Frontier Knowledge in College and Student Success*

Barbara Biasi† Song Ma‡

October 6, 2025

Abstract

This paper studies the teaching of frontier knowledge in higher education and its impact on students. Using text analysis on 2 million course syllabi and 20 million academic articles, we develop a measure called "frontier knowledge proximity," capturing how closely course content aligns with current scholarly research. We document significant variation in frontier knowledge proximity across courses, even within the same institution, and demonstrate that these differences substantially affect student outcomes. Linking syllabi to individual student records from Texas and leveraging unexpected syllabus updates, we show that increases in proximity improves both educational outcomes (graduation, major retention, and graduate school enrollment) and earnings. Educational gains are notably larger among median-ability and lower-income students, whereas earnings benefits disproportionately accrue to higher-ability and higher-income students. These findings indicate that frontier knowledge exposure can narrow socioeconomic disparities in education but remains complementary to students' existing resources. We conclude by showing that instructors, particularly research-active faculty, are the main drivers of differences in frontier knowledge proximity.

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

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

*This paper subsumes and expands a previous working paper titled "The Education-Innovation Gap." The conclusions of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas. 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 various institutions, 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. Meghna Baskar, Xugan Chen, Yajie Luo, and Xinhui Yu provided outstanding research assistance. We thank the Yale Tobin Center for Economic Policy, The Broad Center at the Yale School of Management, the Yale School of Management Dean's Fund, the Yale Center for Research Computing, Yale University Library, and the Yale International Center for Finance for research support. All errors are our own.

†Yale School of Management and NBER, barbara.biasi@yale.edu;

‡Yale School of Management and NBER, song.ma@yale.edu.

1 Introduction

The importance of frontier knowledge as an engine of innovation and economic growth is well established (Romer, 1986). At the aggregate level, it is essential for generating new ideas that fuel innovation (Jones, 2009; Williams, 2013; Iaria et al., 2018). At the individual level, it promotes the accumulation of human capital and skills critical for labor-market success (Goldin and Katz, 2010; Acemoglu and Autor, 2011). Understanding how individuals gain access to frontier knowledge thus carries important policy implications.

Higher-education institutions (HEIs), such as colleges and universities, are key places of frontier knowledge creation (Andrews, 2023). Are they also places of frontier knowledge dissemination? The answer to this question is ex ante not obvious. Institutions differ substantially in their resources and organization and instructors exercise considerable freedom in designing course content, so students may receive very different exposure to frontier knowledge depending on which university they attend and which courses they enroll in.[1] *Should* universities be places of frontier knowledge dissemination? This answer is also not obvious: While this type of knowledge may make some students more innovative and successful in the labor market, the average student may benefit more from foundational knowledge. Despite the relevance of these questions, the presence of frontier knowledge in university curricula and its impact for students has remained understudied. This is likely due to a major data limitation: the lack of information on course content in standard administrative datasets.

This paper addresses this gap by tapping into a novel source of information on course content: the full text of course syllabi. To quantify the presence of frontier knowledge in each course we introduce a new measure, the *frontier knowledge proximity*, based on the textual similarity between each syllabus and recent high–impact-journal publications.[2] We use this measure to answer the two questions posed above. First, we show that some courses teach a fair amount of frontier knowledge while others—even within the same institution, field, and course level—do not. Second, we show that exposing students to frontier knowledge increases the average student's likelihood of persisting in their major, graduating, pursuing graduate studies, and earning higher wages after graduation. We do this by linking syllabi to individual-level student records and leveraging quasi-random variation in course content. Finally, we show that differences in frontier proximity across

[1] An instructor's right to determine course content, teaching methods, and assessment strategies is grounded in principles of academic freedom, as articulated by the American Association of University Professors (AAUP) in its 1940 Statement of Principles on Academic Freedom and Tenure (of University Professors, 1940).

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

courses are driven primarily by instructors, with research-active faculty most likely to integrate cutting-edge material into their syllabi.

Our analysis builds on two large document corpora. The first corpus contains over two million syllabi: a 5% sample of all courses taught at 800 four-year US institutions between 1998 and 2018 (1.7M syllabi of over 500,000 courses) and about half of all courses taught at seven public universities in Texas between 2011 and 2022 (469,100 syllabi and 28,612 courses). This sample includes syllabi of courses from all fields and levels (undergraduate and graduate). The second corpus contains over 20 million academic articles published in top-tier journals since their inception.

We construct the frontier knowledge proximity of each course in a given year as the average textual similarity of a course's syllabus with recent research articles, relative to the average similarity with older articles. To calculate it, we first represent each document (a syllabus or an article) as a term frequency vector by projecting its text on a "dictionary," a comprehensive list of terms corresponding to knowledge items. Each vector element is the frequency of a given dictionary term within the document, divided by the length of the document. We additionally weight each vector element using the "term-frequency-backward-inverse-document-frequency" (*TFBIDF*) approach (Kelly et al., 2021), which increases the importance of terms that most distinctively capture a document's content. Second, we use these vectors to compute the cosine similarity between each syllabus and each article in our sample. Lastly, we construct the frontier knowledge proximity of a given syllabus as the *ratio* of its average similarities with (a) "frontier" knowledge, i.e., all articles published τ years prior to the syllabus's date and (b) "older" knowledge vintages, i.e., all articles published $\tau' > \tau$ years prior to the syllabus's date. To account for cross-field differences in the speed of knowledge production, we choose τ and τ' to match the temporal patterns of citations to academic articles in each field.

By construction, frontier knowledge proximity is higher for syllabi that cover more of the most recently disclosed knowledge, relative to older knowledge. For example, a 2020 Computer Science course that teaches *Julia* (a programming language developed in 2009) will have a higher proximity than a course that teaches *Visual Basic* (a relatively obsolete programming language). Because our metric is a ratio of similarities, idiosyncratic syllabus features (e.g., length, structure, or writing style) tend to cancel out. Moreover, the TFBIDF adjustment ensures that covering "classic" or "fundamental" topics does not penalize proximity, since widely taught foundational terms receive lower weight. Three empirical checks validate our measure: (i) syllabi that cite more recent articles and books exhibit higher proximity; (ii) proximity is lowest for lower-level undergraduate courses

and highest for graduate courses; and (iii) a simulation that replaces "older" content terms with "newer" ones increases a syllabus's proximity.

Applying this metric to our syllabi sample, we begin by addressing our first question: Do university courses teach frontier knowledge? We find that some do and some do not. Specifically, we uncover large differences in frontier knowledge coverage across courses: moving a syllabus from the 25th to the 75th percentile of the proximity distribution is equivalent to replacing roughly 79% of its content with newer knowledge. Importantly, at least 21% of the total variation in frontier knowledge proximity occurs within schools and across courses taught by different instructors, while differences across institutions, fields, and course levels account for a much smaller share. These patterns confirm that even students enrolled at the same institution may receive vastly different exposure to frontier content, depending on their course choices.

We next address our second question: Should universities teach frontier knowledge? Or in other words, does exposure to frontier knowledge help students succeed? To provide an answer, we link syllabi to individual-level administrative data on degree attainment and post-graduation earnings for students at seven Texas public universities. Using information from academic transcripts, listing all the courses a student ever enrolled in, we are able to assign to each student the exact proximity of the courses they took. Our research design exploits a feature of the Texas course-enrollment process: when registering for a course, students can view the previous term's syllabus but not the updated version. Consequently, a course's content in term t—conditional on its content in term t-1—is effectively unobserved and quasi-random from the student's perspective. We use this source of variation to identify the causal effects of frontier exposure on student outcomes.

We find that exposure to frontier knowledge in college has economically and statistically significant effects on educational attainment for the average student. A one-standard deviation (sd) increase in average course proximity increases the probability of graduating by 4%. The effect is lager for undergraduate students and for students in STEM and Business. The same increase raises major persistence by 3% and boosts the likelihood of pursuing graduate studies by 11%, especially among Business students. These effects appear to operate through heightened motivation and engagement: they are the largest for students with median ability and for transfer students from two-year colleges, groups with sufficient skill to master advanced concepts but a non-trivial risk of non-completion. Effects on graduate school attendance are also larger for students from lower-income families (particularly if high-ability), suggesting that frontier exposure can help close educational attainment gaps across socio-economic groups.

Besides improving educational attainment, exposure to frontier knowledge leads students to earn more. A one-sd increase in frontier knowledge proximity increases students' earnings by 3% on average in the short run (1–3 years post-graduation) and by 7% in the long run (4–6 years post-graduation). Although some of these earnings gains reflect improved educational outcomes, we observe that while low-income students realize the greatest gains in graduate-school attendance, high-income students capture the largest wage increases. This pattern implies a complementarity between individual resources and frontier knowledge in driving labor-market success. More generally, our results highlight the crucial role of frontier knowledge access in shaping multiple dimensions of student success.

In the final part of our analysis, we investigate what drives variation in frontier proximity across courses. A variance decomposition reveals that instructors account for most of the differences in frontier coverage. Event-study estimates around instructor changes confirm that, although course proximity is generally stable over time, it rises sharply when a new instructor takes over. However, not all instructors are the same. Linking syllabi to data on instructors' publications, citations, and grants shows that faculty who are more research-active (i.e., those with higher publication counts, citation rates, and grant awards) teach more frontier knowledge in their courses. This finding implies that research and teaching complement each other. Additionally, instructors whose research topics align with course content tend to incorporate more frontier knowledge. These effects are most pronounced for graduate-level courses, consistent with a model in which the cost of keeping a course up to date depends on the instructor's familiarity with the research frontier.

In sum, our results document substantial heterogeneity in the presence of frontier knowledge across HEI courses and show that these differences have meaningful consequences for student outcomes. Moreover, our findings suggest that strategically deploying faculty across courses can bring curricula closer to the knowledge frontier. As part of this work, we develop and validate a novel text-based method to quantify frontier knowledge coverage in higher education. We aim to make our metric, the underlying algorithm, and all code publicly available so that future research can build on this methodology to analyze educational content in HEIs.

Related literature This paper contributes to several strands of the literature. The closest has examined heterogeneity in the production of human capital, primarily emphasizing differences across majors (Altonji et al., 2012; Deming and Noray, 2020), institutional selectivity (Hoxby, 1998; Dale and Krueger, 2014; Mountjoy and Hickman, 2021), and the skill content of college majors (Hemelt et al., 2021; Li et al., 2021). We contribute to this literature in multiple ways. First, we focus on

differences in the teaching of frontier knowledge, a relevant yet unexplored dimension of human capital production. Second, by using syllabi information, we examine the content of each individual course, rather than majors or schools. Lastly, we propose a novel empirical strategy to estimate the causal impact of each course on student outcomes.

Our work also relates to studies of the causal effects of colleges on students, pioneered by Dale and Krueger (2002). More recent contributions (e.g., Cunha and Miller, 2014; Hoxby and Bulman, 2015; Mountjoy and Hickman, 2021) employ a "school value-added" framework to quantify the causal impact of institution on outcomes such as earnings. While valuable to rank schools according to the labor market returns they generate, value-added estimates are difficult to use for policy. First, they are a "black box" and do do not contain any information on which dimensions of instruction generate positive returns. Second, they are by definition backward-looking: they only become available several years after the instruction has taken place. By linking students' outcomes to the course content they experience, our approach allows us to identify the features of a program that help students the most, a piece of information that can be used for policy without delays.

Additionally, we provide direct evidence on the role of instructors in shaping educational content. Prior work has documented important effects of instructors on student outcomes (Hoffman and Oreopoulos, 2009; Carrell and West, 2010; Braga et al., 2016; Feld et al., 2020), but less is known about the mechanisms by which instructors influence learning (De Vlieger et al., 2020). We show that differences in instructors' research activity, measured by publications, citations, and grants, translate into measurable variation in the amount of frontier knowledge taught. Critically, this finding underscores complementarities between teaching and research, activities that have often been framed as conflicting tasks for faculty (Becker and Kennedy, 2005; Arnold, 2008; Hattie and Marsh, 1996; Courant and Turner, 2020).

