5).

We showed before that the presence of frontier knowledge is stronger in graduate-level courses. The relationship between instructor research productivity and the education-innovation gap is also stronger for these courses. A one-sd increase in instructor publications is associated with a 0.05 decline in the gap for graduate-level courses (or 5 percent of a course's syllabus; Table 5, panel (b), column 1, significant at 1 percent). This relationship is again stronger in the Social Sciences (0.07 decline, column 5) and in STEM (0.06 decline, column 4).

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

Panel (a): all courses, publications	All (1)	Business (2)	Humanities (3)	STEM (4)	Soc. Sci. (5)
publications (sd)	-0.035***	-0.004	-0.024	-0.026	-0.065***
	(0.012)	(0.028)	(0.023)	(0.026)	(0.019)
N (Course * year)	581723	60940	156899	195266	150719
# Courses	153731	15155	43046	51844	39166
Panel (b): graduate-level courses, publications	(1)	(2)	(3)	(4)	(5)
publications (sd)	-0.048***	-0.023	-0.010	-0.055	-0.070***
	(0.017)	(0.033)	(0.047)	(0.036)	(0.024)
N (Course * year)	199735	32562	31329	59844	72798
# Courses	54663	8490	9021	17050	19252
Panel (c): all courses, citations	(1)	(2)	(3)	(4)	(5)
citations (sd)	-0.017	0.029	-0.004	-0.004	-0.060***
	(0.012)	(0.025)	(0.022)	(0.023)	(0.020)
N (Course * year)	581723	60940	156899	195266	150719
# Courses	153731	15155	43046	51844	39166
Panel (d): graduate-level courses, citations	(1)	(2)	(3)	(4)	(5)
citations (sd)	-0.037**	0.019	0.010	-0.062*	-0.067***
	(0.016)	(0.029)	(0.043)	(0.034)	(0.025)
N (Course * year)	199735	32562	31329	59844	72798
# Courses	54663	8490	9021	17050	19252
Course FE	Yes	Yes	Yes	Yes	Yes
Field * course level * year FE	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 counts of publications and citations, standardized within each of 69 fields. Panels (a) and (c) are estimated on all courses; panels (b) and (d) are estimated on graduate-level courses. All specifications control for course and field-by-course level-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

In panels (c) and (d) of Table 5 we also test whether instructors whose work receives more cita-

tions teach courses with a lower gap. A one-sd increase in citations is associated with a 0.02 lower gap, although imprecisely estimated (panel (c), column 1, p-value equal to 0.15). This relationship is stronger for graduate-level courses, where the same decline is associated with a 0.04 lower gap (or 3 percent of the median syllabus, panel (d), column 1). The relationship is also stronger for STEM (0.06 lower gap, column 4) and Social Sciences (0.07 lower gap, column 5).

Research funding To explore an alternative measure of research activity, in Table 6 we use data on the number of NSF and NIH grants received by each instructor. A binned scatterplot shows a negative relationship between the gap and the number of NSF and NIH grants (Figure 4, panel (c)). This relationship is confirmed by estimates that control for course and field-by-course level-by-year effects. A switch from an instructor who never received a grant to one with at least one grant is associated with a 0.06 reduction in the gap (column 1, significant at 5 percent). This estimate suggests that public investments in academic research can yield additional private and social returns in the form of more updated instruction.[23]

Fit with the course So far, the evidence supports the hypothesis that research-active instructors are better informed about the research frontier and thus can incorporate this knowledge into a course's content. If this is the case, we should expect the gap to be lower for courses whose topics are more related to the instructor's own research. For example, a labor economist should be better equipped to teach a course on labor economics than a course on industrial organization.

Table 6: The Education-Innovation Gap and Instructors' Research Resources: NSF/NIH Grants

	Field					Course level		
	All	Business	Human.	STEM	Soc. Sci.	Basic	Adv.	Grad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
At least one grant	-0.058**	-0.002	-0.127**	-0.025	-0.073*	-0.063	-0.050	-0.058
	(0.024)	(0.065)	(0.056)	(0.043)	(0.039)	(0.041)	(0.045)	(0.040)
N (Course * year)	581723	60940	156899	195266	150719	210121	171867	199735
# Courses	153731	15155	43046	51844	39166	55594	43474	54663
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field * course level * 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 has received at least one NSF or NIH grant. All specifications control for course and field-by-course level-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

We test this hypothesis by constructing a measure of "fit" between the course and the instruc-

[23] For a review of the role of grant funding as a tool to promote innovation, see Azoulay and Li (2020).

tor's research. This measure is defined as the cosine similarity between the instructor's research in the previous five years (captured using the text of their publications) and the most updated course on the same topic at the same level across *all* schools. For example, we examine the set of courses on Introductory Econometrics and test whether an instructor who does research in econometrics includes more frontier knowledge in the course than one who does research on macroeconomics.[24] Binned scatterplots, obtained controlling for field-by-course level-by-year fixed effects, indicate that a one-sd increase in fit (equal to 0.1) is associated with a 0.3 lower gap (Figure 4, panel (d)). Estimates of equation (6), obtained using the fit measure as the explanatory variable and controlling for field-by-course level-by-year fixed effects, indicate that a one-sd higher instructor-course fit is associated with a 0.09 lower gap, equivalent to a 5 percent update in the median syllabus's content (Table 7, significant at 5 percent). This relationship is particularly strong for Humanities, STEM, and Social Science courses (columns 4 and 5) and for courses at the advanced undergraduate level (column 7).

Table 7: The Education-Innovation Gap and the Fit Between Instructors' Research and Course Content

	Field					Course level		
	All	Business	Human.	STEM	Soc. Sci.	Basic	Adv.	Grad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fit w/top course (sd)	-0.088**	0.279**	-0.184	-0.110*	-0.115	-0.051	-0.149*	-0.054
	(0.044)	(0.113)	(0.234)	(0.059)	(0.074)	(0.093)	(0.086)	(0.063)
N (Course * year)	54591	3293	2270	35814	12626	16743	16224	21139
# Courses	17077	1040	781	11149	3923	5208	4833	6883
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field * course level * year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

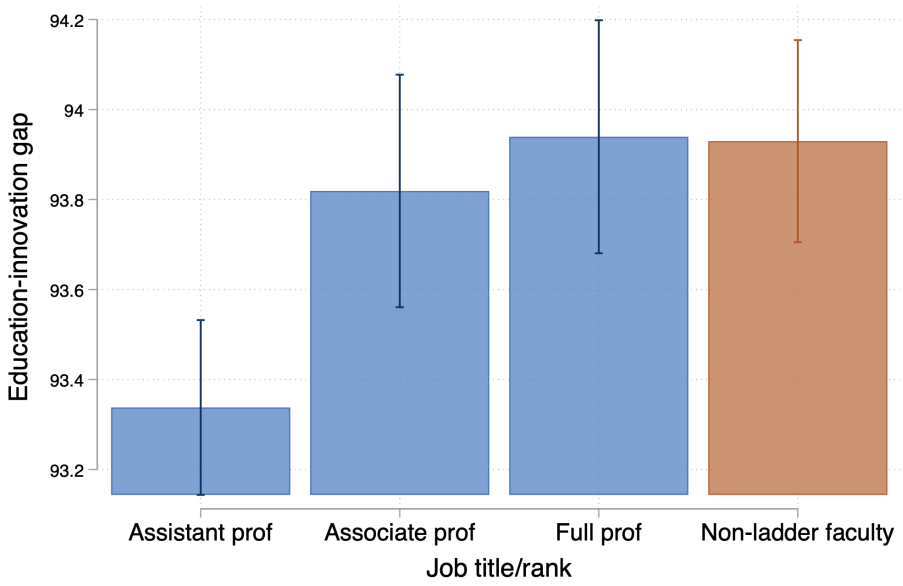
Note: OLS estimates; one observation is a course in a given year. The dependent variable is the education-innovation gap. The variable *Fit w/top course* is a measure of fit between the instructor's research and the content of the course, defined as the cosine similarity between the instructor's research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

Ladder vs non-ladder faculty These results presented so far carry implications for how the education-innovation gap may vary across faculty ranks and tracks (tenure-track vs. non-tenure-track), due to differences in their research activities. Ladder (i.e., tenure-track or tenured) faculty are gener-

[24] An attractive property of this measure is that it does not uniquely reflect the instructor's own syllabus; rather, it aims to capture the content of all courses on the same narrowly defined topic. To construct this measure, we obtained a unique identifier for courses in the same field or topic (e.g., Machine Learning) across schools. Appendix B describes the procedure used to construct these categories.

ally more focused on research compared with non-ladder faculty, whose primary job is to teach. In recent years, universities have started to increasingly rely on non-ladder faculty to meet a rapid rise in enrollment (Goolsbee and Syverson, 2019).[25] Ex ante, one could argue that—by virtue of being specialized in teaching—non-ladder faculty might be better at keeping educational content updated.

Figure 5: 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 and field-by-course level-by-year fixed effects. Estimates are obtained by pooling data for multiple years. Standard errors are clustered at the school level.

We compare the education-innovation gap across job titles, controlling for field-by-course level-by-year effects. We find that, among all faculty, tenure-track assistant professors have the lowest gap at 93.3. Associate professors have a slightly smaller gap than full at 93.8, but still significantly higher than assistant professors. Junior faculty on the tenure track thus appear to teach the courses with the most updated content.

Non-ladder faculty (such as adjunct professors, lecturers, professors in the practice, and visiting professors) also have a higher gap than assistant professors, at 93.9 (Figure 5). A 0.6 difference in the

[25]Colleges have monopsony power on tenure-track (but not non-ladder) faculty, as these earn substantially lower wages and have a much higher elasticity of labor supply. This implies that, when enrollment increases, schools can avoid increasing wages for tenure-track faculty by hiring more non-ladder faculty (Goolsbee and Syverson, 2019).

gap between assistant professors and non-ladder faculty is equivalent to 19 percent of a syllabus's content. One possible explanation for this finding is that assistant professors are more recently trained and therefore better updated about frontier knowledge. Furthermore, they often have the strongest incentives to be active in research. Notably, at 93.9 the gap is indistinguishable between full (tenured) professors and adjuncts.

Taken together, these analyses provide additional evidence that instructors play the most important role in the creation of course content and the dissemination of frontier knowledge, above and beyond the roles played by schools and majors (which so far have received most of the attention from the existing literature). Our findings also suggest that research and teaching are complementary activities: 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 agendas. Proper deployment of faculty across courses can have important impacts on the content of education, and investments in faculty research (both public, in the form of government grants, and institution-specific) can generate additional returns in the form of more updated instruction.

5 The Education-Innovation Gap Across Schools

So far, we have explored differences in the education-innovation gap across courses within the same school. Yet, schools differ substantially in terms of resources, organization, the overall education production function, and the composition of the student body. This can translate into differences in educational content. For example, if research-active faculty teach courses with a lower gap, schools with a stronger focus on research could offer courses with a lower gap. Furthermore, since students from different socio-economic backgrounds tend to enroll in schools with different attributes (Chetty et al., 2020), access to low-gap content may also vary across more and less disadvantaged students.

In this section, we study the relationship between a set of characteristics of the schools and the students they serve and the education-innovation gap of the course these schools offer. We focus on three sets of school attributes: (i) institutional characteristics, including sector (public or private), size, and overall spending and spending on instruction and research; (ii) faculty composition; and (iii) composition of the student body with respect to test scores and socio-economic background, captured by the mean SAT score of admitted students and the median parental income of enrolled students, respectively. Quantifying the differences in the education-innovation gap across schools

with different characteristics can be helpful to better understand the offering of frontier knowledge across the US higher education landscape.

We first examine the relationship of the gap with each of these variables one-by-one, by estimating the following equation:

$$\text{Gap}_i = \beta X_{s(i)} + \psi_{f(i)l(i)t(i)} + \varepsilon_i \tag{7}$$

where Gap_i measures the education-innovation gap of syllabus i, taught in school $s(i)$ and year $t(i)$; the variable X_s is one of the school attributes in school s mentioned above; and field-by-course level-by-year fixed effects ψ_{flt} control for systematic differences in the gap, common to all syllabi in the same field (f) and course level (l), that vary over time (t). We cluster standard errors at the institution level. In this equation, the parameter β captures the pairwise relationship between the gap and each school characteristic.

While pairwise relationships are useful, several of the characteristics we are interested in are correlated with each other. For example, more selective schools on average have higher expenditures and enroll students with higher parental incomes. To account for these correlations, we also estimate a version of equation (7) in which we include all school characteristics on the right-hand side. Both sets of results are displayed in Figure 6: Hollow markers denote pairwise estimates and solid markers denote multivariate estimates. We now describe them in detail.

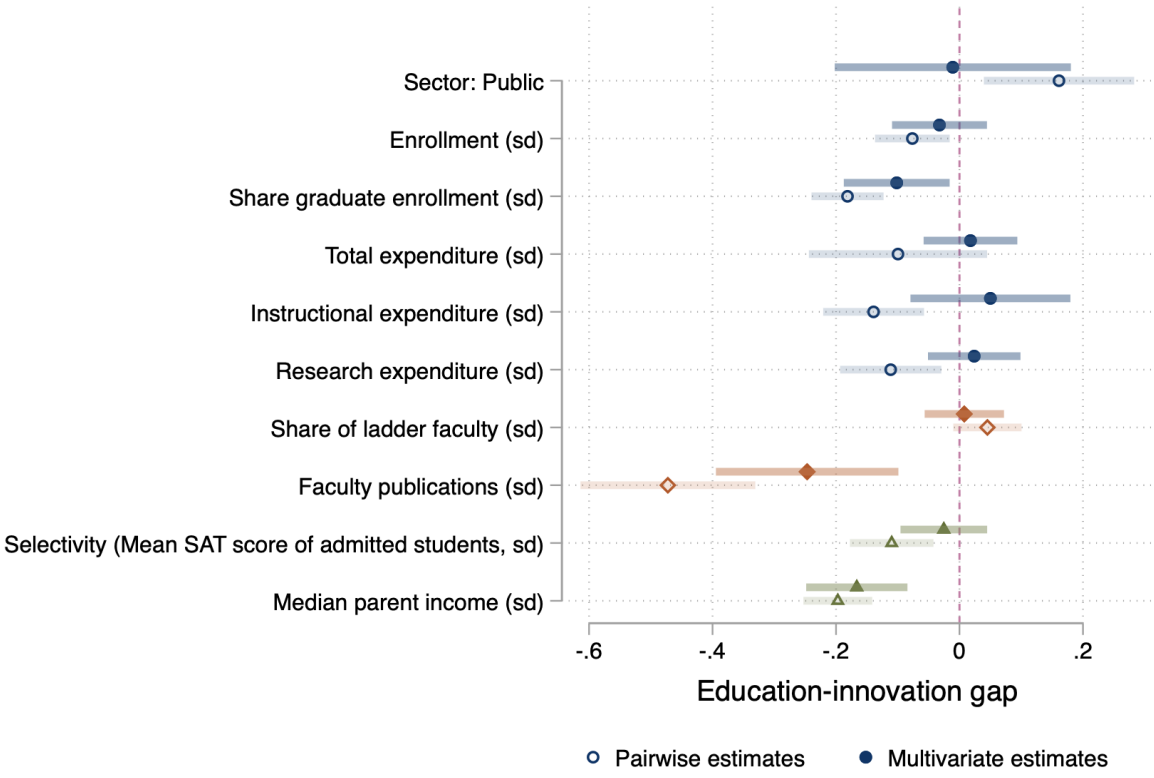
5.1 Institutional Characteristics: Sector, Size, and Spending

Sector On average, public schools have a 0.2 larger gap compared with private schools, akin to a 10 difference in the content of the median syllabus (Figure 6, hollow series). This difference, however, becomes indistinguishable from zero when we include all other characteristics in the equation (Figure 6, solid series). This implies that the observed penalty for public schools can be attributed to differences in size, resources, faculty, and student composition.

Size US schools vary considerably in size, with 12-month enrollment equal to 2,040 on the 10th percentile (Hamilton College, NY) and to more than 34,000 on the 90th percentile (Georgia Institute of Technology, GA). Enrollment size can impact the organization of the instruction, with consequences for the education-innovation gap. The data confirm that larger schools teach courses with a slightly lower gap. A one-sd larger enrollment is associated with a 0.08 lower gap, equivalent to a 6 percent update in the median syllabus. This relationship, though, can be explained by other attributes: It becomes smaller and insignificant when we control for other school characteristics.

School Spending Schools that spend more on instruction and research teach courses with a lower gap. A one-sd increase in instructional spending (approximately $65,000) is associated with a 0.14 lower gap, or an 8 percent change in the syllabus; a one-sd increase in research spending (approximately $18,000) is associated with a 0.11 lower gap, or a 7 percent change in the syllabus. These relationships, though, become small and indistinguishable from zero when we control for all other school characteristics.

Figure 6: The Education-Innovation Gap and School Characteristics



Notes: Hollow markers denote OLS point estimates and 95-percent confidence intervals of β in equation (7), 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 estimate is obtained from a separate regression. Full marks denote OLS estimates of a multivariate regression of the gap on all displayed variables, as well as field-by-course level-by-year fixed effects. All variables on the vertical axis (except for *Public*) are standardized to have mean 0 and variance 1; each coefficient represents the change in the gap associated with a one-sd change in each variable. Enrollment, expenditure, and share minority refer to the year 2018 and are taken from IPEDS. Estimates are obtained by pooling syllabi data for the years 1998 to 2018. Standard errors are clustered at the school level.

5.2 Faculty Composition and Research

The results of the previous section highlight an important role for instructors in shaping educational content and show that research-active instructors tend to teach courses with a lower gap. We now test whether these patterns are confirmed across schools. We estimate the correlation between the gap and (a) the share of ladder faculty, and (b) the average number of faculty publications.

Estimates of these relationships confirm the relationship between faculty's research productivity and the education-innovation gap. A one-sd increase in faculty publications is associated with a 0.5 lower gap, which corresponds to an 18 percent difference in the content of the median syllabus. This estimate becomes smaller at -0.3 but stays statistically significant even when we control for other school characteristics. We instead do not find evidence of a relationship between the share of ladder faculty and the gap: Pairwise and multivariate correlations are small and indistinguishable from zero. Taken together, these findings are in line with the hypothesis that the research productivity of instructors is an important determinant of the education-innovation gap.

Share of graduate enrollment Schools that enroll more research-active faculty typically have a stronger focus on research. This focus is often also reflected in larger graduate programs, such as Master's and PhD. Our data confirm that schools with larger graduate programs offer courses with a lower gap. A one-sd increase in graduate enrollment is associated with a 0.2 lower gap (or a 10 percent change in the median syllabus). This relationship becomes smaller at -0.1, but stays statistically significant when we control for other school characteristics.[26]

5.3 Student Body Characteristics

Socioeconomic background: Parental Income Schools with different characteristics serve different populations of students (Chetty et al., 2020). For example, Ivy Plus and Elite schools are disproportionately more likely to enroll students from wealthier families. Cross-school differences could therefore translate into significant disparities in access to up-to-date knowledge among students with different backgrounds.

To test whether schools serving students from more advantaged backgrounds offer courses with a lower gap, we re-estimate equation (7) using the median parental income as the explanatory variable. This variable is constructed using tax returns for the years 1996 to 2004 (Chetty et al., 2020). The pairwise correlation, shown as a hollow triangle in Figure 6, indicates that courses in schools serving more economically disadvantaged students on average have a higher gap. Specifically, a

[26] This relationship is not simply driven by the fact that schools with more graduate students offer more graduate-level courses. A one-sd increase in graduate enrollment is associated with a 0.15 lower gap for undergraduate syllabi.

one-sd higher median parental income is associated with a 0.2 lower gap, which corresponds to a 10 percent update in the median syllabus.

Students' SAT scores In principle, part of the differences in the gap across schools could be due to a "vertical differentiation" of educational content across schools based on students' ability and preparedness. If better-prepared students can absorb frontier knowledge more easily, cross-school differences in the gap might reflect schools' efforts to provide students with appropriate educational content.

To explore this possibility, we test whether schools that admit students with higher SAT or ACT scores (and therefore tend to be more selective) teach courses with lower gaps. We use each school's average SAT score of all students admitted in 2018, standardized to have mean zero and variance one. The mean 2018 SAT was equal to 1,517 for Yale University, which only admitted 6.3 percent of its applicants, and 1,027 for Southern Connecticut State University, which admitted 66.3 percent of all applicants. We assign non-selective schools an average SAT of zero.

Pairwise estimates of the relationship between the gap and student preparedness, obtained estimating equation (7) with the average SAT score of admitted students as the only right-hand side variable, indicate that schools enrolling better-prepared students offer courses with a lower gap. A one-sd increase in the average SAT score is associated with a 0.11 lower gap, corresponding to a 7 percent difference in the content of the median syllabus (Figure 6, hollow marks). However, this relationship becomes indistinguishable from zero when we control for the other school characteristics (Figure 6, full marks).

Importantly, the relationship between median parental income and the education-innovation gap highlighted above remains negative and significant when we control for all other school characteristics, which include SAT scores (Figure 6, full triangle). This finding is in line with existing evidence on disparities in access to selective schools among more and less advantaged students. Furthermore, they document a new dimension of inequality: access to educational content that is close to the research frontier. Importantly, this inequality cannot be explained by differences in student test scores.

Given this result, one may be left wondering why students from more disadvantaged backgrounds attend schools offering less cutting-edge content. Answering this question is outside the scope of this paper. A possible explanation is the lack of information. Disadvantaged students may be less informed on the type of education provided by each school, which could make them systematically less likely to choose schools with a lower gap (Hoxby et al., 2013; Hoxby and Turner,

2015).[27]

6 The Education-Innovation Gap and Students' Outcomes

Our findings so far reveal significant differences in access to up-to-date knowledge, both within and across schools. To the extent that access to frontier knowledge promotes the creation of new knowledge and human capital, schools that offer courses closer to the knowledge frontiers should have students who are better able to produce innovation and, generally, more successful in the labor market.

In this section, we explore the relationship between the gap and two sets of student outcomes: first, measures of students' innovation activities, such as a school's share of undergraduate students who complete a doctoral degree and the number of patents produced by students; and second, education and labor market outcomes, including graduation rates, earnings, and intergenerational income mobility.

Ideally, one would observe the outcomes of each student and the content of each course they enrolled in. However, in our data, outcomes are measured at the aggregate level—either at the school level or at the school-by-cohort level (with the exception of the share of students who attend graduate school, available separately by macro-field). To match this feature of the data, we follow the school value-added literature (Deming, 2014) and estimate the school component of the gap using the following model:

$$\text{Gap}_i = \theta_{s(i)} + \phi_{f(i)l(i)t(i)} + \varepsilon_i. \tag{8}$$

In this equation, θ_s captures the school component of the education-innovation gap for school s, accounting for flexible time trends that are specific to the level l and 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 AXII shows the distribution of the estimated school-level component, denoted as $\hat{\theta}_s$. Its standard deviation is 0.85, corresponding to a 24 percent change in the average syllabus.

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

$$Y_{st} = \delta\hat{\theta}_s + X_{st}\Gamma + \tau_t + \varepsilon_{st} \tag{9}$$

[27] A randomized controlled trial that provided students with individualized information about colleges' net prices, resources, curricula, students, and outcomes raised students' applications to, admissions at, enrollment, and progress at selective colleges (Hoxby et al., 2013). The intervention also changed students' knowledge about each school and decision-making (Hoxby and Turner, 2015).

where Y_{st} is the outcome for students who graduated from school s in year t; $\hat{\theta}_s$ is the school-level component of the gap (estimated from equation (8) and standardized to have mean zero and variance one); X_{st} is a vector of school observables; and τ_t are year fixed effects. We report bootstrapped standard errors, clustered at the level of the school, to account for the fact that $\hat{\theta}_s$ is an estimated quantity.

We want to stress that the parameter δ does not necessarily capture the causal effect of the gap on outcomes. Other than by this causal effect, a relationship between $\hat{\theta}_s$ and student outcomes could be driven by differences across schools in their organization and educational quality; in the ability of instructors; and in the characteristics of students who attend schools with different gaps. In an effort to control for as many observable differences as possible, we present both unconditional correlations and correlations obtained controlling for a rich set of school observables. 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 (captured by an indicator for institutions who admit all of their applicants, the median SAT and ACT scores of admitted students in 2006, and 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 median parental income.

6.1 Innovation Measures

Invention We begin by studying whether students at schools that offer courses with a lower gap produce more inventions later in their lives, in the form of patents. We do so by using the total number of patents received after graduation by students at each school as the dependent variable in equation (9). Unconditionally, a one-sd lower gap is associated with 24 additional patents at a given school, or 18 percent compared with an average of 130 patents per school (Table 8, panel (a), column 1, significant at 10 percent). The relationship remains largely robust when we control for school observables (Table 8, column 1, panel (b)).

Obtaining a doctoral degree Next, we study the relationship between the gap and the share of students who later obtain a doctoral degree. We construct this variable using data from the NSF

Survey of Doctorate Recipients (SDR), separately for five macro-fields: STEM, Health, Business, Social Science, and Humanities. To match 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 we modify equation (9) so that one observation in our data is a school and by macro-field in a year. The quantity θ_s is also estimated separately for each macro field. In column 2 of Table 8 (panel (a)), we pool data across all macro-fields. The unconditional correlation between the gap and the share of students who obtain a doctoral degree is negative and statistically significant: A one-sd lower gap is associated with a 0.37 percentage point higher share, or 14 percent compared with an average of 2.73 percent. The correlation is particularly strong for Social Science (-0.0108, column 6) and Health (-0.0077, column 4). These correlations remain remarkably robust when we control for school characteristics (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 graduation rates, an outcome that immediately precedes entry into the labor market. Graduation is in part also a function of choices made by the students, which could be impacted by the content of the courses they took.

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-sd decline in the gap is associated with a 5 percentage point higher graduation rate. Compared with an average of 59 percent, this corresponds to an 8 percent increase in graduation rates.

The estimate of δ declines as we control for observable school characteristics, suggesting that part of this correlation can be explained by other differences across schools. However, it remains negative and significant at -0.008, indicating that a one-sd reduction in the gap is associated with a 1.3 percent increase in graduation rates (panel (b), column 1, significant at 5 percent).

Students' income and intergenerational mobility We next examine the relationship between the education-innovation gap and students' economic success after they leave college. In columns 2-8 of Table 9 we estimate the relationship between the gap and various income statistics.

Column 2 shows estimates of the correlation between the gap and the natural logarithm of median student earnings 10 years after graduation, from the College Scorecard. Controlling for the full set of observables, we show that a one-sd lower gap is associated with 0.7 percent higher earnings (column 2, panel (b), significant at 5 percent). The College Scorecard also reports mean earnings, separately for all students and for students with parental incomes in the bottom tercile of

the distribution. Overall, a one-sd lower gap is associated with an increase in mean earnings of 0.6 percent (column 3). For students with parental income in the bottom tercile, it is slightly larger at 1 percent (column 4, panel (b), significant at 5 percent). Exposure to frontier knowledge thus bears the strongest correlation with the earnings of students from less advantaged backgrounds.