Lastly, our results speak to the broader literature on how access to existing frontier knowledge fosters the creation of new ideas and innovation (Moser and Voena, 2012; Williams, 2013; Galasso and Schankerman, 2015; Iaria et al., 2018; Biasi and Moser, 2021).[3] Educational institutions are frequently cited for their critical role in disseminating frontier knowledge, especially in STEM fields (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorcelli, 2019; Akcigit et al., 2020). We extend this work by identifying where, within the higher-education system, students gain exposure to frontier knowledge and by quantifying how that exposure affects students' persistence, graduate-

[3] 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.

school entry, and labor-market outcomes across different socioeconomic and ability groups.

2 Data

Our empirical analysis combines several distinct data sources: course syllabi, academic publications, individual-level student demographics, academic records, and earnings, and detailed instructor characteristics. We provide additional details on the construction of the final data set in Appendix B.

2.1 Course Syllabi

We compiled our syllabi database, containing 2.2 million course syllabi, from two distinct sources. The first source covers the majority of courses taught at seven major public universities in Texas: Stephen F. Austin State University (starting from 2009), Sam Houston State University (2011), Texas A&M University (2013), University of Houston-Clear Lake (2010–2022), University of Texas at Austin (2011), University of Texas at Dallas (2005), and West Texas A&M University (2013). We collected these syllabi directly from each university's website, for a total of 469,100 documents corresponding to 28,612 courses taught up to 2022.[4] These represent approximately 52% of all courses offered at these institutions during our analysis period.

The second source of syllabi is the Open Syllabus Project (OSP), whose records contain 1.7 million U.S. syllabi gathered from publicly available university and faculty webpages. These syllabi span 542,251 courses taught at 767 U.S. institutions from 1998 to 2018. We focus on four-year U.S. institutions with at least 100 syllabi, excluding primarily online universities, and removing documents with fewer than 20 or more than 10,000 words, we retained a sample covering about 5% of all courses at these institutions (see Appendix B.1). Using the Texas dataset allows us to link course content directly to student outcomes, while the OSP dataset enables broader comparisons across diverse institutions.

Most syllabi follow a common structure: they begin with basic course details (code, title, instructor), followed by a description of the course's content, detailed topic outlines, required readings, and evaluation criteria (assignments and exams, and general course policies).

Basic course details We extract course codes, titles, academic terms, years, and instructor names. Using course codes and titles, we categorize each syllabus as lower undergraduate, upper under-

[4] We contacted all public universities in Texas that do not make historical syllabi available online to request access to their records. However, most universities were unable to provide these documents because Texas House Bill 2504 of 2009 only requires public colleges and universities to maintain records for two years following each syllabus' term of instruction.

graduate, or graduate (see Appendix B.1.4). Texas courses are assigned to a field following the National Center for Education Statistics' Classification of Instructional Programs (CIP). OSP syllabi are classified into one of 69 fields, also largely based on CIP. In case of discrepancies, we harmonize the two classifications using a large-language model.[5] For most analyses, we aggregate fields into four broad macro-fields: STEM, Humanities, Social Sciences, and Business (Appendix Table BXI).

Course content We identify the section of a syllabus that provides a description of the course's content by searching for headings such as "Summary," "Description," and "Content."[6] These sections typically include course structure, main concepts, timelines, and reading materials.

Reference list For each syllabus, OSP provides bibliographic details of required and recommended readings. We augment this with in-text citations extracted from syllabi. We successfully compile reference lists for 71% of OSP syllabi and 41% of Texas syllabi. We further expand each reference with bibliographic information from Elsevier's SCOPUS database (title, abstract, journal, keywords, and textbook editions).

Sample description Our syllabi data are described in Table 1, panel (a). In the Texas data, a syllabus contains an average of 197 unique knowledge words—words that belong to a dictionary compiled to capture a document's academic content, defined in greater detail in Section 3—with a standard deviation of 201. The average OSP syllabus contains 420 unique knowledge words. Most Texas syllabi are from STEM fields and upper undergraduate courses, whereas OSP syllabi predominantly belong to STEM and lower undergraduate levels.

2.2 Academic Publications

We construct a database of peer-reviewed articles from Elsevier's SCOPUS, containing all papers published in top journals since their inception. Top journals are defined as those ranking in the top 10 by Impact Factor in SCOPUS's 191 fields at any point since either 1975 or the journal's foundation. Our final dataset includes approximately 20 million articles (around 100,000 per year), with detailed information on titles, abstracts, keywords, authors, and author affiliations.

2.3 Students

A key empirical step for our analysis is to link individual students to the courses they took in college. Individual-level data on enrolled students are provided by the Texas Education Research

[5]The field taxonomy used by OSP draws extensively from the 2010 CIP, available at https://nces.ed.gov/ipeds/cipcode/default.aspx?y=55. We applied the OSP taxonomy to the Texas syllabi using a large-language model, ChatGPT 4.0, and information on the course prefix, title, and first 100 words of the syllabus. Appendix Table BXI lists all OSP and Texas fields and shows the correspondence between fields and macro-fields.

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

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

	Texas sample		OSP sample	
	mean	sd	mean	sd
Panel (a): Syllabi				
Frontier knowledge proximity	105.38	7.54	107.00	19.64
# unique knowledge words	197	201	420	327
Macro-field:				
STEM	0.43	0.50	0.33	0.47
Business	0.07	0.26	0.10	0.30
Humanities	0.26	0.44	0.30	0.46
Social Sciences	0.19	0.39	0.24	0.43
Course level:				
Lower Undergraduate	0.40	0.49	0.39	0.49
Upper Undergraduate	0.48	0.50	0.28	0.45
Graduate	0.12	0.32	0.33	0.47
N syllabi	469,100		1,706,243	
N courses	28,612		542,251	
Panel (b): Students				
Black	0.09	0.28		
Parental income below $20K	0.06	0.24		
Parental income above $80K	0.30	0.46		
Share w/SAT/ACT score	0.64	0.48		
# courses in transcript	20.83	15.35		
Ever graduates	0.77	0.42		
Ever enrolled in graduate school	0.16	0.36		
W/any observed earnings 1-6 years post-grad	0.64	0.48		
Avg quarterly earnings, 1-3 years post-grad ($)	14,990	13,199		
Avg quarterly earnings, 4-6 years post-grad ($)	19,672	23,566		
N students	508,041			
Panel (c): Instructors				
At least 1 publication	0.35	0.48	0.41	0.49
# publications, past 5 yrs	11.02	22.09	6.01	14.89
# citations, past 5 yrs	336.43	1036.54	172.46	887.99
At least 1 grant	0.07	0.26	0.18	0.38
# courses taught (overall)	3.88	4.25	2.31	3.26
N instructors	27,077		332,064	

Note: Summary statistics of the variables used in the analysis. The students' sample is restricted to individuals we can link to at least one syllabus using academic transcripts.

Center (ERC). We use two datasets. Data from the Texas Higher Education Coordinating Board (THECB) provides demographic data (gender, race, parental income), high-school and SAT/ACT scores, university enrollment and degree completion records (starting year, type of degree, initially declared major, graduation outcome, graduation year, and final major), and full academic tran-

scripts. Data from the Texas Workforce Commission (TWC) provides quarterly earnings data for all employed individuals in Texas.

Students' academic transcripts list all the courses each student ever enrolled in, regardless of whether they were completed. We use transcripts to link students to the syllabi of their courses, using course codes and terms (we ignore information on sections since all sections of a course typically share the same syllabus). We successfully link 52% of all courses listed in student transcripts to at least one syllabus.

Our merged student sample covers 508,566 students enrolled at the seven universities with linked syllabi. Among these, 53% are female, 9% are Black, 6% have family income below $20K, and 30% above $80K. On average, students enroll in 21 courses; 77% of them graduate, and 16% pursue graduate studies. Earnings are observed for 64% of students within 1–6 years post-graduation (further details in Appendix D).

2.4 Instructors

Almost all syllabi list the name of the instructor. Our Texas sample includes 27,077 instructors, each teaching 3.9 courses on average, while the OSP sample has 332,064 instructors teaching 2.3 courses each. Instructor reassignment is frequent: in the Texas sample 96% of all instructors change courses at least once, and 69% of all courses have at least one instructor change during our analysis period.

Using fuzzy matching on names and affiliations, we link instructors to publication records from Microsoft Academic Graph (MAG), a platform listing each researcher's publications and citations (details on the matching procedure are in Appendix B).[7] We successfully match 35% of Texas instructors and 41% of OSP instructors. We assume that unmatched instructors never published any article (Table 1, panel (c)).[8] We use counts of publications and citations in the previous five years to measure the quantity and quality of each instructor's research output.[9] Texas instructors averaged 11 publications and 336 citations over the previous five years, compared to 6 publications and 172 citations for OSP instructors (Table 1, panel (c)). In both samples, the median instructor published only one article and received no citations.

We complement data on publications with information on grants received by each instructor from two among the largest government funding agencies: the National Science Foundation (NSF)

[7] Microsoft Academic was discontinued on December 31, 2021.

[8] This assumption is supported by manual random searches.

[9] Using publications in the previous five years, rather than total publications or publications per year, helps account for the life cycle of publications. For example, older instructors may have a higher number of total publications or publications per year even if their productivity has declined over time.

and the National Institute of Health (NIH).[10] We link grants to instructors via fuzzy matching on investigators' names and affiliation (more details can be found in Appendix B). In total, 7% of Texas instructors and 18% of OSP instructors are linked to at least one grant.

3 Measuring Proximity to the Knowledge Frontier

We now outline how we use the text of syllabi and academic publications to construct our measure of frontier knowledge proximity and present a series of checks to validate it. Additional details on the construction of the measure are in Appendix C.

3.1 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 c_{dw} is the number of times term w appears in d and the denominator is the total number of terms in d. To focus specifically on academic knowledge content, we restrict these vectors to terms contained in a comprehensive dictionary W with $|W|$ terms (as a result, each term vector contains $|W|$ elements). Our primary dictionary is the list of all unique terms used as keywords in our academic publications sample.[11]

3.1.2 Adjusting for Term Relevance

To effectively capture the knowledge content of each document, we want to assign higher weights to terms that better represent a document's content. Since unadjusted **TF** vectors measure a term's frequency within the document, frequent terms across *all* documents mechanically receive more weight regardless of their ability to capture the document's content. For example, terms such as "Programming" or "Animals"—very common among Computer Science and Biology syllabi,

[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. Our 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).

[11] 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.

respectively—receive a higher weight but are usually less informative of content than terms such as "Natural Language Processing" or "CRISPR."[12]

To address this issue, we use the term-frequency-inverse-document-frequency (TFIDF) weighting (Kelly et al., 2021), which assigns each term a weight inversely proportional to its frequency across all documents. Specifically, we calculate the inverse-document-frequency vector **IDF**, 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 syllabi and articles, and the denominator counts all documents containing term w. IDF_w is thus the inverse of the share of all documents containing word w. The term-frequency-inverse-document-frequency vector $\mathbf{TFIDF_d}$ thus has elements:

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

Accounting for changes in term relevance over time Traditional TFIDF weighting ignores temporal changes in term relevance, which is problematic for identifying novel content. Terms that became popular only recently might wrongly receive low weights if their current widespread use is considered. 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.

To address this issue, we define a backward-looking IDF measure, $\mathbf{BIDF_t}$, which calculates the inverse-document-frequency based only on documents published prior to year t:

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 $\mathbf{BIDF_t}$, with elements:

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

where D_t is the set of documents published prior to t. Using $\mathbf{BIDF_t}$, we construct a term-frequency-

[12]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.

backward-inverse-document-frequency vector $\mathbf{TFBIDF_d}$, with elements:

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

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

3.1.3 Building Textual Similarities Between Syllabi and Articles

Equipped with weighted vectors $\mathbf{TFBIDF_d}$, we calculate textual similarities between pairs of documents d and d' using the cosine similarity (for simplicity, we denote $\mathbf{TFBIDF_d}$ as $\mathbf{V_d}$):

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

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 Proximity to the Knowledge Frontier

To capture the similarity between each syllabus d and different vintages of research, we calculate 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 $\Omega_\tau(d)$ includes all articles published in the period $[t(d) - \tau - 1, t(d) - \tau + 1]$, and $|\Omega_\tau(d)|$ is the number of these articles.[13]

Our frontier knowledge proximity measure is defined as the ratio between the average syllabus similarity with recent articles (published in the interval $[t(d) - \tau - 1, t(d) - \tau + 1]$) and older articles (published in $[t(d) - \tau' - 1, t(d) - \tau' + 1]$, with $\tau' > \tau$), multiplied by 100 for readability.:

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

This definition implies that a syllabus taught in t has a higher proximity if it is more similar to recent research (published in $[t(d) - \tau - 1, t(d) - \tau + 1]$) than to older research (published in $[t(d) - \tau' - 1, t(d) - \tau' + 1]$).