Information on mean student earnings at the school level is also reported by Chetty et al. (2020) for the graduating cohorts of 2002-04. Unconditional estimates (which omit year effects due to the cross-sectional structure of the data) indicate that a one-sd decline in the gap is associated with a 7 percent increase in students' mean earnings (panel (a), column 5, significant at 1 percent). This estimate is equal to 1.2 percent when we control for institutional characteristics (panel (b), column

Table 8: The Education-Innovation Gap and Innovation Measures: Share of Undergraduate Students Who Obtain a Doctoral Degree and Total Number of Patents

	Nr	Share of students who obtain a doctoral degree, by field					
	Patents	All	STEM	Health	Business	Soc. Sci.	Humanities
Panel (a): no controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gap (sd)	-23.7275*	-0.0037**	-0.0003	-0.0077**	-0.0003	-0.0108***	0.0049
	(12.6023)	(0.0016)	(0.0025)	(0.0035)	(0.0005)	(0.0041)	(0.0055)
Mean dep. var.	130.4513	0.0270	0.0461	0.0255	0.0022	0.0340	0.0232
N	1695	62143	13887	8810	12006	13819	13621
Panel (b): w/ controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gap (sd)	-18.0227*	-0.0038**	0.0008	-0.0067**	-0.0004	-0.0084**	0.0058
	(9.4977)	(0.0017)	(0.0019)	(0.0030)	(0.0005)	(0.0038)	(0.0055)
Mean dep. var.	131.8495	0.0273	0.0466	0.0261	0.0022	0.0346	0.0231
N	1595	45770	10225	6453	8862	10188	10042

Note: OLS estimates of the coefficient δ in equation (9). In column 1, *Gap (sd)* is estimated at the school level pooling data from all fields. In columns 2-7, the variable *Gap (sd)* is a school-by-macro field-level education-innovation gap (estimated as $\theta_{s(i)}$ in equation (8), separately for each macro-field), standardized to have mean zero and variance one. In column 1, the dependent variable is the total number of patents filed by students at each school, from Chetty et al. (2020), in columns 2-7, it is the share of undergraduate students at each institution-field who eventually complete a doctoral degree (from the NSF Survey of Doctorate Recipients, year 2000). All columns in panel (b) control for sector (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 2-7 control for year effects. Column 2 also controls for macro-field fixed effects. Bootstrapped standard errors in parentheses are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

5, significant at 5 percent). Notably, this correlation is larger than the one we obtain when we use earnings data from the College Scorecard. This difference is likely due to the age at which earnings are measured, equal to 24-30 in the College Scorecard (6-10 years after enrolling in college) and 32 in the Chetty et al. (2020) data. The gap is thus more strongly correlated with later-life earnings.

In columns 6 through 8 of Table 9 we investigate the relationship between the gap and the probability that students' earnings reach the top echelons of the distribution. Estimates with the full set of controls indicate that a one-sd decline in the gap is associated with a 0.8 percentage point increase in the probability of reaching the top quintile of the income distribution (2.1 percent, panel (b), column 6, significant at 1 percent), a 0.5 percentage point increase in the probability of reaching the top 10 percent (2.3 percent, column 7, significant at 5 percent), and a 0.3 percentage point increase in the probability of reaching the top 5 percent (2.5 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 earnings.

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 -0.0282, indicating that a one-sd lower gap is associated with a 2.8 percentage point increase in intergenerational mobility (9.4 percent, panel (a), column 9, significant at 1 percent). This correlation becomes smaller at -0.005 when we control for school observables (column 9, panel (b), with a p-value equal to 0.11).

Summary Our findings indicate that students from schools that offer courses with a lower education-innovation gap produce more innovation and have better academic and economic outcomes, even accounting for a wide range of observable characteristics. This finding supports several explanations. For example, it could be the case that students with a higher ability or a higher propensity to innovate self-select into schools with lower gaps. Alternatively, exposure to frontier knowledge in higher education could be beneficial for students, helping them innovate and thrive in the labor market.[28] While a formal test of the causal link between the education-innovation gap and student outcomes is beyond the scope of this paper, both of these explanations indicate that schools could benefit from expanding the coverage of frontier knowledge in the courses they offer, either by attracting better students or by improving students' outcomes.

[28]This explanation is also in line with the idea that access to frontier knowledge spurs the creation of new knowledge (Moser and Voena, 2012; Williams, 2013; Galasso and Schankerman, 2015; Iaria et al., 2018).

Table 9: The Education-Innovation Gap and Student Outcomes

		Earnings (College Scorecard)			Earnings (Chetty et al., 2020)				
	Grad rate	Median	Mean		Mean	P(earnings in top...)			P(top20%
			Overall	$P_y \leq 33$ pctile		top 20%	top 10%	top 5%	|$P_y$bottom 20%)
Panel (a): no controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap (sd)	-0.0500***	-0.0502***	-0.0583***	-0.0457***	-0.0657***	-0.0311***	-0.0251***	-0.0178***	-0.0282***
	(0.0075)	(0.0100)	(0.0110)	(0.0093)	(0.0125)	(0.0056)	(0.0052)	(0.0036)	(0.0053)
Mean dep. var.	0.5784					0.3737	0.2119	0.1169	0.2987
N	14620	3473	3265	3473	730	730	730	730	730
# schools	760	746	720	746					
Panel (b): w/ controls	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Gap (sd)	-0.0077**	-0.0069*	-0.0058	-0.0101**	-0.0119**	-0.0080***	-0.0052**	-0.0030*	-0.0050
	(0.0031)	(0.0041)	(0.0049)	(0.0043)	(0.0050)	(0.0030)	(0.0024)	(0.0017)	(0.0033)
Mean dep. var.	0.5873	10.8428	10.7768	10.7225		0.3753	0.2138	0.1185	0.2999
N	10883	1835	1690	1835	688	688	688	688	688
# schools	732	707	681	707					

Note: OLS estimates of the coefficient δ in equation (9). The variable *Gap (sd)* is the school-level education-innovation gap (estimated as $\theta_{s(i)}$ in equation (8)), standardized to have mean zero and variance one. The dependent variables are graduation rates (from IPEDS, years 1998-2018, column 1); the log of median student earnings 10 years after graduation, from the College Scorecard (column 2); the log of mean earnings 10 years after graduation, also from the College Scorecard, for all students (column 3) and for students with parental income in the bottom tercile (column 4); the log of mean earnings for students who graduated between 2002 and 2004 (from Chetty et al. (2020), column 5); the probability that students have earnings in the top 20, 10, and 5 percent of the national age-specific income distribution (from Chetty et al. (2020), 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 sector (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. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

7 Alternative Measures of Course Content

The richness of the syllabi data allows us to explore alternative measures to describe a course's content. In this section, we propose a few of them and examine the robustness of our results to the use of these alternative measures.

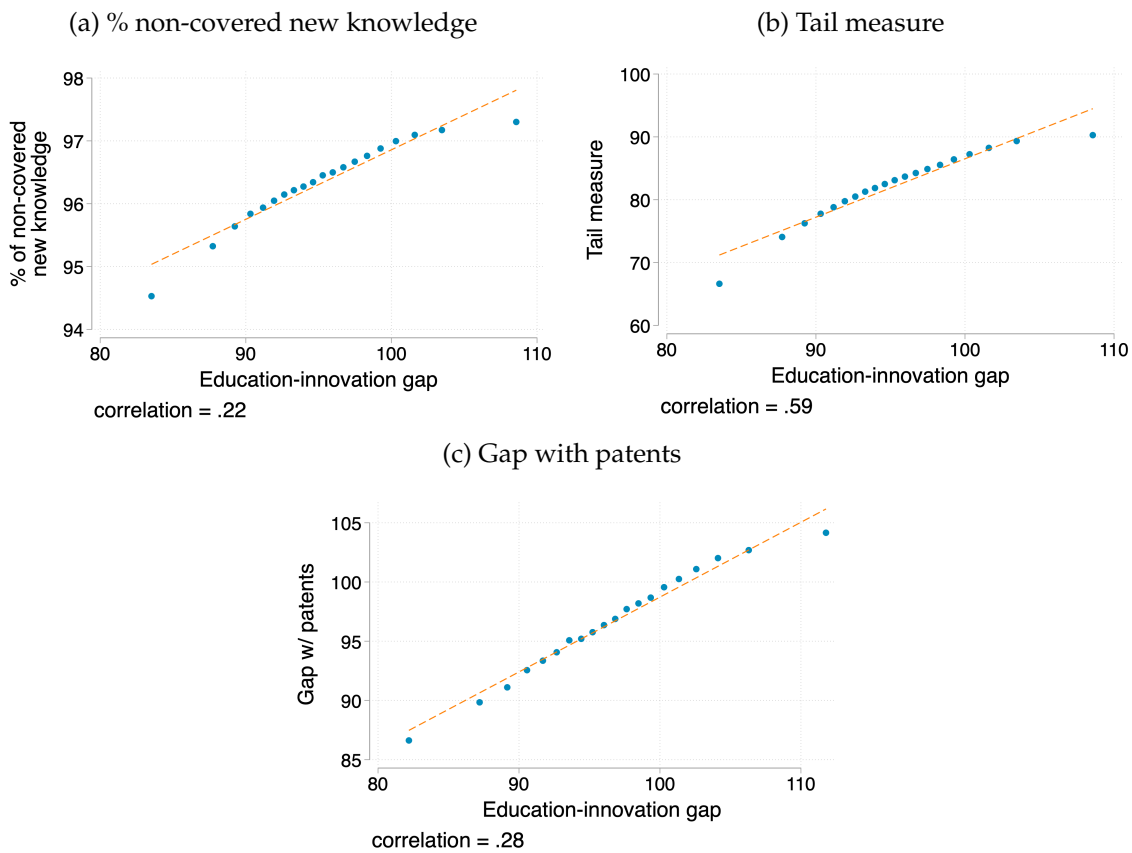
Probing the robustness of our results to these alternative measures is important because, in spite of its many desirable properties, our measure of the education-innovation gap has some limitations. For example, it captures the average distance of content from the knowledge frontier, thus penalizing courses that also include old content. This implies that, among two courses that cover the exact same new content, the one that *also* covers older knowledge will have a higher gap. Similarly, among courses with the same gap, it is unable to identify courses with extremely novel content. Lastly, the gap relies on academic publications to capture the knowledge frontier. In some fields, such as STEM, frontier knowledge could also be disclosed in other forms, such as patents for new technologies.

7.1 Presence of old vs new knowledge

The education-innovation gap measures the presence of new content relative to old content. Consider two syllabi that cover the *same amount* of frontier research; the first syllabus only contains this new content, while the second one also contains some old content. Our measure would assign a larger gap to the second syllabus compared to the first due to the presence of old content, even though both do an equal job in covering frontier knowledge.

To address this limitation, we construct an alternative metric: the share of new knowledge covered by a course, defined as the ratio between the number of "new words" in each syllabus and the number of all new words. New words are 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 $t-13$. Intuitively, this measure captures the portion of all new knowledge covered by the course, regardless of the presence of old knowledge. For clarity, we show our results using one minus the share of covered new knowledge, which we refer to as the *share of non-covered new knowledge*. This allows us to work with a metric that, like the education-innovation gap, is larger when the content of a course is more distant from frontier knowledge. The correlation between the share of non-covered new knowledge and the education-innovation gap is 0.22 (Figure 7, panel (a)), and our main results hold if we use this alternative metric to capture the novelty of a syllabus's content (see panel (a) of Figure AX for

Figure 7: The Education-Innovation Gap and Alternative Measures of Novelty: Binned Scatterplots



Notes: Binned scatterplots of the education-innovation gap and three alternative measures of novelty of each syllabus, as defined in the text: the percentage of non-covered new knowledge (panel (a)); a "tail measure," (panel (b)); and the education-innovation gap calculated using the text of all patents as a benchmark for frontier knowledge (panel (c)).

the correlation with school-level characteristics, panel (a) of Figure AXI for the correlation with instructors' research productivity, and panels (a) and (b) of Table AII for the relationship with student outcomes).

7.2 Right tail of academic novelty

The 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 random subsets of 20 percent of the syllabus's words, and calculate the corresponding education-

innovation gap. The gaps of these 100 sub-syllabi form a distribution; we use the 5th percentile of this distribution for each syllabus as a tail measure of new content.[29] 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 0.59 (Figure 7, panel (b)). All our results hold when we use the tail measure as a metric for syllabus novelty (see panel (b) of Figure AX for the correlation with school-level characteristics, panel (b) of Figure AXI for the correlation with instructors' research productivity, and panels (c) and (d) of Table AII for the relationship with student outcomes).

7.3 Gap with patents

The education-innovation gap is defined using new academic publications as the frontier of knowledge. For STEM fields, knowledge advancements are also documented in the form of patents. To incorporate this information in our analysis, we construct a version of the education-innovation gap for STEM courses that uses patents in lieu of academic publications. This measure is positively correlated with the standard education-innovation gap (Figure 7, panel (c)), and our main results hold when we use the patent-based gap (see panel (c) of Figure AX for the correlation with school-level characteristics, panel (c) of Figure AXI for the correlation with instructors' research productivity, and panels (e) and (f) of Table AII for the relationship with student outcomes).

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

8 Conclusion

This paper uses the text of HE course syllabi to quantify the distance between the content of each course and frontier knowledge. Our approach centers around a new measure, the "education-innovation gap," defined as the textual similarity between course syllabi and knowledge from older vintages, relative to newer ones. We construct this measure by applying NLP techniques to the full text of 1.7 million syllabi and 20 million academic publications. Our empirical approach combines a large-scale novel data source with textual analysis to shed new light on some key aspects of higher education.

Using our measure, we document a set of new findings about the dissemination of frontier knowledge across HE programs. Across and within schools, significant differences exist in the

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

extent to which frontier knowledge is taught to students. Instructors play the largest role in shaping the content of the courses they teach. Courses taught by more research-active instructors have lower gaps. Access to updated knowledge is highly unequal across students from different backgrounds: Schools that enroll more socio-economically advantaged students offer courses with a lower gap. The education-innovation gap is strongly correlated with students' innovation and labor-market outcomes. In schools offering courses with lower gaps, students are more likely to graduate, earn a PhD, and produce patents. They also earn more once they enter the labor market. Taken together, our findings indicate that the education-innovation gap can be an important metric for quantifying how frontier knowledge is produced and disseminated, and they could help shed new light on the way in which schools and instructors impact students' lives.

For future research, a careful analysis of the causal impacts of a low-gap education on students' later life outcomes represents an important and fruitful avenue. The use of novel alternative data, such as the text of various documents, could open the opportunity for researchers to investigate questions related to higher education which would otherwise be difficult to study.

References

Acemoglu, Daron, and David Autor, 2011, Skills, tasks and technologies: Implications for employment and earnings, in *Handbook of labor economics*, volume 4, 1043–1171 (Elsevier).

Akcigit, Ufuk, Jeremy G Pearce, and Marta Prato, 2020, Tapping into talent: Coupling education and innovation policies for economic growth, *NBER Working Paper* .

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

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

Angrist, Joshua, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu, 2017, Economic research evolves: Fields and styles, *American Economic Review* 107, 293–97.

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

Arnold, Ivo JM, 2008, Course level and the relationship between research productivity and teaching effectiveness, *Journal of Economic Education* 39, 307–321.

Azoulay, Pierre, and Danielle Li, 2020, Scientific grant funding, in *Innovation and Public Policy* (University of Chicago Press).

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

Becker, William E, and Peter E Kennedy, 2005, Does teaching enhance research in economics?, *American Economic Review* 95, 172–176.

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, *Quarterly Journal of Economics* 134, 647–713.

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, and Petra Moser, 2021, Effects of copyrights on science: Evidence from the wwii book republication program, *American Economic Journal: Microeconomics* 13, 218–60.

Bloom, Nicholas, Tarek Alexander Hassan, Aakash Kalyani, Josh Lerner, and Ahmed Tahoun, 2021, The diffusion of disruptive technologies .

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.

Bullen, Elizabeth, and Simon Robb, 2004, *Innovation and Tradition: the arts, humanities, and the knowledge economy*, volume 21 (Peter Lang).

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, 2020, Income segregation and intergenerational mobility across colleges in the United States, *Quarterly Journal of Economics* 135, 1567–1633.

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

Dale, Stacy B, and Alan B Krueger, 2014, Estimating the effects of college characteristics over the career using administrative earnings data, *Journal of Human Resources* 49, 323–358.

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

Deming, David J, and Kadeem Noray, 2020a, Earnings dynamics, changing job skills, and stem careers, *Quarterly Journal of Economics* 135, 1965–2005.

Deming, David J, and Kadeem L Noray, 2020b, Earnings dynamics, changing job skills, and stem careers, *Quarterly Journal of Economics* .

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.

Galasso, Alberto, and Mark Schankerman, 2015, Patents and cumulative innovation: Causal evidence from the courts, *The Quarterly Journal of Economics* 130, 317–369.

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

Goolsbee, Austan, and Chad Syverson, 2019, Monopsony power in higher education: A tale of two tracks, *NBER Working Paper* .

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.

Hattie, John, and Herbert W Marsh, 1996, The relationship between research and teaching: A meta-analysis, *Review of Educational Research* 66, 507–542.

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, *NBER Working Paper* .

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

Hoxby, Caroline, Sarah Turner, et al., 2013, Expanding college opportunities for high-achieving, low income students, *Stanford Institute for Economic Policy Research Discussion Paper* 12, 7.

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, and Sarah Turner, 2015, What high-achieving low-income students know about college, *American Economic Review* 105, 514–517.

Huettner, Frank, Marco Sunder, et al., 2012, Rego: Stata module for decomposing goodness of fit according to owen and shapley values, in *United Kingdom Stata Users' Group Meetings 2012*, number 17, Stata Users Group.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger, 2018, Frontier knowledge and scientific production: evidence from the collapse of international science, *The Quarterly Journal of Economics* 133, 927–991.

Israeli, Osnat, 2007, A shapley-based decomposition of the r-square of a linear regression, *The Journal of Economic Inequality* 5, 199–212.

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, 2021, Measuring technological innovation over the long run, *American Economic Review: Insights* 3, 303–20.

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* .

Moser, Petra, and Alessandra Voena, 2012, Compulsory licensing: Evidence from the trading with the enemy act, *American Economic Review* 102, 396–427.

Rambachan, Ashesh, and Jonathan Roth, 2019, An honest approach to parallel trends, *Unpublished manuscript, Harvard University* .

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* .

Sun, Liyang, and Sarah Abraham, 2021, Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics* 225, 175–199.

Tartari, Valentina, and Scott Stern, 2021, More than an ivory tower: The impact of research institutions on the quantity and quality of entrepreneurship, *NBER Working Paper* .

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.

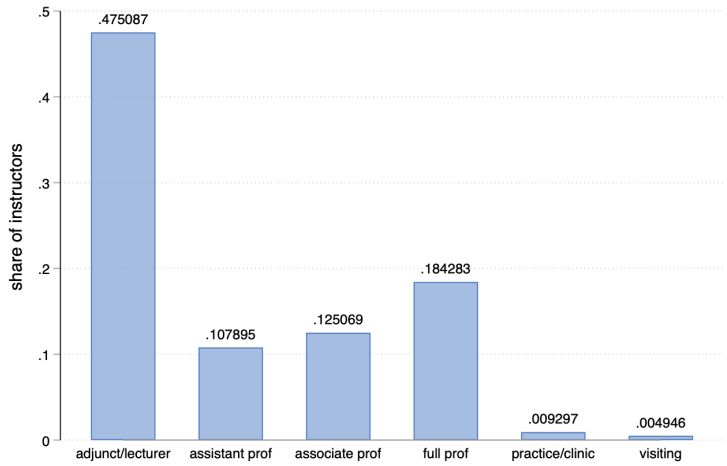
Williams, Heidi L, 2013, Intellectual property rights and innovation: Evidence from the human genome, *Journal of Political Economy* 121, 1–27.

Appendix

For online publication only

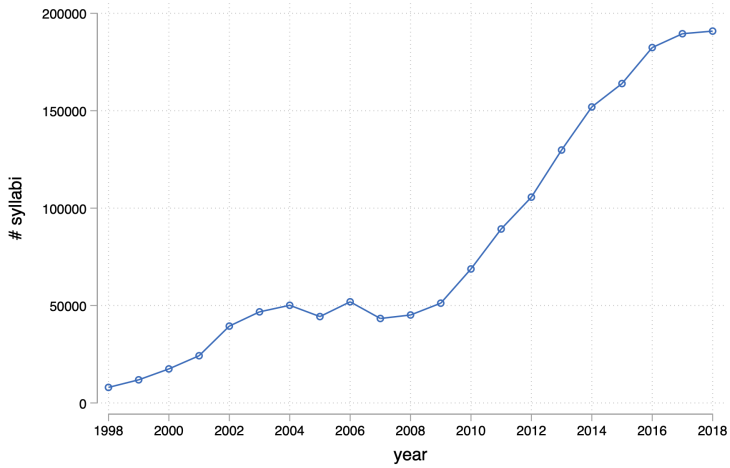
Appendix A Additional Tables and Figures

Figure AI: Distribution of Instructor Job Titles



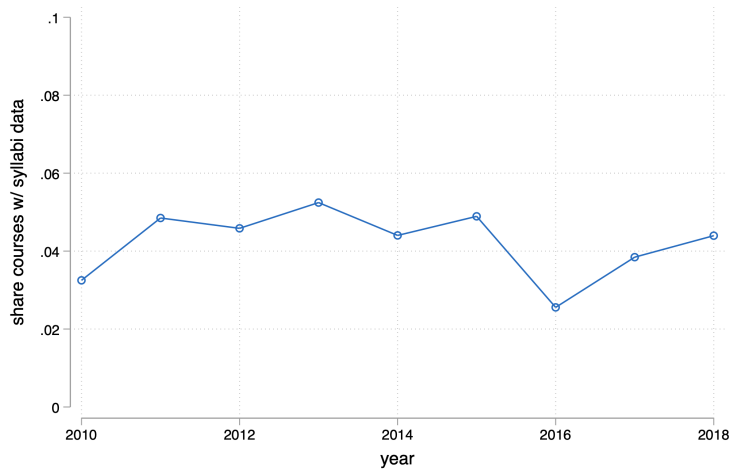
Note: Share of syllabi instructors by job title. The sample is restricted to 32,090 instructors in public institutions for whom title information is available.

Figure AII: Number of Syllabi in the Sample, By Year



Note: Number of syllabi included in final sample, by year.

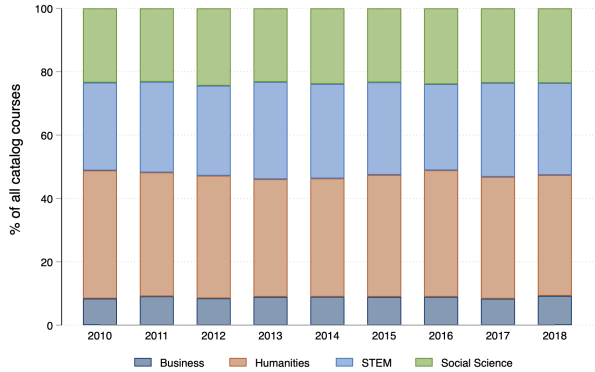
Figure AIII: Share of Catalog Courses in the Syllabi Sample



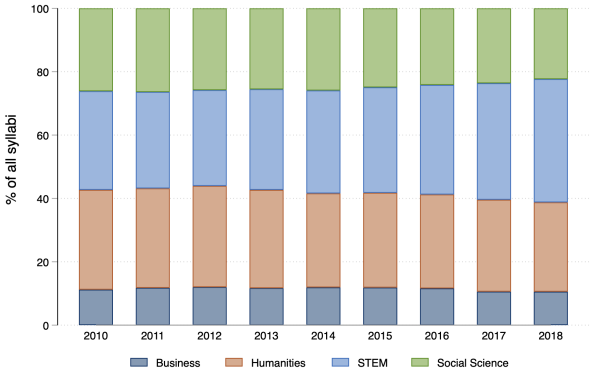
Note: Share of courses from full course catalogs whose syllabi are included in the syllabi sample.

Figure AIV: Macro-Field Coverage, Course Catalogs, and Syllabi Sample

(a) Course Catalog



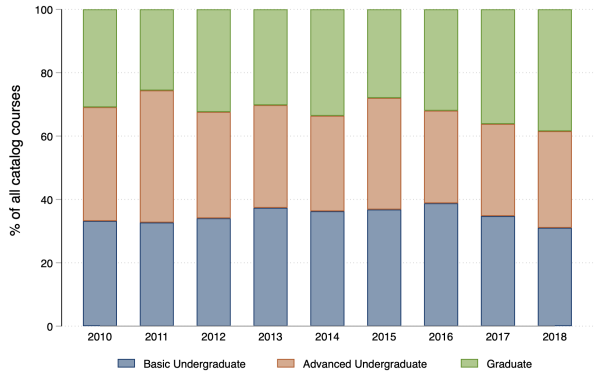
(b) Syllabi Sample



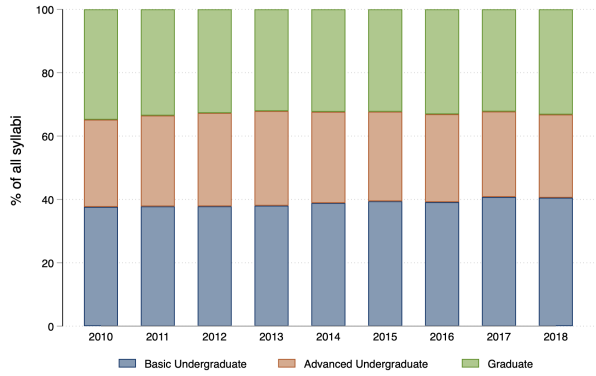
Note: Composition across macro fields, for all courses included in a sample of school catalogs (panel (a)) and for courses included in the syllabi sample (panel (b)).