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

In setting the parameters τ and τ', which define new and old knowledge vintages, we want to account for field-specific rates of knowledge production. In fields where knowledge production is fast, τ and τ' are both small; in fields where it is slower, τ' (and possibly τ) will be large. To account for this feature, we consider the distribution of citation age within each field and select τ and τ' to be the 5th and 90th percentiles of citation ages across all the articles in that field. The median recent knowledge vintage (5th percentile) is approximately 2 years, while older vintages (90th percentile) have a median of 16 years, ranging from 12 years in fast-moving fields like Medicine to 36 years in slower-moving fields such as Religion (Appendix Figure A1).

Our measure has two attractive properties. First, being a ratio, it is not affected by stylistic differences across syllabi (such as length or verbosity) that could otherwise affect similarity scores. For example, two courses covering the same materials could have different similarities to research publications if one syllabus is more detailed or uses more academic terms. Second, by employing the $TFBIDF$ weighting scheme, it appropriately reduces the influence of common foundational terms, ensuring that syllabi covering widely taught, fundamental concepts (i.e., the "classics" of a field, such as *Ordinary Least Squares* in applied microeconomics courses) are not unfairly penalized.

3.3 Validating The Measure and Interpreting Its Magnitude

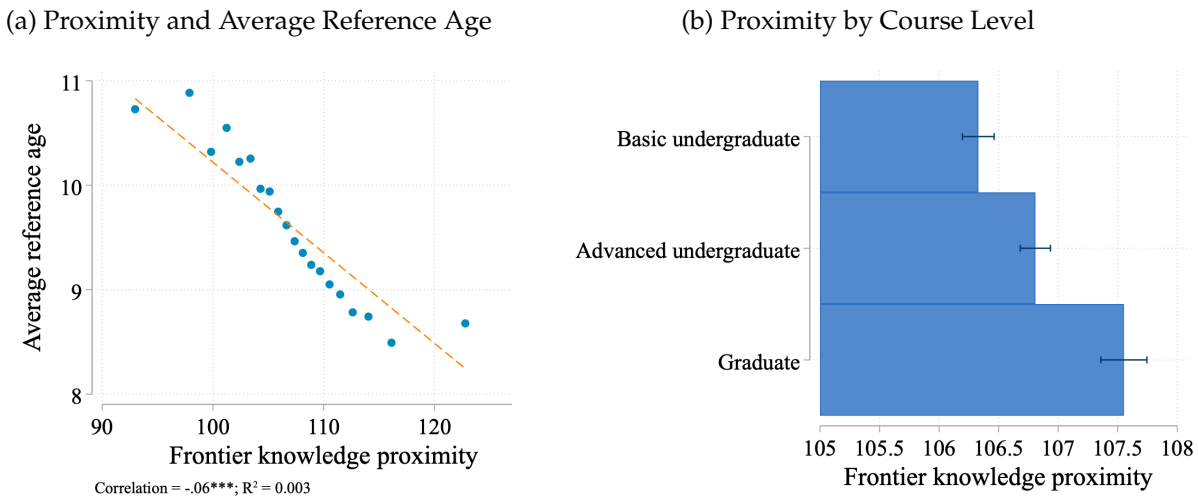
We validate our measure's ability to capture the distance between course content and the research frontier using three tests.

First, we verify that frontier proximity correlates negatively with the average age of a syllabus's references, calculated as the average difference between the syllabus year and the publication years of its cited sources (Figure 1, panel (a)). The correlation, however, is modest in magnitude (-0.06). This weak relationship is unsurprising, as many syllabi primarily reference textbooks whose publication dates might not accurately reflect course content. Thus, while reference age is simple to compute, our text-based measure better captures actual course content and is available even for syllabi lacking explicit references or relying primarily on a single textbook.

Second, we confirm that proximity varies as expected across course levels. Graduate courses and upper undergraduate courses have higher proximity scores compared to lower undergraduate courses (Figure 1, panel (b)), consistent with the intuition that more advanced courses incorporate more frontier knowledge.

Third, we use a simulation exercise to demonstrate that our measure can detect incremental changes in a syllabus's coverage of knowledge vintages. We begin with a random subsample of 100,000 syllabi from the combined Texas and OSP samples. In each of these, we progressively

Figure 1: Validating the Frontier Knowledge Proximity



Note: Panel (a) shows a binned scatterplot of the proximity to the knowledge frontier 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 proximity by course level, controlling for field-by-year effects. We combine data from the Texas and OSP samples.

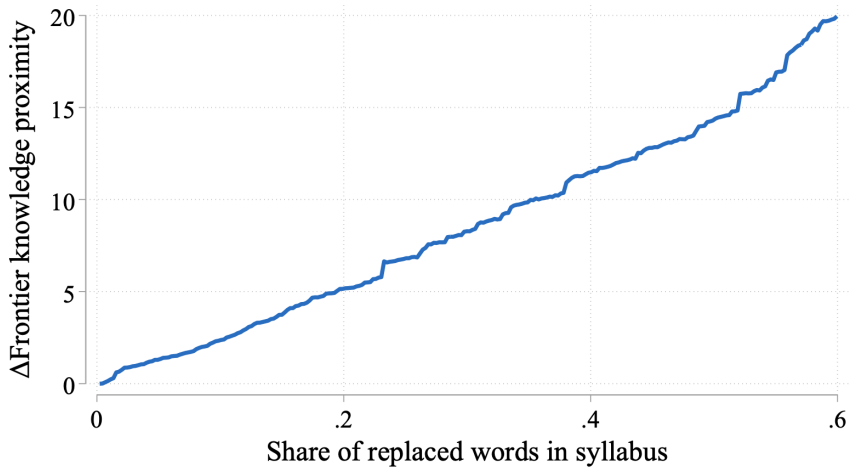
replace terms more frequently appearing in older articles ("old terms") with those more frequent in recent articles ("new terms"). We then recalculate proximity at each incremental replacement. Old terms are defined as either: (a) within the top 5% frequency among older articles (published between $t - \tau' - 1$ and $t - \tau' + 1$), or (b) appearing in the old corpus but not in the recent corpus (published between $t - \tau - 1$ and $t - \tau + 1$), where t denotes the syllabus year, and τ and τ' are field-specific as defined in Section 3.2. New words are defined in a symmetric way, as either (a) in the top 5% in terms of frequency in the new publication corpus, or (b) in the new publication corpus but not in the old publication corpus.

Proximity increases monotonically as we replace more old words with new ones. This is shown in Figure 2, which plots the median change in proximity across all syllabi for a given number of replaced words. A 10-unit increase in frontier knowledge proximity is equivalent to replacing about 35% of the old content of the median syllabus with new content.

4 Differences in Frontier Knowledge Proximity Across Courses

We begin our analysis by examining the variation in frontier knowledge proximity across courses in our samples. Figure 3 (blue bars) shows histograms of the proximity measure in the Texas and OSP samples. In the Texas sample, the average proximity is 105.2, with a standard deviation of 7.6,

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



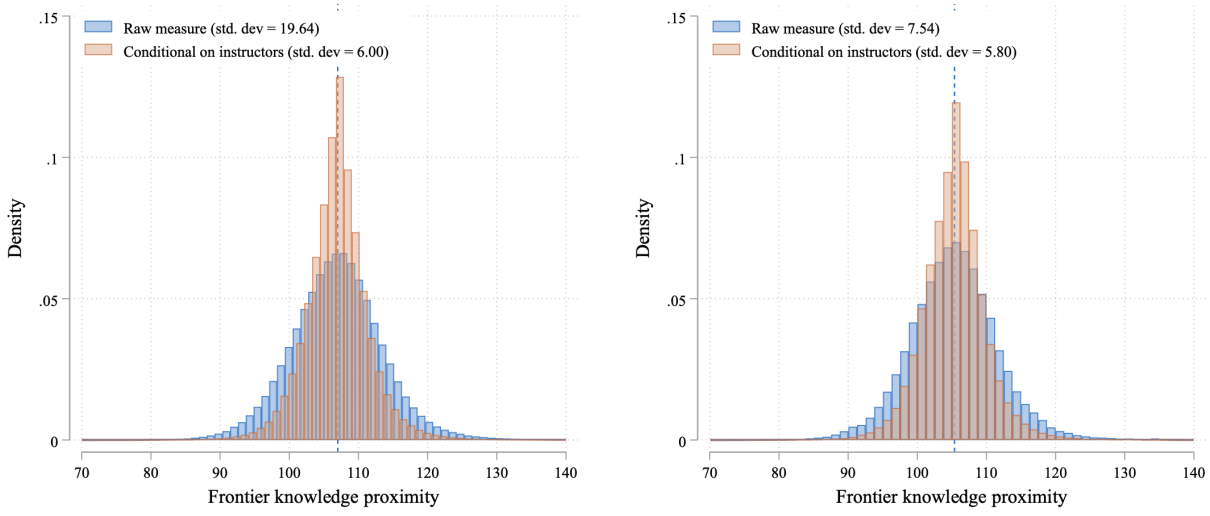
Note: This figure illustrates the median change in proximity from a simulation exercise on a random subsample of 100,000 syllabi from the combined Texas and OSP samples, in which we progressively replace "old" knowledge words with "new" knowledge words.

a 25th percentile of 101.2, and a 75th percentile of 109.1. In the OSP sample, the average proximity is 106.9, with a standard deviation of 19.6, a 25th percentile of 102.7, and a 75th percentile of 111.0.

Figure 3: Distribution of Frontier Knowledge Proximity in the Texas and OSP Samples

(a) Texas sample

(b) OSP sample



Note: Distribution of frontier knowledge proximity measures in the Texas and OSP samples.

To give these numbers an economic interpretation, we rely on the simulation in Figure 2. A one–standard-deviation (sd) increase in proximity corresponds to replacing 27% of all knowledge

terms in the median Texas syllabus. Moving a Texas syllabus from the 25th to the 75th percentile of proximity (a 7.9 increase) corresponds to replacing approximately 30% of the median syllabus content. These findings illustrate substantial variation in frontier knowledge proximity across courses.

4.1 Explaining the Variation in Frontier Knowledge Proximity

Several factors could explain this variation. Fields may differ in their emphasis on foundational versus frontier knowledge; for example, mathematics courses might emphasize more foundational content compared to computer science courses. Institutional characteristics, such as teaching philosophy, resources, or faculty expertise, might also matter; for instance, less-resourced institutions may face challenges integrating cutting-edge material. Differences between instructors, as well as variations over time, might further influence proximity.

To quantify the contribution of each of these factors to explaining the variation, we perform a Shapley-Owen decomposition of variance (Israeli, 2007; Huettner et al., 2012). Specifically, we estimate the adjusted R^2 of a regression of proximity on fixed effects for fields, schools, years, instructors, and courses. We then compute each factor's partial contribution (partial-R^2) by measuring the average decline in adjusted R^2 when removing that factor from the regression.

We implement this decomposition in three steps. First, we estimate OLS models of proximity on all possible combinations of factors and record their adjusted R^2.[14] Second, we re-estimate these models excluding one factor at a time and record the decrease in adjusted R^2. Third, we calculate the average decline in adjusted R^2 for each factor across all possible combinations, defining this average as the partial-R^2 attributable to each factor. Formally, for factor j, the partial-R^2 is calculated as:

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

where $R^2(S)$ is the adjusted R^2 of a regression of proximity on the set of factors S, V is the full set of factors, $|T|$ denotes the number of factors in subset T, and $K \equiv |V| = 5$ is the total number of factors. Using adjusted R^2 ensures comparability across factors with different numbers of categories.[15]

[14] 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 frontier knowledge proximity out of the regressions of those combinations of the five attributes that do not include course effects.

[15] To confirm the robustness of this method and 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, we perform a placebo test, randomly scrambling course codes across syllabi. This test ensures that explanatory power is not artificially inflated by the large number of course identifiers. Scrambled course codes explain less than 1% of total variation, confirming the validity of our decomposition.