Figure AV: Course Level Coverage, Course Catalogs, and Syllabi Sample

(a) Course Catalog

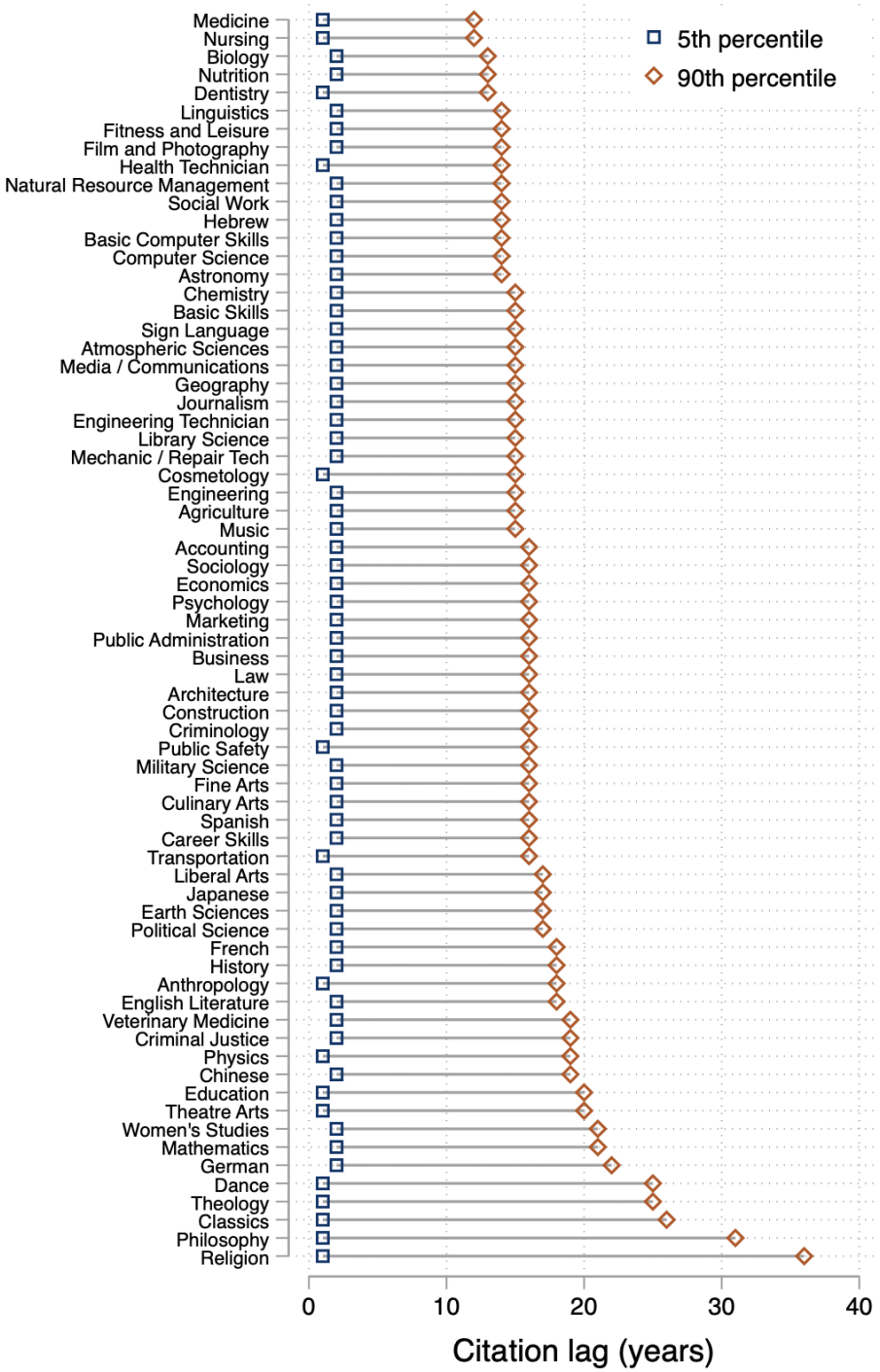


(b) Syllabi Sample



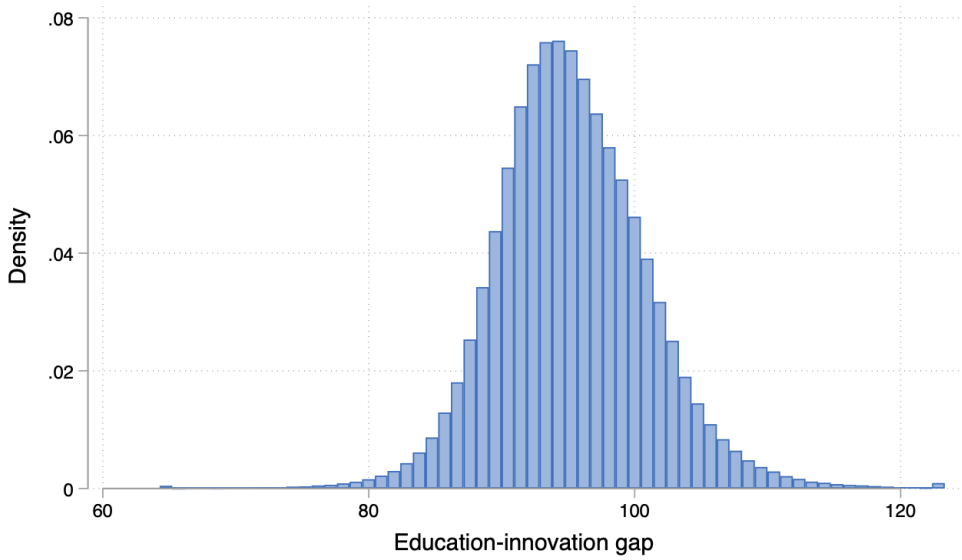
Note: Composition across course levels, for all courses included in a sample of school catalogs (panel (a)) and for courses included in the syllabi sample (panel (b)).

Figure AVI: Citation Lags by Field: 5th and 90th Percentiles



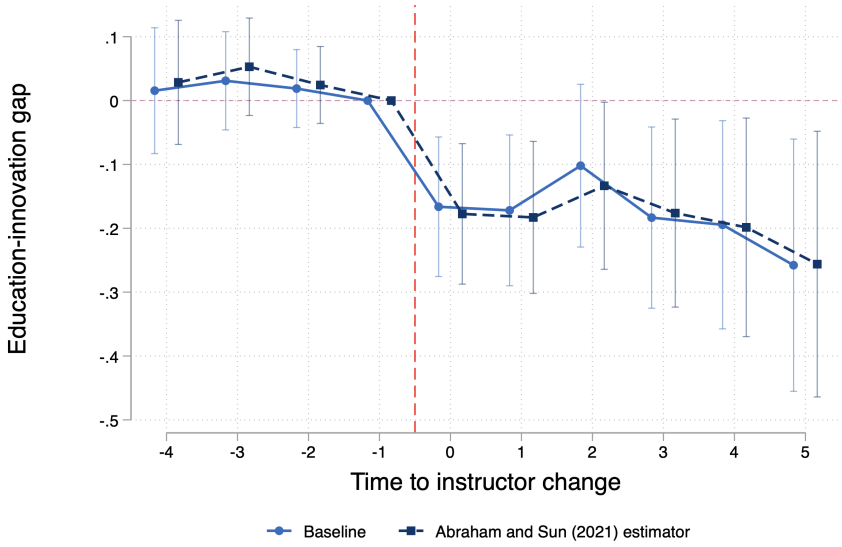
Notes: 5th and 90th percentile of the citation lag distribution in each field, used to calculate the education-innovation gap for the syllabi in each field.

Figure AVII: Education-Innovation Gap: Distribution



Notes: Histogram of the education-innovation gap.

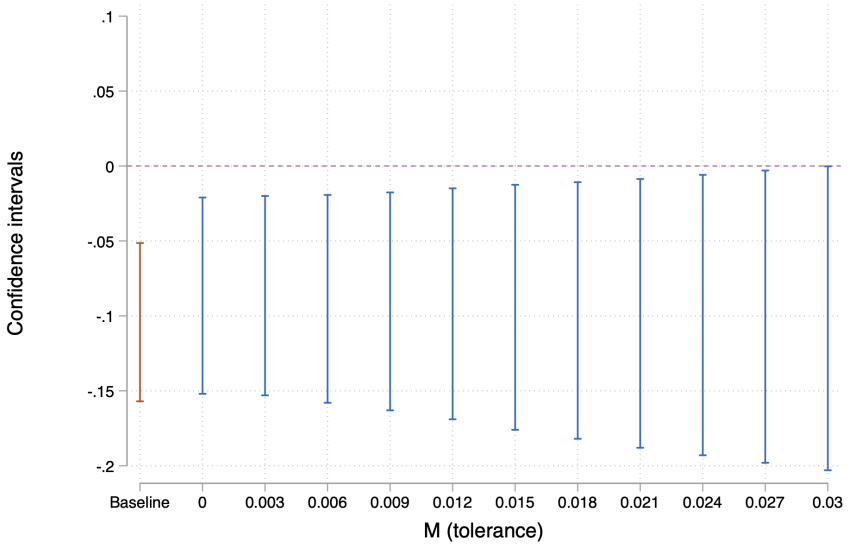
Figure AVIII: Event Study of The Gap Around an Instructor Change: Baseline and Abraham and Sun (2021) Estimator



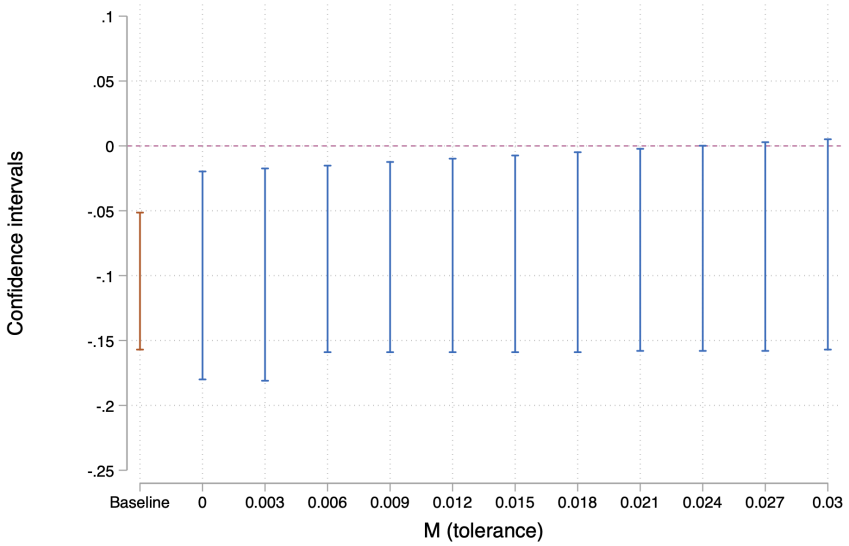
Notes: Estimates of the confidence intervals of δ_k in equation (5), obtained using the baseline approach used in the paper (solid series) and the estimator developed by Sun and Abraham (2021) (dashed series), which accounts for the possibility of heterogeneous treatment effects across cohorts of treated units (in our data, courses that experience an instructor change in different years).

Figure AIX: Event Study of The Gap Around an Instructor Change: Rambachan and Roth (2019) Test for Parallel Trends

(a) Without monotonicity ($\Delta^{SD}(M)$)



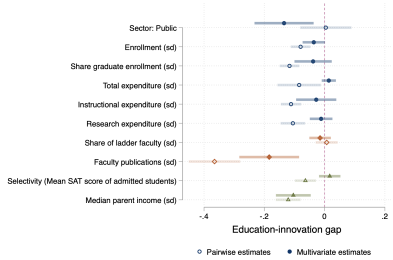
(b) With monotonicity ($\Delta^{SDD}(M)$)



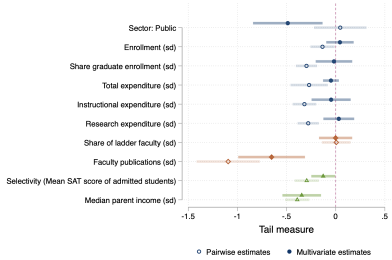
Notes: Sensitivity plots of the confidence intervals of δ_0 in equation (5), constructed following the approach of Rambachan and Roth (2019). The approach tests for violations of the parallel trends assumption and studies their impacts on the point estimates and confidence intervals of interest. Specifically, their proposed test consists in (a) constructing a set Δ of possible deviations from the parallel trends assumption, and (b) constructing the confidence intervals associated with these deviations. In panel (a) we adopt Rambachan and Roth (2019)'s main robustness test, which involves constructing confidence intervals that allow for deviations from linearity up to a tolerance parameter M: defining δ as the trend, $\Delta^{SD}(M) := \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq M, \forall t\}$. In panel (b) we also show confidence intervals for deviations in $\Delta^{SDD}(M)$, analogous to $\Delta^{SD}(M)$ but with the additional assumption that the pre-trend be decreasing. In both panels, the orange series represents baseline OLS confidence intervals; the blue series show confidence intervals as M grows. We allow M to range from zero (linear pre-trends) to the standard error of the coefficient of interest.

Figure AX: School Characteristics and Alternative Measures of Course Novelty

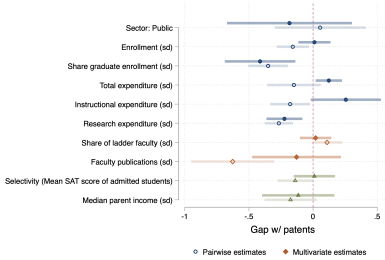
(a) % non-covered new knowledge



(b) Tail measure

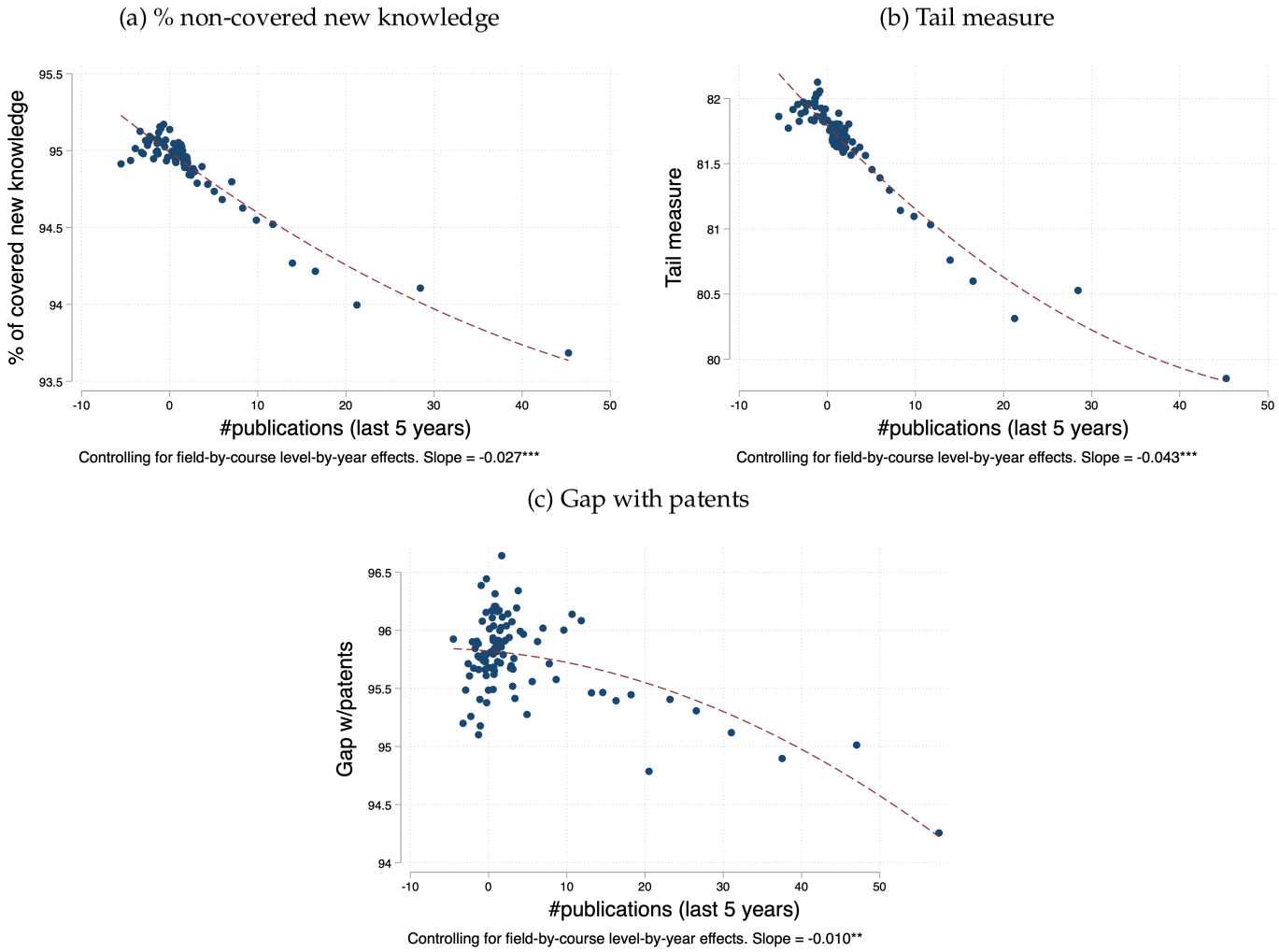


(c) Gap with patents



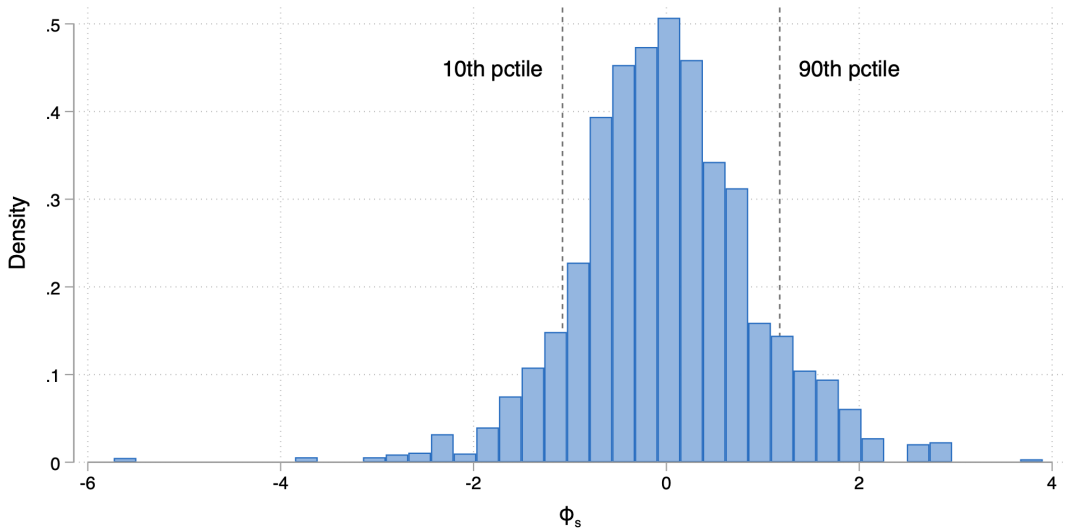
Notes: Point estimates and 95-percent confidence intervals of coefficient β in equation (7), using three alternative measures of course novelty: a measure of non-covered new knowledge, defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top five percent of the word frequency among articles published between $t-3$ and $t-1$ or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$, panel a); a "tail measure," calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution (panel b); and the education-innovation gap calculated using the text of all patents as a benchmark, instead of academic articles (panel c). 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 by pooling syllabi data for the years 1998 to 2018. Standard errors are clustered at the school level.