Table 2: Decomposing the Variation in the Frontier Knowledge Proximity: Schools, Years, Fields, Courses, and Instructors

	Partial R^2	
	Texas sample (1)	OSP sample (2)
School	0.003	0.003
Field	0.046	0.007
Year	0.079	0.042
Course	0.258	0.301
Instructor	0.21	0.387
All	0.597	0.740

Note: This table shows a Shapley-Owen decomposition of the adjusted R^2 of a regression of proximity on fixed effects for schools, years, fields, instructors, and courses into the contribution of each set of fixed effects. *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. Column 1 uses data from the Texas sample and column 2 uses data from the OSP sample.

The results of this decomposition exercise, shown in Table 2, indicate that differences across instructors explain the largest portion of the overall variation in frontier knowledge proximity, 21% in the Texas sample (column 1) and 39% in the OSP sample (column 2). Differences across courses also explain a substantial portion (26% in the Texas sample and 30% in the OSP sample), indicating a significant amount of persistence in content within courses over time. By contrast, field effects explain only 5% (Texas) and 0.7% (OSP), and school effects contribute just 0.3% in both samples.

5 The Effects of Frontier Knowledge on Students: Research Design

Do differences in proximity to frontier knowledge in a student's education have an impact on their educational and labor-market outcomes? To estimate these causal effects, we use data on the Texas syllabi sample, taking advantage of its linkage to individual student records. A key challenge in this empirical estimation is the endogeneity of course selection at the student level. In this section, we present our core empirical specification, identification strategy, and estimation. In the next section, we present results on student outcomes, including graduation rates, major persistence, graduate-school enrollment, and post-graduation earnings.

5.1 Empirical Model

Our goal is to estimate the following model:

$$y_i = \beta p_i + \gamma X_i + \theta_{s(i)m(i)c(i)} + \epsilon_i, \tag{5}$$

where y_i is student i's outcome (e.g., average earnings 1-3 years post graduation) and p_i is the average frontier knowledge proximity of all courses on i's transcript. The vector X_i includes individual attributes that may influence outcomes, such as gender, race, family income, and high school academic achievement. The vector θ_{smc} contains school s-by-major m-by-starting cohort c fixed effects. The inclusion of these fixed effects implies that we compare students in the same program and entry year.

Our parameter of interest is β, which captures the causal effect of average course proximity p_i on outcomes. A key challenge in estimating β is that p_i depends on the student's course selection, which could be correlated with unobserved traits that also affect outcomes. For example, more motivated students may be more likely to enroll in courses with a higher proximity and work harder to achieve higher earnings post-graduation. This implies that $\mathbb{E}(p_i \epsilon_i) \neq 0$, leading to bias in OLS estimates.

5.2 Identification Strategy

To address the identification challenge, we exploit an institutional feature that generates variations in p_i that are unobservable to the student at the time of course selection. In the schools in our sample, when registering in a course, students cannot see the latest updated syllabus, but only the one from the prior term (see Appendix Table A1 for details). As a result, changes in course content (Δp_i) are unobserved and therefore quasi-random from the student's perspective.

To illustrate this argument, suppose a student i chooses between courses A and B at time t, with proximity $p_{A,t}$ and $p_{B,t}$. The student's experienced proximity, p_i, equals $p_{A,t}$ if the student chooses A and $p_{B,t}$ if she chooses B. Define the change in proximity for course c as $\Delta_{c,t} \equiv p_{c,t} - p_{c,t-1} \forall c \in \{A, B\}$. We can express the realized proximity as

$$p_i = g(\overbrace{\alpha_i, \nu_i, \underbrace{p_{A,t-1}, p_{B,t-1}}_{\text{course content}}}^{\text{course selection}}, \Delta_{A,t}, \Delta_{B,t}),$$

where α_i represents the student's (unobserved) preference for high-proximity courses and ν_i in-

cludes other unobserved factors. In words, students' choice of courses depends only on $\{\alpha_i, \nu_i, p_{A,t-1}, p_{B,t-1}\}$, whereas actual course content is instead determined by $\{p_{A,t-1}, p_{B,t-1}, \Delta_A, \Delta_B\}$. Endogeneity arises if ν_i and α_i are correlated with ϵ_i.

Our identification argument relies on the *conditional* independence of p_i and ϵ_i given $\{p_{A,t-1}, p_{B,t-1}\}$. Since students cannot observe $\{\Delta_{A,t}, \Delta_{B,t}\}$, their choice only depends on $\{p_{A,t-1}, p_{B,t-1}\}$. Hence, conditional on these lagged values, any variation in actual proximity p_i is independent from α_i and ν_i and thus exogenous. This argument leads us to augment the model in equation (5) with controls for $\{p_{A,t-1}, p_{B,t-1}\}$:

$$y_i = \beta p_i + \gamma X_i + \delta_A p_{A,t-1} + \delta_B p_{B,t-1} + \theta_{s(i)m(i)c(i)} + \epsilon_i.$$

5.2.1 Implementation

We now map the illustrative identification strategy to the real sample. In our sample, students choose multiple courses from a large set of potential courses; the average student in our data reports 21 courses on her transcript (Table 1). Our identification argument relies on the conditional independence of p_i and ϵ_i given the lagged proximity of all courses in i's choice set. Directly including the lagged proximity of all courses in a student's choice set as controls in our estimating equation is infeasible (on average, $\mathcal{C}_i$ contains over 1,000 courses).[16] Instead, we summarize this information with two aggregate measures: $p_{i,-1}$, the average lagged proximity of courses that the student takes, and $p_{-i,-1}$, the average lagged proximity of courses in $\mathcal{C}_i$ that the student never takes. Incorporating these aggregates, our estimating model becomes:

$$y_i = \beta p_i + \gamma X_i + \delta_1 p_{i,-1} + \delta_2 p_{-i,-1} + \theta_{s(i)m(i)c(i)} + \epsilon_i, \tag{6}$$

Robustness checks indicate that our results do not depend on the specific way we summarize the lagged proximities (Appendix Tables A5 and A6).

Our framework assumes that the decision to take course k at time t, captured by d_{ikt}, is made independently for each course and term. This assumption rules out the possibility that students strategically coordinate a sequence of courses. While in practice, course choices may be interdependent (e.g., due to prerequisites), this simplification aids in the estimation and interpretation of our estimates.

[16] On average, programs (defined as major-institution-degree combinations) offer 215 courses during our time period, with a standard deviation of 579 and a median of 37.

5.2.2 Identification assumptions

Our identification strategy rests on two main assumptions. The first is the exogeneity of $\Delta_{k,t}$, which requires that students cannot predict syllabi updates when choosing courses. Even if students cannot see updated course syllabi at the time of enrollment (a possibility we rule out in Appendix A1), this assumption could be violated if they can predict content changes via observable changes in other course attributes. The most obvious is a change in instructor, shown in Section 7 to be linked to a significant change in course proximity. To address this issue, we explicitly control for instructors by regressing our proximity measure on instructor fixed effects and using the standardized residuals as the explanatory variable in our models.[17]. We additionally control for changes in course content that are common across courses in a given program (e.g., those driven by the directives of a new department chair) by including school-by-major-by-cohort fixed effects in all our specifications. The assumption could also be violated if students are able to drop courses after having observed $\Delta_{k,t}$. To avoid this, we assign courses to students based on enrollment status on the first day of the term, regardless of whether the course was completed.[18]

The second identifying assumption is course design independence, requiring that instructors design their courses regardless of the characteristics of future students (so that both $p_{k,t-1}$ and $\Delta_{k,t}$ are orthogonal to ϵ_i). This assumption is plausible given that syllabi are generally compiled before instructors meet their students. In support of it, Appendix Table A2 shows that it is not possible to predict $\Delta_{k,t}$ with observable characteristics of all students who take k in t, considered one at a time or jointly: The p-value of an F-test of joint significance is 0.46.

6 The Effects of Frontier Knowledge on Students: Results

Frontier-knowledge proximity may affect student outcomes through multiple channels. It may make coursework more engaging and thus boost the odds of graduating and persistence in the major. By exposing students to the latest theories, methods, and technologies in a field, it can spark interest in graduate study. It could also foster critical-thinking and problem-solving skills that enhance productivity and earnings. We now investigate these effects.

Our baseline models use a proximity measure net of instructor fixed effects, standardized to have mean zero and unit variance; a one-standard-deviation change in this measure, equal to 5.7, is equivalent to a 63% change in the content of the median syllabus. We control for school-major-

[17]The distribution of these residuals, rescaled to have the same mean as the raw measure, is shown in Figure 3. Estimates obtained using the raw proximity measure as the explanatory variable are shown in Appendix A3 and A4.

[18]Our estimates of β can thus be interpreted as the "intent-to-treat" (ITT) effect of course proximity on outcomes.

starting cohort fixed effects and by gender, race, family income categories (below $20K, $20K-$40K, $40K-$80K, and above $80K), and quintiles of SAT/ACT scores. Unless noted otherwise, we account for course selection with two aggregates: the student's average lagged proximity of courses taken and of courses not taken.

Table 3: Frontier Knowledge Proximity and Students' Educational Attainment

	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Graduation	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.008***	0.028***	0.043***	0.011***	0.028***	0.028***
	(0.002)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Mean of Dep. Var.	0.77	0.77	0.77	0.80	0.75	0.77
Adj. R^2	0.16	0.16	0.16	0.18	0.15	0.17
N	458,773	458,773	396,098	62,675	241,738	214,907
Panel (b): Major persistence	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.019***	0.016***	0.024***	0.006***	0.021***	0.014***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
Mean of Dep. Var.	0.56	0.56	0.53	0.76	0.56	0.56
Adj. R^2	0.39	0.39	0.38	0.36	0.43	0.43
N	355,810	355,810	306,687	49,123	192,927	160,766
Panel (c): Graduate school attendance	Undergraduate				Females	Males
Proximity (sd)	-0.004**	0.016***			0.014***	0.017***
	(0.002)	(0.003)			(0.004)	(0.005)
Mean of Dep. Var.	0.15	0.15			0.17	0.13
Adj. R^2	0.10	0.10			0.10	0.09
N	396,098	396,098			212,132	182,512
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students who graduate with the same major initially declared upon enrollment (panel (b)) and an indicator for undergraduate students attending graduate school within Texas (panel (c)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, residualized using instructor fixed effects and calculated as the average across all courses in the student's transcript for which we observe a syllabus. In column 3 of panels (a) and (b) and in panel (c), the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. In panel (b), the sample is restricted to students who graduate. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below $20K, between $20-40K, between $40-60K, between $60-80K, and above $80K. Columns 2-6 additionally control for the lagged average proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

6.1 Graduation

Exposure to frontier knowledge significantly raises the likelihood of program completion. A one-sd increase in proximity increases the likelihood of graduating by 2.8 pp, or 4% relative to a mean probability of 0.77 (Table 3, panel (a), column 2, significant at 1%). Omitting lagged proximity controls yields a much smaller estimate (0.8 pp), suggesting that endogenous course selection would bias our effect towards zero. This could occur, for instance, if the students most likely to graduate select "easier" (lower-proximity) courses (column 1).

The effect of course proximity is larger for undergraduates (4.3 pp or 6%, Table 3, panel (a), column 3) than for graduate students (1.1 pp or 1%, column 4). This pattern may reflect a stronger motivational boost from frontier knowledge exposure among younger students. Effects are comparable across genders (columns 1 and 2) but differ across macro-fields: they are largest for STEM (3.8 pp or 5%), followed by Business (2.9 pp or 4%), Humanities (1.7 pp or 2%), and Social Sciences (1.3 pp or 2%, Appendix Figure A2, panel (a)).