Figure AXI: Instructor Productivity (# Publications) and Alternative Measures of Course Novelty



Notes: Binned scatterplots of a measure of instructor productivity (the number of publications in the prior five years) and three alternative measures of course novelty: a measure of non-covered new knowledge, defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top five percent of the word frequency among articles published between $t-3$ and $t-1$ or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$, panel a); a "tail measure," calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution (panel b); and the education-innovation gap calculated using the text of all patents as a benchmark instead of academic articles (panel c). Relationships are plotted controlling for field-by-course level-by-year effects.

Figure AXII: Distribution of School-Level Gap



Note: Distribution of the school-level component of the gap, denoted by $\theta_{s(i)}$ in equation (8).

Table AI: Characteristics of Schools Included and Not Included in the Random Catalog Sample

Schools:	In Sample N = 161	Out of Sample N = 1,956	t-stat	p-values
ln Expenditure on instruction (2013)	8.693	8.601	-1.725	0.085
ln Endowment per capita (2000)	6.857	6.483	-1.304	0.193
ln Sticker price (2013)	9.197	9.153	-0.520	0.603
ln Avg faculty salary (2013)	8.890	8.850	-1.897	0.058
ln Enrollment (2013)	8.708	8.634	-0.685	0.494
Share Black students (2000)	0.109	0.112	0.153	0.879
Share Hispanic students (2000)	0.063	0.065	0.183	0.855
Share alien students (2000)	0.025	0.022	-1.030	0.303
Share grad in Arts & Humanities (2000)	7.581	7.958	0.382	0.703
Share grad in STEM (2000)	14.861	14.050	-0.772	0.440
Share grad in Social Sciences (2000)	21.068	19.202	-1.342	0.180

Note: Balance test of universities included and not included in the catalog sample.

Table AII: Alternative Measures of Novelty and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2020)				
	Grad rate	Mean	$P_y \leq 33$ pctile	Median	Mean	P(top 20%)	P(top 10%)	P(top 5%)	P(top 20%| $P_y \leq 20$ pctile)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel (a): Share of non-covered new knowledge, no controls									
Gap (sd)	-0.0401***	-0.0528***	-0.0580***	-0.0434***	-0.0675***	-0.0300***	-0.0265***	-0.0197***	-0.0273***
	(0.0079)	(0.0102)	(0.0111)	(0.0086)	(0.0122)	(0.0063)	(0.0047)	(0.0034)	(0.0061)
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): Share of non-covered new knowledge, with controls									
Gap (sd)	-0.0032	-0.0043	-0.0028	-0.0027	-0.0117**	-0.0053*	-0.0040*	-0.0027*	-0.0005
	(0.0032)	(0.0047)	(0.0059)	(0.0048)	(0.0046)	(0.0031)	(0.0023)	(0.0015)	(0.0034)
Mean dep. var.	0.5816					0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					
Panel (c): Tail measure, no controls									
Gap (sd)	-0.0537***	-0.0644***	-0.0703***	-0.0572***	-0.0886***	-0.0389***	-0.0336***	-0.0248***	-0.0372***
	(0.0083)	(0.0097)	(0.0114)	(0.0087)	(0.0119)	(0.0057)	(0.0046)	(0.0034)	(0.0057)
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 (d): Tail measure, with controls									
Gap (sd)	-0.0020	-0.0094**	-0.0140***	-0.0107**	-0.0172***	-0.0101***	-0.0081***	-0.0052***	-0.0095***
	(0.0034)	(0.0043)	(0.0053)	(0.0046)	(0.0048)	(0.0028)	(0.0021)	(0.0013)	(0.0031)
Mean dep. var.	0.5816					0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					
Panel (e): Gap w/patents, no controls									
Gap (sd)	-0.0232***	-0.0323***	-0.0434***	-0.0282***	-0.0404***	-0.0144**	-0.0140**	-0.0120***	-0.0146**
	(0.0068)	(0.0116)	(0.0122)	(0.0099)	(0.0138)	(0.0067)	(0.0059)	(0.0042)	(0.0064)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763

(Continued)

Table AII. Continued

	Grad rate	Mean	$P_y \leq 33$ pctile	Median	Mean	P(top 20%)	P(top 10%)	P(top 5%)	P(top 20%| $P_y \leq 20$ pctile)
# schools	761	760	734	760					
Panel (f): Gap w/patents, with controls									
Gap (sd)	-0.0049	-0.0003	-0.0023	-0.0007	-0.0039	0.0004	-0.0015	-0.0023*	-0.0014
	(0.0032)	(0.0038)	(0.0044)	(0.0042)	(0.0046)	(0.0025)	(0.0020)	(0.0012)	(0.0029)
Mean dep. var.	0.5816					0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					

Note: OLS estimates of the coefficient δ in equation (9). The variable *Gap (sd)* is a school-level alternative measure of the education-innovation gap (estimated as $\theta_{s(i)}$ in equation (8)), standardized to have mean zero and variance one. In panels (a) and (b), the alternative measure is the share of non-covered new knowledge; in panels (c) and (d) it is a "tail measure;" and in panels (e) and (f) it is the education-innovation gap calculated using the text of all patents as a benchmark for frontier knowledge. The dependent variables 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. (2020), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2020), 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 panels b, d, f, and h control for sector (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. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Appendix B Dataset Construction

B.1 Syllabi

We obtained data on the text of university and college syllabi from the Open Syllabus Project (OSP).[30] The dataset includes nearly 7 million syllabi, collected from 7,365 institutions across the world. OSP provided us with basic information on each syllabus, the full text, and the list of references (papers, textbooks, articles, etc.) included in each syllabus, for a total of 1.8 million unique titles.

We use the following variables from the OSP database:

- `id`: The unique identifier assigned to each syllabus.
- `text`: The text of the syllabus.
- `textmd5`: The md5sum of the text, which can also be used as a unique identifier.
- `language`: The language of the document.
- `year`: The academic year when the syllabus was taught.
- `fieldname`: The name of the academic field most associated with the syllabus.
- `institutionid`: The unique identifier for the institution of the course.
- `unitid`: The IPEDS identifier for the institution.
- `countrycode`: The ISO 3166-1 alpha-2 code of the country the syllabus was taught in.
- `institutionname`: The name of the institution of the course.

In the paper, we focus on syllabi that satisfy the following criteria:

(i) Taught in a four-year, non-online university based in the US (`countrycode` equal to "US") with at least 100 syllabi in the data;

(ii) Taught in English;

(iii) Taught between 1998 and 2018;

(iv) With a word length between 20 and 10,000.

[30] https://opensyllabus.org

The number of syllabi we keep in each step, and the associated syllabi characteristics, are shown in Table BIII.

Table BIII: Summary Statistics of Open Syllabus Project

	# of records	Syllabus word length (raw)	Syllabus word length ("knowledge content")
Original data	6,852,971		
Keep syllabus based in the United States (Syllabus language is English)	3,995,483		
Keep syllabus from four-year university	1,951,933	2,725.41	1,435.09
Year from 1998 to 2018	1,937,284	2,732.09	1,436.77
Extracted syllabus length must be in [20, 10000]	1,901,367	2,279.66	1,057.35
Number of syllabi per institution larger than 100	1,882,224	2,274.55	1,056.77
Remove syllabi from online-only universities	1,706,319	2,226.08	1,010.82

Note: Counts of syllabi, raw word length, and knowledge content (number of words remaining after the cleaning process is complete).

Course catalog data To complement the syllabi data and determine selection patterns into this sample, we also obtained the entire list of course offerings from university catalogs for a sample of US institutions. We begin by randomly selecting 10% of all universities in our sample (212 universities). Then, we manually search and download electronic copies (usually in the PDF format) of university catalogs for those universities for all years available, which list all courses offered in that institution and year. Out of the 212 universities selected, 161 have at least one catalog available. We downloaded and processed a total of 2,348 catalogs for these 161 universities (14.5 catalogs per university). Due to random selection, these schools are representative of the full sample on the basis of standard school-level characteristics. A balance test of characteristics between the full sample and the catalog sample is shown in Table AI.

University catalog data provide the following information: course code, course name, and course level (classified into Basic, Advanced, and Graduate). Some course catalogs also provide a brief course description.

B.1.1 Extracting A Course's Content From Its Syllabus

The full text of a syllabus is contained in the variable `text` of the OSP database. To transform text into usable content, we (i) clean it by removing html language left over from web scraping or correcting obvious errors from OCR procedures; (ii) identify the various sections of the syllabus in it; and (iii) remove text unrelated to content (e.g., course policy, absence policy, accommodation rules). We now explain these steps in more detail.

B.1.2 Cleaning The Raw Text

To clean the text of each syllabus, we proceed as follows:

(i) We use the Unidecode Python Package[31] to convert Unicode text into ASCII text. This includes legacy code that does not support Unicode, non-Roman names on a US keyboard, and ASCII approximations for symbols and non-Latin alphabets.

(ii) We remove browser information, often present in the header of a syllabus, by searching for keywords such as "Internet Explorer", "Newer Browser", "JavaScript Enabled", "Cookies Are", "Download Info", "Login", "Log In", "Print", and "Search".

B.1.3 Identifying Syllabi Sections

Most syllabi contain a set of sections, only some of which are relevant for our analysis. The relevant sections include: instructor and course information (such as code, course level, and title); course description, requirements, and objectives; an outline; homework, exams, and other evaluation methods; and other policies. A syllabus often also includes other information that we do not use in the analysis and, as such, we want to remove. This includes the honor code, policies related to disability, classroom laptop and cellphone policies, and others.

To parse among sections, we developed a supervised algorithm based on a set of section title keywords. The algorithm identifies a section type by searching through a set of keywords belonging to each category. Table BIV provides section types along with the corresponding keywords.

Using these keywords, the algorithm separates the text into different sections of the syllabus by combining keywords with the formatting rules of each syllabus. In Figure BXIII, we use part of a syllabus as an example to present our process step by step.

[31] https://pypi.org/project/Unidecode/

Table BIV: Section Title Keywords List

Section type	Keywords
Course Description	Syllabi, Syllabus, Title, Description, Method, Instruction, Content, Characteristics, Overview, Tutorial, Intro, Abstract, Methodologies, Summary, Conclusion, Appendix, Guide, Document, Module, Introduction, Approach, Lab, Background
Requirements	Requirement, Applicability, Required
Objectives	Objectives, Achievement, Outcome, Motivation, Purpose, Statement, Skill, Competency, Performance, Goal
Outline	Outline, Schedule, Timeline, Guideline
Materials	Text, Material, Resource, Recommend, Reference, Book, Calendar, Textbook, Guidebook
Instructor information	Instructor, About, Email, Phone, Contact, Professor, Staff, Faculty, Information
Projects, homwork, papers, and exams	Personal, Total, Individual, Exercise, Essay, Submission, Assign, Homework, Paper, Final, Examing, Midterm, Term, Semester, Proposal, Application, Demonstration, Program, Task, Report, Pracical, Drafting, Project, Plan, Deadline, Makeup, Advising, Advisor, Survey, Assignment, Planning, Practice, Group, Participation, Team, Research, Activity, Complaint, Design, Analysis, Strategy, Procedure, Working, Work, Exam, Examination, Training, Professional, Test, Case, Discussion, Grade, Presentation, Quiz, Essay, Layout, Sample, Rewrite
Grades	Assessment, Point, Scope, Evaluation, Record, Grading, Composition, Review
Other Policies	Academic, Justice, Administration, Rule, Discipline, Disclaimer, Regulation, Standard, Affair, Dishonesty, Plagiarism, Misconduct, Offence, Medical, Absent, Absence, Trip, Religious, Observance, Ttendance, Honesty, Origination, Originator, Help, Technology, Attendance, Accessing, Service, Oppotunity, Administrative, Accommodation, Support, Policy, Right, Responsibility, Disability, Weather, Integrity, Copyright
Notes	Remark, Notice, Additional, Acknowledgement, Absolutely, Absolute, Important, Note, Cannot, Can, Must, Should, Will, Please, No
Other Words	Course, Lecture, Catalog, Campus, Commuity, Class, Classroom, College, Univerity, Discussion, Seminar

Note: Keywords used to identify the corresponding section types of a syllabus. In the implementation, we use both the singular and plural versions of each term.