6.1.1 Differences by Ability

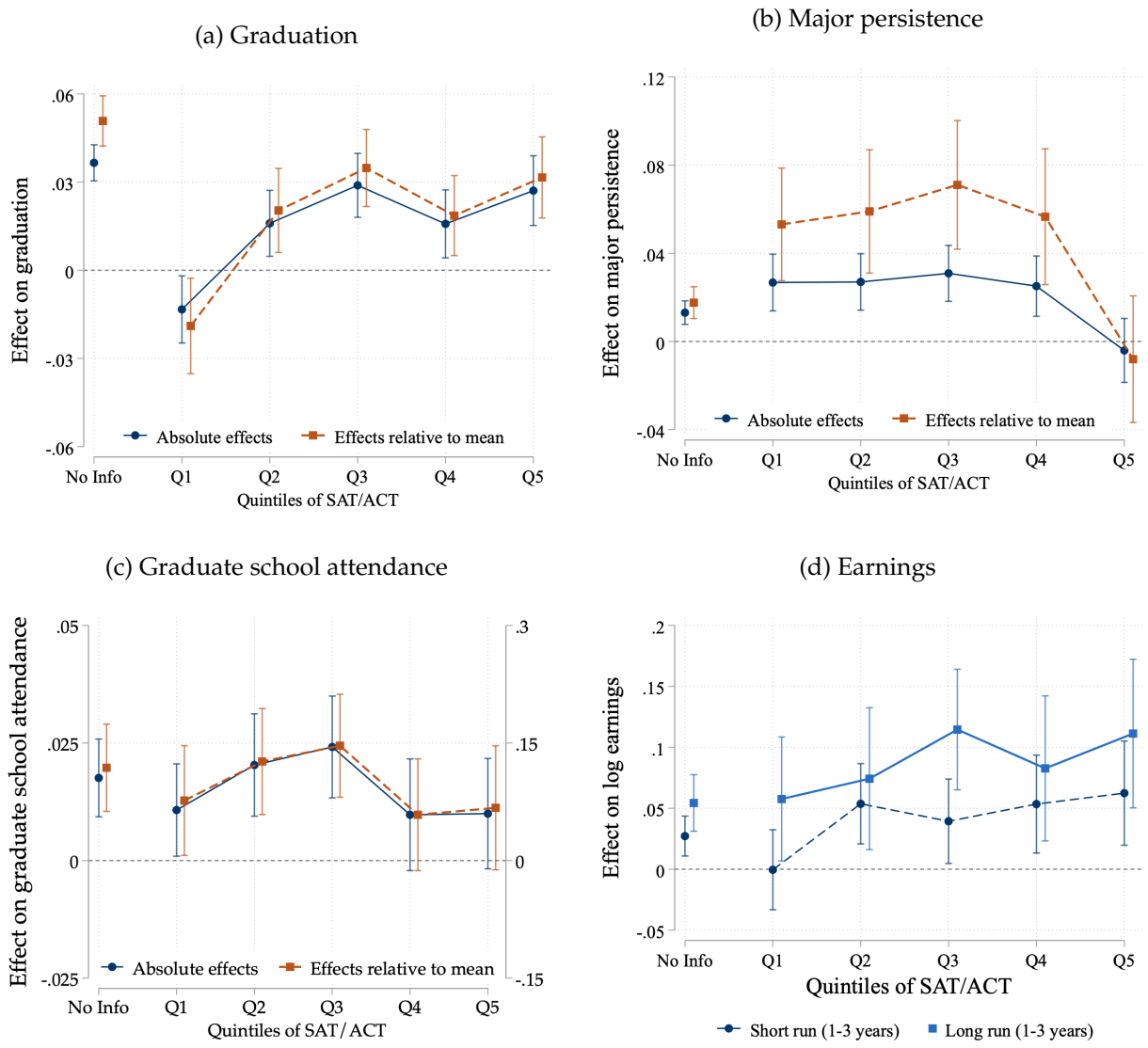
The impact of frontier knowledge may vary by students' academic ability, in ways that are ambiguous ex ante. On the one hand, students with lower ability might struggle to master advanced concepts, so exposure to frontier knowledge could fail to help or even hurt them. On the other hand, if frontier content primarily increases graduation rates by boosting motivation, then high-ability students—who typically already possess strong intrinsic motivation—might see smaller marginal gains.

To test for these hypotheses, panel (a) of Figure 4 shows estimates from a version of equation (6) where β is allowed to vary by quintiles of each student's SAT/ACT score. We assign students without a reported score to a separate "No Info" category. This group is mainly composed of transfer students from community colleges, who are typically not required to submit a test score upon transfer.[19]

The results support both hypotheses. Among students who report a score, the effect of frontier knowledge is negative for students at the bottom of the ability distribution and largest for those in the middle and at the top. A one-sd increase in proximity *reduces* the probability of graduating for

[19] All four-year public institutions in our sample (and virtually all in Texas) require SAT/ACT scores for first-year admissions. Transfer students, however, need to submit scores only if they have not completed enough academic credits; the minimum credit threshold varies by school. Consequently, students without reported SAT/ACT scores almost certainly started at a two-year college, for which scores are not required. Until 2014 one of the schools in our sample, the University of Houston–Clear Lake, admitted only upper-class transfers, who were not required to provide test scores.

Figure 4: Frontier Knowledge Proximity and Students' Outcomes, by Ability



Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students graduating with the same major declared upon enrollment (panel (b)), an indicator undergraduate students attending graduate school within Texas (panel (c)), and average log quarterly earnings 1-3 years and 4-6 years after each student's predicted graduation year (panel (d)). Each coefficient is an estimate of β interacted with indicators for quintiles of SAT/ACT scores, obtained using a measure of average proximity that is residualized using instructor fixed effects and then standardized. "N/A" refers to students without a test score. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. In panel (b), the sample is restricted to students who graduate; in panel (c), it is restricted to undergraduate students. Robust standard errors in parentheses.

students in the bottom quintile of the SAT/ACT distribution, by 1.3 pp or 2% of the mean for this group (Figure 4, top left panel, significant at 5%). The effect becomes positive at 1.6 pp (2%) for the second quintile and peaks at 2.9 pp (3.5%) for the third quintile. It then falls to 1.6 pp (2%) for the fourth quintile and is 2.7 pp (3%) for the top quintile (all significant at 1%).

Most notably, transfer students (the "N/A" category) exhibit the largest effect: a 3.7 pp increase in graduation probability, equivalent to 5% of their mean. Together, these results corroborate the hypothesis that frontier knowledge primarily enhances graduation by increasing students' motivation, with stronger effects for those who both have sufficient ability to benefit and, in the case of transfer students, lack prior access to such content at two-year institutions.

6.1.2 Differences by Family Income

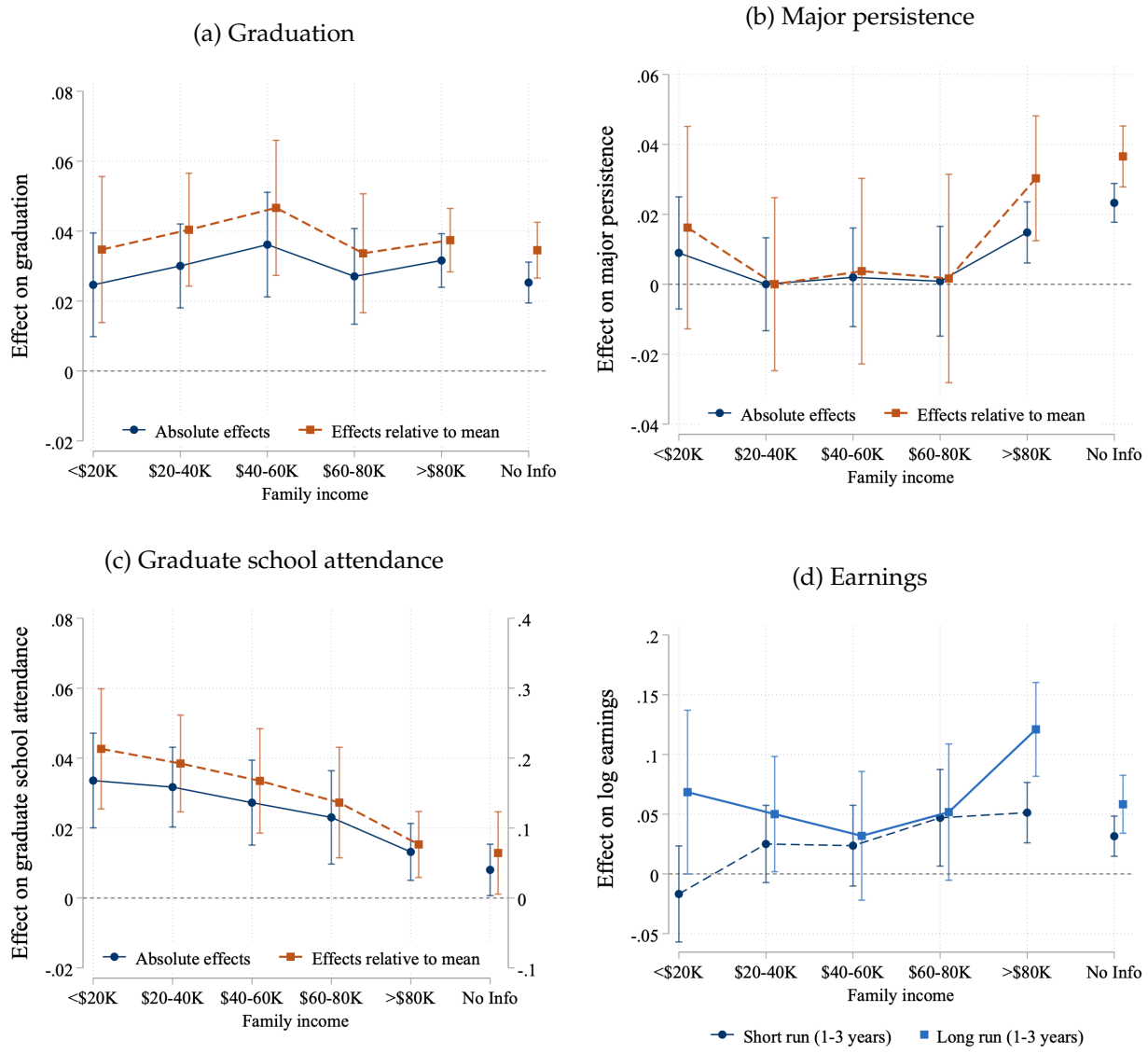
The relationship between the effects of frontier knowledge exposure and students' socio-economic background is a priori ambiguous. On the one hand, students from lower-income families may face higher perceived opportunity costs and more uncertain returns on college.[20] Conditional on ability, frontier content could particularly benefit these students by raising the perceived value of completing a degree. On the other hand, frontier knowledge may complement resources such as tutoring, discretionary time, or other academic supports that higher-income students typically access more easily. To test these competing hypotheses, panel (a) of Figure 5 reports estimates from a version of equation (6) in which β is allowed to vary across five family-income categories: below \$20 K, \$20–40K, \$40–60K, \$60–80K, and above \$80K (we keep students without a reported family income as a separate category).

We find modest evidence for both channels. Although the impact of frontier proximity is positive, statistically significant, and fairly stable across income groups, it is slightly larger in the middle of the distribution. A one–standard–deviation increase in proximity raises graduation rates by 2.5 pp (3.5% of the group mean) for students with family income below \$20K and by 3.0 pp (4%) for those in the \$20–40K range. It peaks at 3.6 pp (4.7%) for the \$40–60K group and is 2.7 pp (3.4%) for \$60–80K and 3.2 pp (3.7%) for above \$80K. These results imply that frontier knowledge exposure may be especially effective at boosting persistence for students with incomes near the median.

Panel (a) of Appendix Figure A3 further decomposes graduation effects by ability quintiles, separately for students with family incomes above and below \$80K. These estimates show that higher-income students benefit from frontier knowledge exposure across the entire ability distri-

[20] In our sample, 71% of students with family income below \$20K graduate, compared to 86% of those with family income above \$80K.

Figure 5: Frontier Knowledge Proximity and Students' Outcomes, by Family Income



Note: The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students graduating with the same major declared upon enrollment (panel (b)), an indicator undergraduate students attending graduate school within Texas (panel (c)), and average log quarterly earnings 1-3 years and 4-6 years after each student's predicted graduation year (panel (d)). Each coefficient is an estimate of β interacted with indicators for family income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K, obtained by also interacting lagged proximity with these indicators and using a measure of average proximity residualized from instructor fixed effects. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. In panel (b), the sample is restricted to students who graduate; in panel (c), it is restricted to undergraduate students. Robust standard errors in parentheses.

bution, and their effects are higher for higher-ability students. In contrast, lower-income students exhibit negative effects in the bottom ability quintile, with the impact peaking in the middle and top quintiles. These results suggest that resources amplify the payoff to ability in absorbing advanced material.