1. For each syllabus, we identify the section titles based on the word list described above and the formatting features. We mark all cases in which the section title phrases appear as all uppercase or consecutive initial capital letters using regular expressions.

    - In Figure BXIII, underlined sentences satisfy the features of a section title, such as "Course Description".

2. We divide the syllabus into parts, and we use Arabic numerals to mark them out. Finally, we select sections with relevant titles and extract the cleaned text.

    - In Figure BXIII, we focus on highlighted sections, such as "Course Objective," "Prerequisites," and "Text".

B.1.4 Extracting Additional Information

Instructor Names To extract the name of the instructor from each syllabus, we build a neural network model based on the BiLSTM-CNNs-CRF model for named entity recognition (NER).[32] The training/test dataset is built via the following three steps:

(i) We select syllabi that contain at least one keyword such as "Doctor", "Doctors", "Dr", "Professor", "Prof", "Instructor", "Instructors", "Tutor", "Tutors" in the first 3,500 characters.

(ii) We use the Spacy[33] package to identify whether the words following those keywords are names of people (entity label is "PERSON").

(iii) We process the syllabus text sentence by sentence as the training and test data of the model.

We also apply a few additional filters: (a) we remove single letter names; (2) all the words in the name are required to appear in the Python Library *English First and Last Names Data Set*[34]; (c) after the first two filters, we only keep the first instructor name. With this algorithm, we are able to assign an instructor name to 86.23% of all syllabi. The out-of-sample precision of this algorithm is 85.18%.

Course Level: Basic, Advanced, Graduate To assign a course level (basic undergraduate, advanced undergraduate, and graduate) to each syllabus, we trained a Natural Language Processing (NLP) algorithm. Our training sample consists of 56,831 syllabi taught in universities for which we

[32] BiLSTM-CNNs-CRF model for named entity recognition (NER), Ma and Hovy (2016).
[33] https://spacy.io/
[34] https://github.com/philipperemy/name-dataset

have catalog information and for which we can manually code the course levels. Specifically, in the catalog data, we label a course as basic undergraduate if the course belongs to the undergraduate catalog of a university and the course code starts with 1 or 2; we label the course as advanced undergraduate if the course belongs to the undergraduate catalog and the course code starts with 3 or 4; finally, we label the course as graduate if the course belongs to the graduate catalog or the first digit of the course code is larger than 4. We link syllabi to catalog information using institution and course code. Once we have obtained course levels for these syllabi, we use course levels as labels and the text of each syllabus as input in the training model. The model we use is Distilled BERT[35] (Sanh et al., 2019), accessed via the transformers library.[36] The out-of-sample prediction precision is 85.04%.

Course code Our data extraction process allows us to obtain the course code corresponding to each syllabus. However, these courses are institution-specific and often vary over time. To be able to identify courses of the same level (e.g., basic undergraduate) covering the same topic (e.g., Principles of Microeconomics), both within and across schools, we proceed as follows. First, we construct a unified within-school course code using the raw course code and the course name. We do so as follows: (a) we remove the punctuations and multiple whitespaces from codes and names; (b) for course names, we further remove stop-words and isolate the course stem name (the common base form of the words). We then consider two courses as sharing a course code if (a) they share the same name and code or (b) they share the same name, even if the course code changes over time. This procedure accounts for the possibility that the course code system might have changed within a school over time.

Once we have a disambiguated identifier for courses within the same school, we assign courses a cross-school identifier. Specifically, we assign two courses the same cross-school identifier if they share the same standardized course name.

B.1.5 References and Recommended Readings in Each Syllabus

In addition to syllabi text and metadata, OSP provided us with two additional datasets: "Matches" and "Catalog." "Matches" allows us to link syllabi to records in "Catalog." "Catalog" is the set of 1.8 million bibliographic records assigned to at least one syllabus. We use the following variables from the "Matched" dataset:

- `MatchID`: The unique identifier of the match

[35] https://arxiv.org/abs/1910.01108
[36] https://huggingface.co/transformers/index.html

- `ID`: The id of the syllabus
- `WorkID`: The id of the catalog record

We use the following variables from the "Catalog" dataset:

- `WorkID`: The id of the catalog record
- `Publicationtype`: The type of publication ("journal" or "book")
- `Publicationyear`: The year of publication

B.1.6 Syllabi Field

The OSP database classifies syllabi into one of 69 fields. For some of our analyses, we group these into macro-fields. The grouping is illustrated in Table BV.

Table BV: Categorization of Course (Macro-)Fields

Macro-field	Fields
Business	Business, Accounting, Marketing, Public Administration
Humanities	English Literature, Media / Communications, Philosophy, Theology, Criminal Justice, Library Science, Classics, Women's Studies, Journalism, Religion, Sign Language, Liberal Arts, Music, Theatre Arts, Fine Arts, History, Film and Photography, Dance, Anthropology, Japanese, French, Chinese, German, Spanish, Hebrew
Science	Mathematics, Biology, Chemistry, Physics, Earth Sciences, Astronomy, Atmospheric Sciences, Dentistry, Medicine, Nutrition, Nursing, Veterinary Medicine, Natural Resource Management
Engineering	Computer Science, Engineering, Architecture, Agriculture, Basic Computer Skills, Engineering Technician, Transportation
Social Sciences	Psychology, Political Science, Economics, Law, Social Work, Geography, Education, Linguistics, Sociology Education, Criminology
Other	Fitness and Leisure, Basic Skills, Mechanic / Repair Tech, Cosmetology, Culinary Arts, Health Technician, Public Safety, Career Skills, Construction, Military Science

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

Figure BXIII: Dividing A Syllabus Into Sections: An Example

Econ 561a Yale University Fall 2005
Prof. Tony Smith (Part I) Prof. Michael Keane (Part II)
Syllabus for **COMPUTATIONAL METHODS FOR ECONOMIC DYNAMICS** ECON 561a

Course Objectives:
Most of the dynamic economic models used in modern quantitative research in economics do not have analytical (closed-form) solutions. For this reason, the computer has become an indispensable tool for conducting research in dynamic economics. The goal of this two-part course is precisely to teach students computational tools for conducting numerical analysis of dynamic economic models. The focus of the first half of the course, taught by Prof. Tony Smith, is on solving dynamic programming problems and on computing competitive equilibria of dynamic economic models. The first half of the course also provides an introduction to some of the basic tools of numerical analysis, including minimization, root-finding, interpolation, function approximation, and integration. The focus of the second half course, taught by Prof. Michael Keane, is on solving and estimating discrete-choice dynamic programming models of economic behavior. Taken together, the two halves of the course provide students with a thorough introduction to the numerical analysis of dynamic economic models in both microeconomics and macroeconomics.

Contact Information (Prof. Tony Smith)
Office: 28 Hillhouse, Room 306 Office phone: (203) 432-3583
Email address: tony.smith@yale.edu Course Web site: www.econ.yale.edu/smith/econ561a
Office hours: Thursdays from 10AM–noon, or by appointment

Class Meetings:
The course meets on Mondays and Wednesdays from 2:30PM to 3:50PM in a room to be determined.

Prerequisites:
This course is designed for graduate students in economics who have taken first-year graduate courses in microeconomics, macroeconomics, and econometrics. No prior knowledge of either numerical methods or computer programming is assumed, but some familiarity with a programming language would prove helpful.

Texts:
The required textbook for this course is:
Numerical Recipes in Fortran 77: The Art of Scientific Computing, Second Edition (Volume 1 of Fortran Numerical Recipes) by William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery (Cambridge University Press, 1992). This book, as well as its (optional) companion Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing, Second Edition (Volume 2 of Fortran Numerical Recipes), is available online at: www.library.cornell.edu/nr/.
Other (optional) books that students might find useful are:
- Numerical Methods in Economics by Kenneth L. Judd (MIT Press, 1998).
- Handbook of Computational Economics (Volume 1), edited by Hans M. Amman, David A. Kendrick, and John Rust (North-Holland, 1996).
- Computational Methods for the Study of Dynamic Economies, edited by Ramon Marimon and Andrew Scott (Oxford University Press, 1999).
- Dynamic Economics: Quantitative Methods and Applications by Jérôme Adda and Russell Cooper (MIT Press, 2003).
- Applied Computational Economics and Finance by Mario J. Miranda and Paul L. Fackler (MIT Press, 2002).

Grading:
The course grade will be based on two (equally-weighted) projects, one for the first part of the course and one for the second part of the course. Each project consists of writing a program in Fortran to solve an assigned problem. Students must submit their code as well as a brief (roughly five pages) description of their numerical findings. The first project will involve solving for the competitive equilibrium of a dynamic macroeconomic model; the second project will involve solving and estimating a discrete-choice dynamic programming model. Fortran is the language of choice for most researchers in computational economics; requiring that the code for the projects be written in Fortran will help students to become proficient in this powerful and useful language. The first project is due on Monday, November 14 and the second project is due at the end of the semester. Occasional short programming problems may also be assigned as the course proceeds. The purpose of these assignments is to help students develop the skills they need to complete the projects; these assignments will not be graded.

Approximate Schedule of Lectures (Part I)
I. **INTRODUCTION**
Lecture 1 Introduction to numerical dynamic programming (built around the stochasticgrowth model and the Aiyagari (1994) model). General considerations in numerical analysis: convergence, roundoff error, truncation error. Numerical differentiation.
Readings:
- Aiyagari, S.R. (1994), "Uninsured Idiosyncratic Risk and Aggregate Saving," Quarterly Journal of Economics 109, 659–684.
- Numerical Recipes: Chapters 1 and 5.7
- Judd: Chapters 1, 2, and 7.7

II. **BASIC NUMERICAL METHODS**
Lecture 2 Root-finding in one or more dimensions: bisection, secant method, Newton's method, fixed-point iteration, Gauss-Jacobi, Gauss-Seidel, Brent's method.
Readings:
- Numerical Recipes: Chapter 9

……

Note: Example of a syllabus from OSP, in its original format. Subsections are identified using the algorithm described in this appendix.

B.2 Academic Publications

To construct the education-innovation gap, we collect a large sample of academic articles from top journals. We describe here how this sample is defined, constructed, and collected.

B.2.1 List of Top Journals

We begin by compiling a list of top academic journals within each discipline. Our starting point is the Journal Citation Reports (JCR), an annual report published by Thomson Reuters (formerly ISI) to provide citation and publication data of academic journals in the science and social science fields by means of the impact factor.[37] We consider as top journals those that were ranked within the top ten of their respective field at least once since their establishment. This leaves us with 3,962 journals in 223 fields.

B.2.2 Collecting Academic Articles

Having compiled a list of top journals, we collect information on all the articles ever published in these journals. These data come from Scopus, an Elsevier-owned database containing abstracts and citations of academic articles.[38] To extract the metadata of journal articles, we access Scopus's API and search for the ISSN of each journal ("ISSN(0022-1082)"). We then extract all the metadata of all articles of the relative journal for all available years. We focus our attention on the following variables:[39]

- `EID`: electronic ID, used as the unique identifier of each article;
- `title`: title of the article;
- `ISSN`: ISSN of publisher;
- `coverdate`: publication date;
- `description`: abstract;
- `authkeywords`: keywords.

Our initial search yielded 20,779,713 articles, of which we discarded those without an abstract.

[37] https://jcr.clarivate.com/

[38] https://www.scopus.com

[39] The full list of variables available through Scopus is available at https://dev.elsevier.com/guides/ScopusSearchViews.htm

B.2.3 Data Cleaning

The main information from academic articles that we use in our analysis is the abstract, contained in the variable `description` of the SCOPUS database. We further clean the content of this variable to remove copyright disclaimers, usually present at the beginning or at the end of each abstract and unrelated to content. We do this using keyword recognition techniques. Starting from the first sentence of an abstract, we remove it if it contains at least one of the following words: "copyright", "©", "published", "publisher", "all right", or "all rights reserved". We repeat this procedure until the first sentence does not contain any of these words. We then repeat the same procedure starting from the next sentence.

B.3 Research Productivity

We use information from Microsoft Academic (MA) to measure the research productivity of all people listed as instructors in the syllabi. We download these data from Microsoft Academic Knowledge Graph (MAKG).[40] MAKG is a large resource-description framework (RDF) knowledge graph with over eight billion triples containing information about scientific publications and related entities, including authors, institutions, journals, and fields of study. The dataset is based on the Microsoft Academic Graph and licensed under the Open Data Attributions license. For each researcher, Microsoft Academic lists publications, working papers, other manuscripts, and patents, together with the counts of citations to each of these documents. Due to differences in counting citations, Microsoft Academic citations do not necessarily match those from similar services such as Web of Science or Google Scholar. The correlations between all these services' citations numbers, however, are very high.

We link instructor records from the text of the syllabi to Microsoft Academic records using names, a person's history of institutions, and research fields. In the sample of syllabi used in our analysis, 44.23% (697,756 / 1,487,820) have an instructor record, covering 332,063 unique instructors. Of these instructors, 40.76% (135,364 / 332,063) are matched to a Microsoft Academic profile.

B.4 Patents

We obtain data on patents from the publicly available Patent Full-Text Database (PatFT)[41] of the US Patent and Trademark Office (USPTO). This database provides records for all patents ever issued since 1976. We use a web crawler to collect the text content of patents over this period, which

[40] We download the data based on the Microsoft Academic Graph data as of 2020-05-29 from https://zenodo.org/record/3936556#.YFndr2Qza3J

[41] http://patft.uspto.gov/netahtml/PTO/index.html

includes patents with numbers ranging from 3,850,000 to 10,279,999. We use the following variables for each patent record:

- `PatentNumber`: The unique identifier assigned to each patent record
- `Abstract`: The abstract in each patent filings
- `Year`: The year that the patent was issued
- `Class`: The International Patent Classification (IPC) assigned to each patent

B.5 National Science Foundation and National Institute of Health Grants

We collect information on grants awarded by the National Science Foundation (NSF)[42] and the National Institutes of Health (NIH)[43] to construct measures of research investment and productivity. These data are provided directly by the respective organizations; the versions used in the paper were accessed on May 25, 2021.

The NSF grant data include 480,633 grants with effective starting years ranging from 1960 to 2022. The NIH grant data include 2,566,358 grants with effective years ranging from 1978 to 2021. Both NSF and NIH grant data contain information on the host institution (institution name, country, state, and city) and the investigator (investigator name and role). In the NSF data, investigators can be listed under four figures: principal investigator (PI), co-PI, former PI, and former co-PI. In the NIH data, they can be listed under two figures: contact and non-contact.