6.2 Persistence in The Major

In our sample, 41% of all students graduate with a different major from the one declared when enrolling. Survey evidence indicates that interest in a topic is one of the main determinants of major choice (Malgwi et al., 2005). By making courses more interesting, frontier knowledge could therefore increase the likelihood that students remain with their initially declared major, a behavior we refer to as persistence in the major. We test for this hypothesis in panel (b) of Table 3, which displays estimates of equation (6) using an indicator for graduating with the same major as the one declared upon enrollment as the dependent variable. Due to this variable construction, these estimates are obtained on the subsample of students who graduate.

These estimates indicate that exposure to frontier knowledge significantly increases the likelihood of major persistence. A one-sd increase in average course proximity increases this probability by 1.6 pp, or 3% of the mean (Table 3, panel (b), column 2, significant at 1%). This estimate is primarily driven by undergraduate students, who experience an increase of 2.4 pp (4.6% or the group mean) compared to 0.6 pp for graduate students (columns 3 and 4). Effects are larger for women (2.1 pp, column 4) than for men (1.4 pp, column 5). They are significant across all fields and larger for students who start in Business (4.6 pp or 8.7% of the group mean), followed by Social Sciences (2.0 pp or 3%), Humanities (1.3 pp or 4%), and STEM (0.8 pp or 1.4%, Appendix Figure A2, panel (b)).

Estimates by academic ability show the largest impacts on students in the middle and at the bottom of the distribution. For students in the second, third, and fourth quintiles, a one-sd increase in proximity increases major persistence by 2.7 pp, 3.1 pp, and 2.5 pp respectively (6%, 7%, and 6% of their respective group means; Figure 4, top right panel, significant at 1%). Effects are slightly similar for the bottom quintile (2.7 pp or 5%) and zero for the top quintile. Transfer students without a reported score, who display the highest rate of major persistence at baseline (75%, compared to 56% for the full sample) see a smaller but still significant increase, at 1.3 pp or 1.8% of the mean. These results suggest that frontier knowledge exposure may improve the odds of graduating by increasing major persistence, particularly among marginal students.

The effects of frontier knowledge on major persistence are driven by students at the very top

of the family income distribution. They are indistinguishable from zero for students with incomes below $80K and equal to 1.5 (or 3%) for those with incomes above $80K (Figure 5, panel (b)).

6.3 Graduate School Attendance

Exposure to frontier knowledge significantly increases the likelihood of pursuing graduate studies. Among undergraduates, a one-sd increase in proximity raises in-state graduate-school attendance by 1.6 pp (11% of the mean, equal to 15%; Table 3, panel (c), column 2). Male students experience slightly larger gains than female students, 1.7 pp (13%) compared to 1.4 pp (8%, Table 3, panel (c), columns 5 and 6). By macro-field, Business majors benefit the most (7.5 pp or 41%), followed by Humanities (1.5 pp or 11%), and Social Sciences (1.1 pp or 5%). The effect is insignificant for STEM (Appendix Figure A2, panel (c)).

Examining effects by academic ability shows the largest impacts in the middle of the SAT/ACT distribution. For students in the third quintile, a one-sd increase in proximity boosts graduate-school attendance by 2.4 pp (15% of the group mean; Figure 4, panel (c), significant at 1%). Effects are smaller for the first and second quintiles—1.1 pp (8%) and 2.0 pp (13%), respectively— and become insignificant for the top two quintiles. Transfer students without a reported score also see a substantial increase of 1.8 pp (12%, significant at 1%). This pattern is consistent with the hypothesis that frontier content stimulates interest and motivation for graduate studies, especially among students who possess sufficient ability but would not otherwise pursue them.

The effect of frontier proximity on graduate-school attendance is greatest among lower-income students and declines along the income distribution. A one-sd increase raises attendance by 3.4 pp (21%) for those with parental income below $20K, by 3.2 pp (19%) for the $20–40K group, and by 2.7 pp (17%) for the $40–60K group (Figure 5, top right panel). The effect remains significant but declines to 2.3 pp (14%) for $60–80K and 1.3 pp (8%) for above $80K. These findings indicate that frontier knowledge can be particularly effective at promoting graduate-school access among less advantaged students.

Appendix Figure A3 shows graduate-school attendance effects by ability, separately for high- and low-income students. Among high-income students, effects are largest for lower-ability and transfer students (who report no score). For low-income students, effects increase with ability and are also substantial for transfer students. These results suggest that frontier exposure is especially important for attracting low-income, high-ability students—and those who began at two-year institutions—to graduate studies. They further underscore the role of frontier content in motivating students and the complementarity between advanced material and students' available resources.

6.4 Earnings

Exposure to frontier knowledge may affect students' earnings both directly—by expanding their labor-market relevant knowledge and skills—and indirectly, by increasing their likelihood of graduating and attend graduate school (shown to be linked to higher earnings; Altonji and Zhong, 2021). To estimate these earnings effects, we calculate each student's "short-run" earnings (averaged over 1-3 years post-graduation) and "long run" earnings (averaged over 3-6 years post-graduation). In computing these averages, we include only quarters with earnings above $1,000 (our results are robust if we include all positive earnings; we examine labor market participation in Section 6.6).

Table 4: Frontier Knowledge Proximity and Students' Earnings

	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Short-run earnings	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.046***	0.033***	0.022***	0.037***	0.030***	0.035***
	(0.007)	(0.007)	(0.009)	(0.011)	(0.010)	(0.011)
Adj. R^2	0.14	0.14	0.09	0.34	0.10	0.15
N	135,459	135,459	115,837	19,622	73,045	60,550
Panel (b): Long-run earnings	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.091***	0.069***	0.073***	0.053***	0.061***	0.088***
	(0.009)	(0.010)	(0.012)	(0.016)	(0.014)	(0.015)
Adj. R^2	0.14	0.15	0.11	0.27	0.11	0.15
N	63,436	63,436	53,975	9,461	34,354	27,917
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is the logarithm of average quarterly earnings 1-3 years after each student's predicted graduation year (panel (a)) and the logarithm of average quarterly earnings 4-6 years after each student's predicted graduation year (panel (b)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, residualized using instructor fixed effects and calculated as the average across all courses in the student's transcript for which we observe a syllabus. The sample is restricted to students with observed earnings greater than $1,000 for at least one quarter post-predicted graduation year. In column 3, the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below $20K, between $20-40K, between $40-60K, between $60-80K, and above $80K. Columns 2-6 additionally control for the lagged average proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

We find that exposure to frontier knowledge has statistically and economically significant aver-

age effects on student earnings, which grow over time. A one-sd increase in average course proximity increases students' earnings by 3.4% in the short run (Table 4, panel (a), column 2, significant at 1%) and by 7.1% in the long run (panel (b), column 2, significant at 1%). Omitting the lagged proximity controls yields larger estimates (4.7% and 9.5%, respectively; Table 4, panels (a) and (b), column 1), implying that students with higher earnings potential tend to select higher-proximity courses.

Graduate students experience a larger short-run earnings boost (3.8%) than undergraduates (2.2%; panel (a), columns 3 and 4). However, undergraduates see a greater long-run gain, 7.6% compared to 5.4% for graduates. A possible explanation is that some undergraduates delay full-time work to pursue graduate studies when exposed to frontier knowledge. Short-run effects are similar across gender (panel (a), columns 5 and 6), while long-run effects are larger for men (9.2% compared to 6.3%, panel (b)).[21]

Short-run effects appear driven primarily by Business majors, who experience a 6.5% increase (Appendix Figure A2, panel (d), hollow marks), but they are also significant for the other fields. Long-run effects are again largest for Business (13%), followed by STEM and Humanities (9% and 8%, respectively), and are indistinguishable from zero for the Social Sciences (Appendix Figure A2, panel (d), full marks).

The short-run earnings effects of frontier knowledge are concentrated among students in the top four quintiles of the ability distribution. A one-sd increase in proximity has no effect for students in the bottom quintile (Figure 4, panel (d), circle marks). In contrast, the same one-sd increase raises earnings by approximately 5%-6% for students in each of the other quintiles. The effect is smaller at 3% for students without a reported score.

The long-run earnings effects of frontier knowledge exposure are positive across the entire ability distribution. They are smaller for students without a score and for those in the bottom quintile of the distribution (both equal to 6.0%; Figure 4, panel (d), square marks). They grow to 8% for students in the second quintile, peak at 12% for the third quintile, and are equal to 9% and 12% for the fourth and fifth quintiles, respectively. Larger long-run gains for students in the middle and top of the ability distribution mirror the patterns found for graduation rates, indicating that improvements in educational attainment are important channels for the earnings effects of frontier knowledge. However, the modest effects for students without a score imply that this group may

[21] A possible explanation for this pattern is part-time employment, which rises over the course of a woman's career but remains stable (and low) for men. In 2022, 20% of U.S. female college graduates were employed part-time, compared to only 11% for men (U.S. Bureau of Labor Statistics, 2023).

face additional barriers in capturing the benefits of frontier material.

The earnings effects of frontier knowledge exposure are larger for students from higher-income families, both in the short and in the long run. In the short-run, effects are indistinguishable from zero for students with family incomes below $60K. They rise to 5% for those with incomes above $60K (Figure 5, panel (d), circle marks. Long-run effects are significant at 7% for students with incomes below $20K. They are insignificant for those with incomes between $20K and $80K, and are the highest at 13% for those with family incomes above $80K, respectively (Figure 5, panel (d), square mark). These findings indicate that, while low-income students gain from frontier knowledge exposure, most of the benefits accrue to higher-SES students, suggesting a complementarity between the skills provided by frontier content and individual resources in shaping students' labor market success.

Appendix Figure A3 examines effects by ability, separately for high- and low-income students. In the long-run, the ability gradient is steeper for lower-income students: Among them, frontier exposure has no measurable effect in the bottom ability quintile but has a significant impact in the top quintile, equal to the one of higher-income students. For higher-income students, the largest effect occurs in the middle of the ability distribution. These findings suggest that frontier exposure can substantially improve labor-market outcomes for low-income, high-ability students, and that family resources enhance the value of advanced material even for students not at the top of the ability distribution.

6.5 Robustness

Not controlling for instructor fixed effects Appendix Tables A3 and A4 show estimates using our raw proximity measure, not controlling for instructor effects. Estimates tend to be larger, in line with the hypothesis that instructors who teach courses with higher proximity also have positive direct effects on student outcomes.

Using alternative statistics of lagged proximity Appendix Tables A5 and A6 show estimates of our baseline models where we control for selection using the 25th, 50th, and 75th percentiles of the lagged proximity of courses taken and courses not taken by each student. Estimates are largely unchanged.

6.6 Selection into the Earnings Sample

Our earnings estimates are based on the sample of individuals with quarterly earnings above $1,000. We thus exclude from our analysis non-participants to the labor force, the unemployed, and those

who leave the state after graduation. Other than affecting external validity, this sample restriction could lead to endogenous sample selection, since these margins of labor market participation may themselves be affected by frontier-knowledge exposure. For example, exposure to frontier knowledge could increase in-state labor-force attachment and in-state job opportunities (raising a student's likelihood of appearing in the earnings sample) or improve out-of-state opportunities (lowering that likelihood).

To investigate these channels, we estimate our empirical models using two alternative dependent variables: (a) the probability of never being observed in the earnings sample 1-6 years following graduation (a proxy for having moved out of state), and (b) the probability of having no-earnings spells 4 quarters long or longer, preceded and followed by at least a quarter with positive earnings (a proxy for a lack of labor market attachment or unemployment).

These estimates indicate that exposure to frontier knowledge has a sizable positive effect on the probability of never being observed in the earnings data, 0.06 pp or 15% of the mean (equal to 0.38; Appendix Table A7, column 1, significant at 1%). A possible interpretation for this finding is that frontier knowledge motivates students to seek job opportunities out of state. In contrast, we find no significant effect on the probability of a four-quarter (or longer) earnings hiatus (Appendix Table A7, column 2).

6.7 Taking Stock

The results in this section demonstrate that exposure to frontier knowledge substantially improves students' persistence in higher education and their propensity to pursue graduate studies. These improvements appear to stem from increased motivation and engagement generated by frontier content, especially for students who possess sufficient ability to benefit from advanced material yet still face a meaningful risk of non-completion. In particular, low-income, high-ability students exhibit the largest gains in educational outcomes.

Frontier knowledge also yields sizable positive effects on post-graduation earnings. Although better educational outcomes explain part of this earnings boost, they cannot account for the full magnitude of the effect. Indeed, while low-income students experience the greatest improvements in persistence and graduate-school attendance, high-income students enjoy the largest increases in earnings. This pattern indicates a complementarity between a student's resources and access to frontier content in driving labor-market success.

7 The Role of Course Instructors

The findings presented in the previous section indicate that expanding access to frontier knowledge can significantly reduce disparities in educational achievement and labor-market outcomes. Closing gaps in the coverage of cutting-edge material across courses could help more students—particularly those from disadvantaged backgrounds—succeed in college and beyond.

To achieve this goal, it is critical to understand the sources of these disparities. Course instructors emerge as a likely determinant, given their central role in education production (De Vlieger, Jacob, and Stange, 2020). Indeed, instructors account for the majority of the variation in courses' proximity to frontier knowledge (Table 2). Motivated by these findings, we now examine how instructors affect the frontier knowledge content of their courses, with the ultimate aim of identifying strategies to maximize students' exposure to cutting-edge knowledge.

7.1 Estimating the Causal Effect of Instructors

To measure the causal impact of instructors on course content, we exploit instructors' turnover across courses in an event-study framework. Specifically, we estimate the change in a course's proximity to frontier knowledge when the instructor changes. We use the following two-way fixed-effects model:

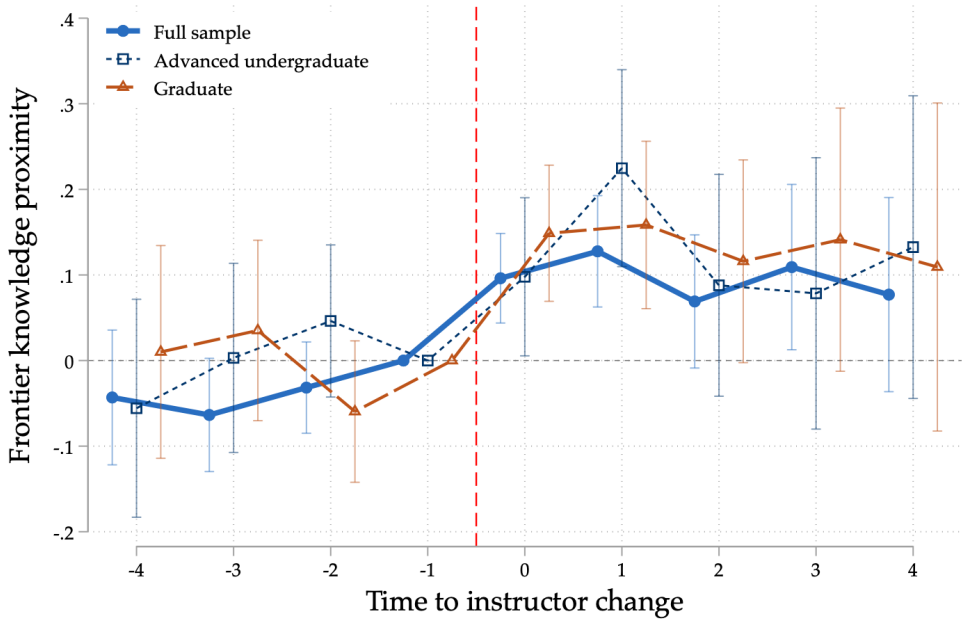
$$p_{ct} = \sum_{k=-4}^{4} \delta_k \mathbb{1}(t - T_c = k) + \gamma_c + \eta_t + \varepsilon_{ct}, \tag{7}$$

where p_{ct} is the frontier knowledge proximity measure for course c in year t, averaged across sections; T_c denotes the year of the first instructor change for course c in our data; γ_c are course fixed effects; and η_t are year fixed effects.[22] We set the indicator function to zero for all courses without an instructor change, which serve as the comparison group, and normalize δ_{-1} to 0. We cluster standard errors at the course level. The parameters δ_k represent the change in proximity k years after an instructor change, relative to the year before the instructor 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 frontier knowledge proximity.

Estimates on the combined Texas-OSP syllabi sample show a marked increase in proximity immediately following an instructor change. Estimates of δ_k are indistinguishable from zero in the

[22] 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 6: Event Study: Frontier Knowledge Proximity When the Course Instructor Changes



Notes: Estimates and confidence intervals of the parameters δ_k in equation (7), obtained 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. Standard errors clustered at the course level.

years leading to an instructor change. Following the change, proximity increases by about 0.1 (Figure 6, solid line). As indicated by Figure 2, this increase is equivalent to a 5% change in the content of the median syllabus. Estimates remain robust when we account for heterogeneous effects across treatment cohorts, using the approach of Sun and Abraham (2021) (Appendix Figure A4). Effects are slightly larger for upper undergraduate and graduate courses (Figure 6, dashed lines).

In Table 5, we investigates differences across macro-fields, pooling together years preceding and following an instructor change (the variable *After change* equals one in years following the change). The proximity increase is the largest for Business courses (0.14, column 2, significant at 10%).

These results indicate that instructors play an active role in determining the content of the courses they teach. New instructors who take over a course significantly update its content, bringing it closer to the knowledge frontier. A flat trend prior to the change suggests that instructors who teach the same course for many years tend to leave content unchanged.

7.1.1 Instructors' Research Activity and Frontier Knowledge

Yet, instructors differ among themselves in many dimensions. One of the most relevant for our analysis is their research activity. Ladder (or tenure-track) faculty combine teaching with research.

Table 5: Frontier Knowledge Proximity When the Course Instructor Changes: Estimates by Macro-Field

		Field			
	All	Business	Humanities	STEM	Social Sciences
	(1)	(2)	(3)	(4)	(5)
After change	0.058*	0.137*	0.071	0.047	0.064
	(0.031)	(0.079)	(0.067)	(0.056)	(0.052)
N (Course x year)	430461	41447	116419	152736	105725
# Courses	140743	13199	39213	49011	34998
Course FE	Yes	Yes	Yes	Yes	Yes
Field * year FE	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. The dependent variable is the proximity to the knowledge frontier. The variable *After change* is an indicator for years following an instructor change. All specifications control for course and year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Non-ladder faculty (such as lecturers, adjunct professors, or clinical professors) are mostly tasked with teaching. We now investigate whether instructors' research activity affects the presence of frontier knowledge in their courses.

The relationship between instructors' research activity and frontier knowledge proximity is ex ante unclear. On the one hand, research and teaching compete for the instructor's time. Research-active instructors may face a higher opportunity cost of keeping a course up to date and thus teach courses with a lower proximity. On the other hand, research-active instructors are more familiar with the knowledge frontier and may find it easier to keep a course up to date.

In our data, measures of instructors' research activity and quality (such as publications, citations, and grants) are positively correlated with the frontier knowledge proximity of the courses they teach (Appendix Figure A5, panel (a)). This relationship, though, does not necessarily indicate a causal link. For example, it could be the case that schools assign more research-active instructors to courses that cover more frontier knowledge. To isolate the effect of instructors on course content, we estimate the change in a course's frontier knowledge proximity driven by changes in the characteristics of its instructor. We use the following model:

$$p_{ct} = \beta q_{ct} + \gamma_c + \phi_{f(c)l(c)t} + \varepsilon_{ct} \tag{8}$$

where q_{ct} is a measure of research activity of the instructors of course c in year t (such as indicators for having at least one publication or grant and the standardized number of publications and citations). The inclusion of course fixed effects γ_c implies that our estimates are obtained off of changes in course instructors over time. Field-by-course level-by-year fixed effects ϕ_{flt} account for any determinants of proximity or research activity specific to courses of a given level belonging to a given field and year. To the extent that course-specific factors, which might drive the allocation of instructors across courses, do not change systematically over time, β captures the causal effect of within-course changes in the instructor's research activity on the course's frontier knowledge proximity.

Estimates of β show that an instructor's research activity positively affects the frontier knowledge proximity of the courses she teaches. Switching from an instructor without publications to one with publications increases proximity by 0.16 (Table 6, panel (a), column 1, significant at 1%). This effect is primarily driven by Social Sciences and STEM courses (columns 5 and 4, respectively) and is larger for graduate courses (column 7).

The intensity of an instructor's research activity matters, too. A one-sd increase in instructor publications increases proximity by 0.04 (equivalent to updating 3% of a course's syllabus; Table 6, panel (b), column 1, significant at 1%). This relationship is particularly pronounced for Social Sciences, where the same increase is associated with a 0.07 increase in proximity (4% of a course's syllabus, panel (b), column 5), and for graduate courses (column 7). It is instead more moderate for STEM (column 4), suggesting that for this macro-field the presence of a ladder faculty matters more than the intensity of research activity per se.

Instructors' research quality, measured by citation counts, also affects proximity, but only for Social Sciences and graduate-level courses. A one-sd increase in citations increases proximity by 0.02 in the main sample, although imprecisely estimated (panel (c), column 1, p-value equal to 0.24). This estimate is larger at 0.082 for Social Sciences and at 0.034 for graduate courses (columns 5 and 7, significant at 5 and 10%, respectively).

Using grant receipts as a measure of research activity and quality paints a similar picture. A switch from an instructor who never received a grant to one with at least one grant leads to a a 0.11 increase in proximity (Table 6, column 1, significant at 1%). This estimate suggests that public investments in academic research can yield additional private and social returns in the form of more updated instruction.[23]

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

Table 6: Frontier Knowledge Proximity and Instructors' Research Activity: Publications, Citations, and Grants

Panel (a): any publications	All (1)	Business (2)	Humanities (3)	STEM (4)	Soc. Sci. (5)	Upper Ugrad (6)	Grad (7)
At least one publ.	0.077**	-0.036	0.159**	0.066	0.192***	0.048	0.155***
	(0.036)	(0.082)	(0.075)	(0.065)	(0.057)	(0.061)	(0.057)
N (Course x year)	300492	29432	59002	119504	80206	102447	100188
# Courses	79854	7795	16897	32019	21768	25963	29034
Panel (b): # publications	(1)	(2)	(3)	(4)	(5)	(6)	(7)
#publications (sd)	0.043***	0.001	0.018	0.046*	0.087***	0.030	0.069***
	(0.015)	(0.035)	(0.033)	(0.025)	(0.023)	(0.028)	(0.022)
N (Course x year)	645246	64750	167808	220470	162845	202687	212660
# Courses	162299	15912	45228	56414	41490	48980	57495
Panel (c): # citation	(1)	(2)	(3)	(4)	(5)		
#citations (sd)	0.016	-0.066**	-0.015	0.022	0.100***	0.023	0.041**
	(0.015)	(0.030)	(0.040)	(0.025)	(0.032)	(0.032)	(0.021)
N (Course x year)	645246	64750	167808	220470	162845	202687	212660
# Courses	162299	15912	45228	56414	41490	48980	57495
Panel (d): any grants	(1)	(2)	(3)	(4)	(5)	(6)	(7)
At least one grant	0.102***	0.039	0.152**	0.050	0.091**	0.055	0.067
	(0.030)	(0.076)	(0.062)	(0.054)	(0.042)	(0.048)	(0.046)
N (Course x year)	645246	64750	167808	220470	162845	202687	212660
# Courses	162299	15912	45228	56414	41490	48980	57495
Panel (e): fit with course	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fit w/top course (sd)	0.119***	-0.033	-0.013	0.111*	0.086	0.141**	0.096
	(0.042)	(0.096)	(0.162)	(0.063)	(0.063)	(0.072)	(0.062)
N (Course x year)	106485	6952	12162	59637	24016	45332	32657
# Courses	27561	1803	2948	15930	6326	10652	9772
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field * level * year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates; one observation is a course in a given year. The dependent variable is the frontier knowledge proximity; the independent variables are an indicator for positive instructor publications (panel (a)), counts of publications (panel (b)), counts of citations (panel (c)), an indicator for having received a government grant (panel (d)), and a measure of fit between the instructor's research agenda and the content of the course (panel (e)). All specifications control for course and field-by-course level-by-year fixed effects. Standard errors in parentheses are clustered at the course level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

7.2 Instructor-Course Fit

A possible explanation for the effect of instructors' research activity on frontier knowledge proximity is that research-active instructors are more familiar with this type of knowledge and can more effectively incorporate it in their courses. This mechanism should be especially strong for instruc-

tors who teach courses closely related to their own research agenda. For example, a labor economist should be better equipped to teach a course on labor economics than a course on industrial organization.