B.5.1 Linking NSF/NIH Institutions to Syllabi Institutions

To link grants to institutions in the syllabi data and IPEDS, we use information on the institution's name and location (country, state, and city). To do so, we first perform an exact match using institution names as listed in the NSF/NIH data and in IPEDS, stripped of punctuation marks and stop words (including "and" and "the"). Then, for the remaining unmatched NSF/NIH institutions, we conduct a fuzzy matching based on name and location. We require the matching algorithm to meet the following two conditions: (1) the two institutions must be in the same city; (2) the fuzzy matching ratio must be larger than a certain threshold (specifically, we use partial ratio and token set ratio in the FuzzyWuzzy Package).[44] This method sometimes leads us to match a NSF/NIH

[42] https://www.nsf.gov/awardsearch/download.jsp

[43] https://exporter.nih.gov/ExPORTER_Catalog.aspx

[44] The package uses Levenshtein Distance to calculate the differences between sequences; its homepage is https://github.com/seatgeek/fuzzywuzzy, and we use a threshold of 80.

institution to multiple IPEDS institutions. In this case, we consider the IPEDS institution with the largest average matching ratio .

We are able to match 11.30% (2,402) of NSF institutions to IPEDS, covering 82.05% (= 394,383 / 480,633) of all NSF grants. Similarly, we are able to match 6.73% (1,573) of NIH schools to IPEDS, covering 66.53% (= 1,707,498/2,566,358) of all NIH grants. The unmatched NSF and NIH institutions are mostly non-academic, private, or not-for-profit research institutes.

B.5.2 Linking NSF/NIH Investigators to Instructors

Next, we match grant investigators to course instructors in the syllabus data. We do this via a fuzzy matching algorithm using names. The NSF and NIH data provide different investigator information to be used in the fuzzy matching, so the matching methods differ slightly between the two datasets.

NSF To match NSF investigators to instructors, we first remove duplicates within NSF based on first name, last name, email, and institutions since NSF does not provide investigator unique identifiers. We consider two investigators to be the same person if (1) they share the same email or (2) they have exactly the same first name and last name in the same school in a certain year. Next, we perform a many-to-one fuzzy matching between NSF investigators and syllabi instructors based on the names and history of institutions at which the researcher was employed. We proceed in three steps:

(i) After removing any punctuation marks from name strings, we fuzzy-match each NSF investigator name with syllabus instructor names. We calculate matching scores using the Whoswho Package[45], a Python library for determining whether two names belong to the same person.

(ii) If a match has a score of 100, we consider it successful. For matches with scores larger than 95 who have ever worked at the same school, assign an investigator to one and only one instructor as follows.

   (a) If an NSF investigator and a set of syllabi instructors have spent some common period of time at the same institution as we can observe it, we link the investigator to the instructor with the highest matching score.

   (b) If they have not spent any common period of time at the same institution, we link the investigator to the instructor with the highest matching score and lowest temporal distance between the time spent at each institution.

[45] https://github.com/rliebz/whoswho

(iii) For matches with a matching score larger than 95 but in different schools,

(a) If an instructor and an investigator are observed for the same period of time in the two datasets, we choose the match with the highest matching score.

(b) Otherwise, we choose the matching with the highest matching score and shorter time distance between observed periods between the two datasets.

This procedure leaves us with 232,206 unique investigators, 23.31% (= 54,118 / 232,206) of whom can be matched to one syllabus instructor, and corresponding to 44.28% (= 208,857 / 471,646) of all grants.

NIH Data from NIH contain investigator unique identifiers, which implies that we do not have to remove duplicates. We use these to perform a one-to-one matching between each NIH investigator and a syllabus instructor. We follow the same process as with NSF grant data. This procedure leaves us with 298,687 unique investigators, 10.07% (= 30,087 / 298,687) of whom can be matched to one syllabus instructor, corresponding to 17.69% (= 450,339 / 2,546,123) of all grants.

Our final grant data combine information from NSF and NIH grants. The syllabi sample used in our analysis covers 332,063 instructors, of whom 17.51% (= 58,136 / 332,063) have at least one NSF or NIH grant, accounting for 20.93% (= 311,350 / 1,487,820) of all syllabi.

B.6 Instructors' Job Titles and Salaries

We are able to collect the salaries of instructors employed at 490 public colleges and universities in 16 states. As the regulations on the disclosure of public-sector employees' salaries vary across states and over time, the temporal coverage of our data differs across states. Table BVI describes the coverage and source of the salary data.

Together with the salary data, the job title of each employee is also disclosed. We are able to identify the following titles: assistant professor, associate professor, full professor, lecturer, adjunct professor, clinical professor, professor of practice, and visiting professor. This information is available for 32,090 instructors in our syllabi sample (9.7 percent of all instructors and 13 percent of public-sector instructors), employed in 278 public institutions in 13 states. Table BVII describes how we assign job titles based on the information available in the data.

Table BVI: Coverage and Source of Salary and Job Title Data

State	Data available for	Source
CA	2011-2018	https://transparentcalifornia.com/agencies/salaries/
CT	2010-2018	http://transparency.ct.gov/html/searchPayroll.asp
GA	2010-2018	https://open.ga.gov/openga/salaryTravel/index
IA	2009-2018	https://www.legis.iowa.gov/publications/fiscal/salaryBook
IL	2010-2018	https://salary.bettergov.org/
IN	2012-2018	https://gateway.ifionline.org/default.aspx
KS	2012-2018	http://kanview.ks.gov/DataDownload.aspx
MA	2010-2018	https://cthrupayroll.mass.gov/
MD	2012-2018	https://salaries.news.baltimoresun.com/
MI	2014-2018	https://www.mackinac.org/salaries
MN	2011-2018	https://mn.gov/mmb/transparency-mn/payrolldata.jsp
NV	2009-2018	https://transparentnevada.com/
NY	2008-2018	https://www.seethroughny.net/payrolls
OK	2010-2018	https://data.ok.gov/dataset
RI	2011-2018	http://www.transparency.ri.gov/payroll/
WA	2016-2018	http://fiscal.wa.gov/salaries.aspx

Note: States for which instructor salary and job title data are available, together with available year and source.

Table BVII: Assigning Job Titles

Job Title	Definition
Adjunct Professor	Any word of the job title starts with ”adjunct”, ”adj”, ”temporary”, ”temporari”, ”temporar”, or ”part time”.
Clinical Professor	Any word of the job title starts with ”clinic” or ”clin”.
Professor of Practice	Any word of the job title starts with ”practic” or ”pract”.
Visiting Professor	Any word of the job title starts with ”visiting” or ”visit”.
Lecturer	(1) Any word of the job title starts with ””lectur”, ”lect”, ”instructor”, ”instruct”, ”instr”, ”teacher”, or ”teach”; (2) AND any word of the job title does not end with ”ship”; (3) AND job title is not identified as adjunct professor, clinical professor, professor of practice, and visiting professor.
Professor	(1) Any word of the job title starts with ””professor”, ”prof”, or ”tenur”; (2) OR any word of the job title includes ”tenr trk” or ”tenur track”; (3) AND any word of the job title does not end with ”profession”; (4) AND job title is not identified as adjunct professor, clinical professor, professor of practice, or visiting professor.
Assistant Professor	(1) Job title is identified as professor; (2) AND any word of the job title starts with ””assist”, ”asst”, or ”assi”.
Associate Professor	(1) Job title is identified as professor; (2) AND any word of the job title starts with ””associ”, ”assoc”, or ”asso”.
Full Professor	(1) Job title is identified as professor; (2) AND detailed job title is not identified as assistant professor or associate professor.

Note: Procedure used to assign job titles to salary records.

Appendix C Calculating The Education-Innovation Gap: Additional Details and A Simulation Exercise

We now explain in detail the process employed to identify the knowledge terms used in our analysis, extract them from the text of syllabi and academic publications, and calculate the gap.

C.1 Extracting Knowledge Terms From Each Document

Dictionary The first step is to build a dictionary, i.e., a list of all knowledge terms. We use the list of all unique words and expressions ever used as a keywords in academic publications. We extract these keywords from the data described in Section B.2.

Term Extraction Next, we convert the text content of each document (syllabi and academic papers) into numerical data for statistical analyses. To do so, our starting point is to clean the text. First, we convert the text of each document into ASCII text using the Unidecode Python Package.[46] This allows us to handle host legacy code that does not support Unicode, non-Roman names on a US keyboard, and ASCII approximations for symbols and non-Latin alphabets. Next, we convert all capitalized characters to lowercase and use the NLTK Python Toolkit to strip out all non-word text elements, such as punctuation marks, numbers, and HTML tags. We also remove all occurrences of 280 "stop words", which include propositions, punctuation marks, pronouns, and other words that carry little semantic content.[47]

Once we have cleaned the text, we convert it into numerical data using a term-extraction algorithm called NGramMatch. This algorithm performs exact string matching of the text of each document, consisting in N-grams with N ranging from 1 to 7, with the dictionary. To do so, the algorithm extracts N-grams from text to form a basic term set. Then, it filters out all the terms which cannot be linked to any dictionary entry. In the final set, the algorithm assigns each document a frequency vector based on matched dictionary words.

C.2 A Simulation Exercise

To better understand how the education-innovation gap captures the academic novelty of a syllabus's content and to illustrate its properties, we perform a simulation exercise. In this simulation, we manually construct a set of syllabi by combining dictionary words that can be found in academic

[46] https://pypi.org/project/Unidecode/

[47] We create a list of stop words as the union of all single letters and Stanford CoreNLP package: https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt.

publications. Each syllabus is characterized by a year (t, ranging from 1998 to 2018 to match our data), a known gap (*gap*, ranging between 0 and 1), and a parameter governing its style (*style*). For each of these syllabi, we calculate the education-innovation gap with the procedure described in the text. We then compare it with the known gap to assess its performance.

The three parameters characterizing each syllabus govern the way the terms in it are drawn from three different buckets of words: new knowledge terms, old knowledge terms, and style words.

- New knowledge terms are (i) in the top 5% of the word frequency distribution among articles published between $t-3$ and $t-1$ or (2) words that appear in articles published between $t-3$ and $t-1$ but not those published between $t-15$ and $t-13$.
- Old knowledge terms are (i) in the top 5% of the word frequency distribution among articles published between $t-15$ and $t-13$ or (2) words that appear in articles published between $t-15$ and $t-13$ but not those published between $t-3$ and $t-1$.
- Style words are those terms that appear in academic articles but do not belong to the previous two groups.
- gap is the ratio between the share of old and new knowledge words in a syllabus.

To generate each syllabus, we use the following algorithm:

- We assign the syllabus a length of L, where $L = 10*U$ and U is drawn from a discrete uniform distribution between 1 and 50 (so that L lies between 10 and 500, with increments of 10, and can therefore take 51 possible values).
- We assign the syllabus a number $L_s = L \times style$ style words, where $style$ ranges between 0.01 and 0.1 in increments of 0.01 (and can therefore take 11 possible values).
- The remaining $L - L_s = L_k$ words in the syllabus are drawn from the new and old knowledge terms buckets. Among these, $L_k \times (1 + gap)^{-1}$ are from the new knowledge terms bucket and $L_k \times gap \times (1 + gap)^{-1}$ are from the old knowledge terms bucket.

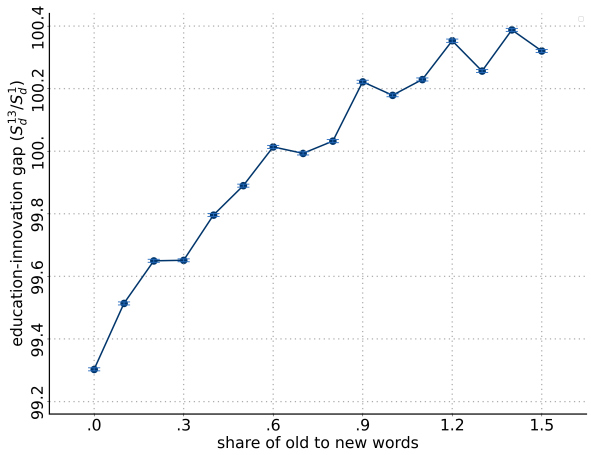
With this algorithm, we generate 10 syllabi for each set of parameters $\{t, L, style, gap\}$. The total number of generated syllabi is thus $= 10 \times 21 \times 46 \times 11 \times 16 = 1,700,160$, which is close to the sample size in our data.

Figure BXIV (panel (a)) shows the relationship between gap and our estimated education-innovation gap. The correlation between these variables is strong and equal to 0.96. By contrast, in panel (b)

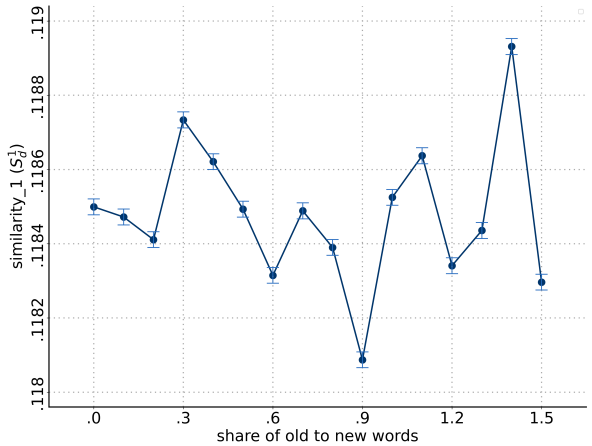
we show the relationship between gap and the cosine similarity between the syllabus and new publications (appeared in $t-3$ to $t-1$), i.e., the denominator of the education-innovation gap. This relationship is much noisier. This is likely to occur because a simple cosine similarity is likely to be affected by the overall style of the syllabus, whereas the gap is not.

Figure BXIV: Simulated Syllabi and Their "True" Gap Measure

(a) gap and The Education-Innovation Gap



(b) gap and Cosine Similarity with New Publications



Note: Panel (a) shows the relationship between gap and the education-innovation gap as defined and constructed in the paper. Panel (b) shows the relationship between gap and the cosine similarity between the syllabus and new publications (appeared in $t-3$ to $t-1$).