We test this hypothesis by constructing a measure of "fit" between the course and the instructor's research, 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 instructors doing research in econometrics teach courses with lower proximity compared with instructors doing research in macroeconomics.[24]

Our measure of instructor-course fit is positively correlated to course proximity (Appendix Figure A5, panel (d)). This positive correlation bears a causal interpretation: Estimates of equation (8) using the standardized instructor-course fit as the explanatory variable indicate that a one-sd higher instructor-course fit leads to a 0.12 higher course proximity to the knowledge frontier, equivalent to a 6% update in the content of the median syllabus (Table 6, significant at 1%).

Together, our analysis indicates that instructors play the most important role in the creation of course content and the dissemination of frontier knowledge. 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.

8 Discussion and Conclusion

This paper has documented differences in the coverage of frontier knowledge across university courses and their implications for student success. Using the text of course syllabi as data and developing a new measure of course proximity to the knowledge frontier, we have uncovered a significant amount of variation in frontier knowledge across courses, even within the same university. We have also demonstrated that these differences are important for students: Exposure to course content closer to the knowledge frontier makes students more likely to persist in their major,

[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.

graduate, and attend graduate school, likely by increasing their motivation and enthusiasm for the material. Exposure to frontier knowledge also increases students' earnings, to an extent that cannot fully be explained by improvements in educational attainment. This suggests that exposure to cutting-edge material fosters the development of skills with high returns on the labor market.

An important implication of our findings is that ensuring that courses cover updated material has the potential of boosting student outcomes. A promising avenue to achieve this goal is a proper deployment of faculty across courses. Indeed, instructors explain most of the differences in frontier knowledge proximity across courses. Research-active instructors, as well as those with research interests more in line with the topic of the course, are particularly effective at keeping materials up to date, especially for graduate-level courses.

We end with a note on data and methods. To conduct our analysis, we have assembled a novel dataset of course syllabi linked to individual student records and developed a new empirical design. This setup can be used to study many other dimensions of course content, a valuable exercise to better understand what constitutes high-returns higher education. We plan on making our syllabi data available to all researchers to facilitate this line of research.

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.

Altonji, Joseph G, and Ling Zhong, 2021, The labor market returns to advanced degrees, *Journal of Labor Economics* 39, 303–360.

Andrews, Michael J, 2023, How do institutions of higher education affect local invention? evidence from the establishment of us colleges, *American Economic Journal: Economic Policy* 15, 1–41.

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.

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.

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.

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

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

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

Cunha, Jesse M, and Trey Miller, 2014, Measuring value-added in higher education: Possibilities and limitations in the use of administrative data, *Economics of Education Review* 42, 64–77.

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.

Dale, Stacy Berg, and Alan B Krueger, 2002, Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables, *The Quarterly Journal of Economics* 117, 1491–1527.

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, and Kadeem L Noray, 2020, 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).

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, and G Bulman, 2015, Computing the value-added of american postsecondary institutions, *Internal Revenue Service, US Department of the Treasury, Washington, DC* .

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* .

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.

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* .

Malgwi, Charles A, Martha A Howe, and Priscilla A Burnaby, 2005, Influences on students' choice of college major, *Journal of education for business* 80, 275–282.

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

Mountjoy, Jack, and Brent R Hickman, 2021, The returns to college (s): Relative value-added and match effects in higher education, Technical report, National Bureau of Economic Research.

of University Professors, American Association, 1940, 1940 statement of principles on academic freedom and tenure, *AAUP Bulletin* 64, 108–112.

Romer, Paul M, 1986, Increasing returns and long-run growth, *Journal of political economy* 94, 1002–1037.

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.

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

U.S. Bureau of Labor Statistics, 2023, Current Population Survey, Household Data Annual Averages, 2022: Table 8. Employed persons by class of worker and usual full- or part-time status, educational attainment, sex, race, and Hispanic or Latino ethnicity, https://www.bls.gov/cps/cpsaat08.htm, [Accessed: 2025-06-04].

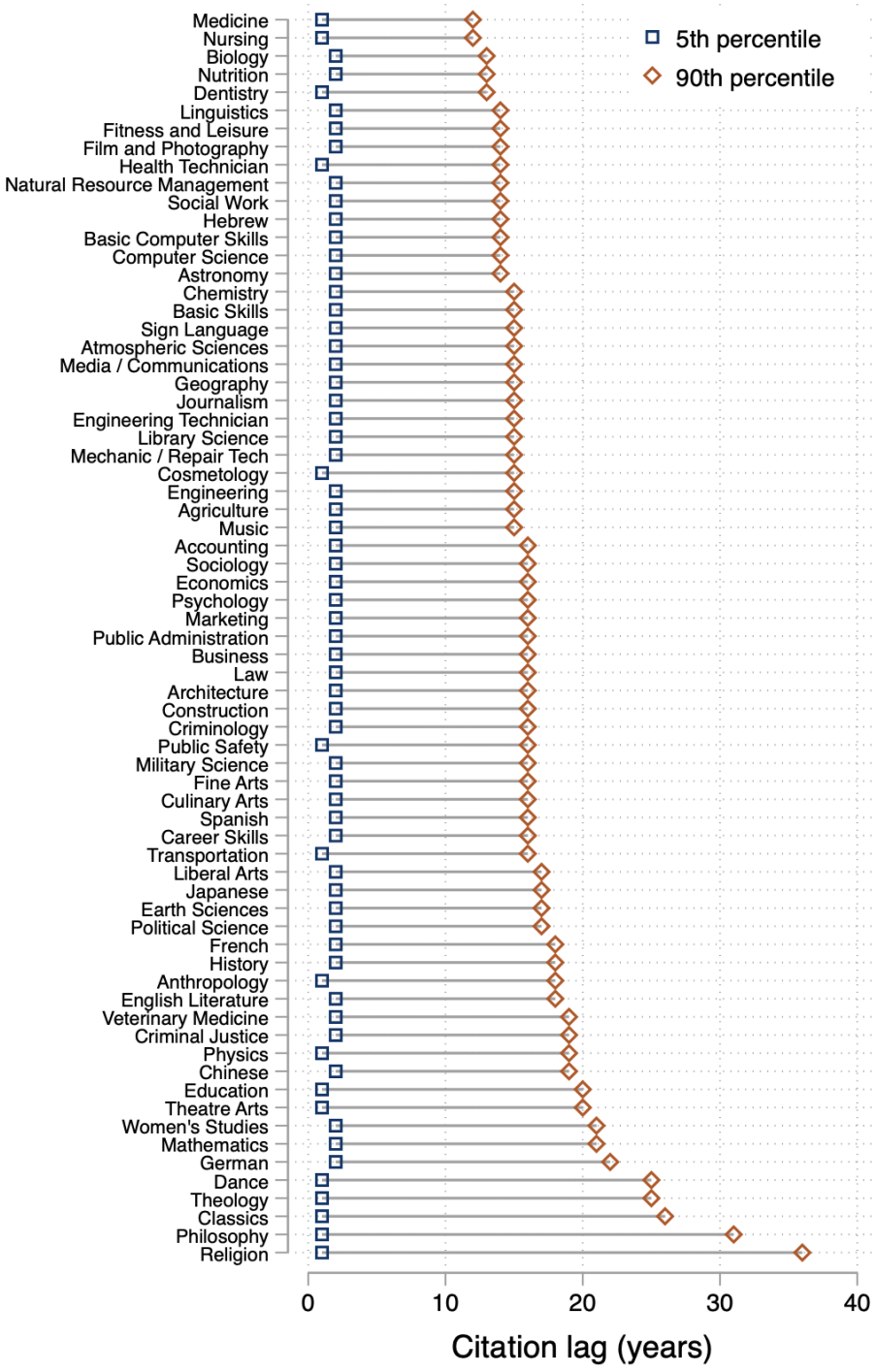
Williams, Heidi L, 2013, Intellectual property rights and innovation: Evidence from the human genome, *Journal of Political Economy* 121, 1–27.

Appendix

For online publication only

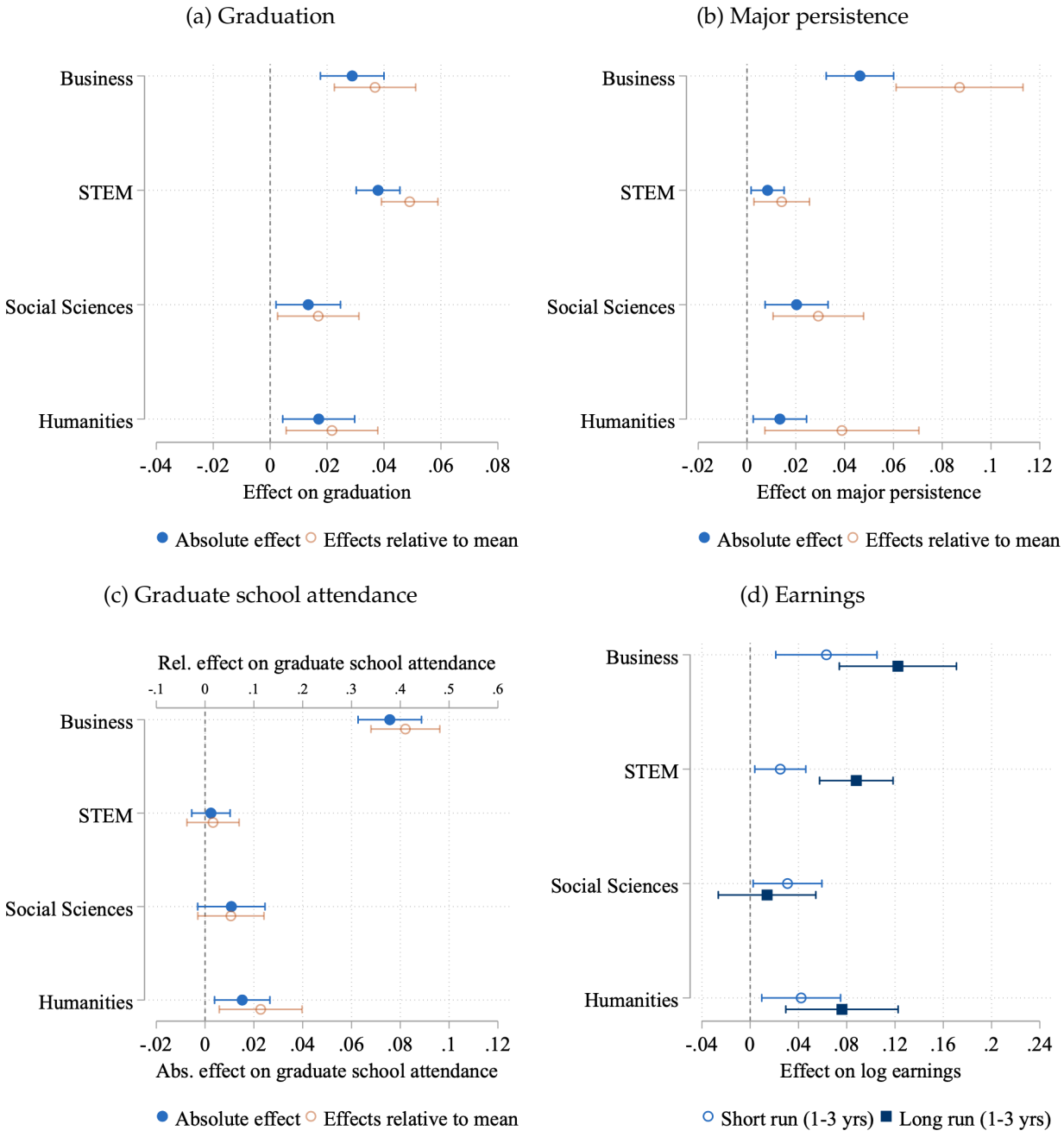
Appendix A Additional Tables and Figures

Figure A1: Citation Lags by Field: 5th and 90th Percentiles, OSP Sample



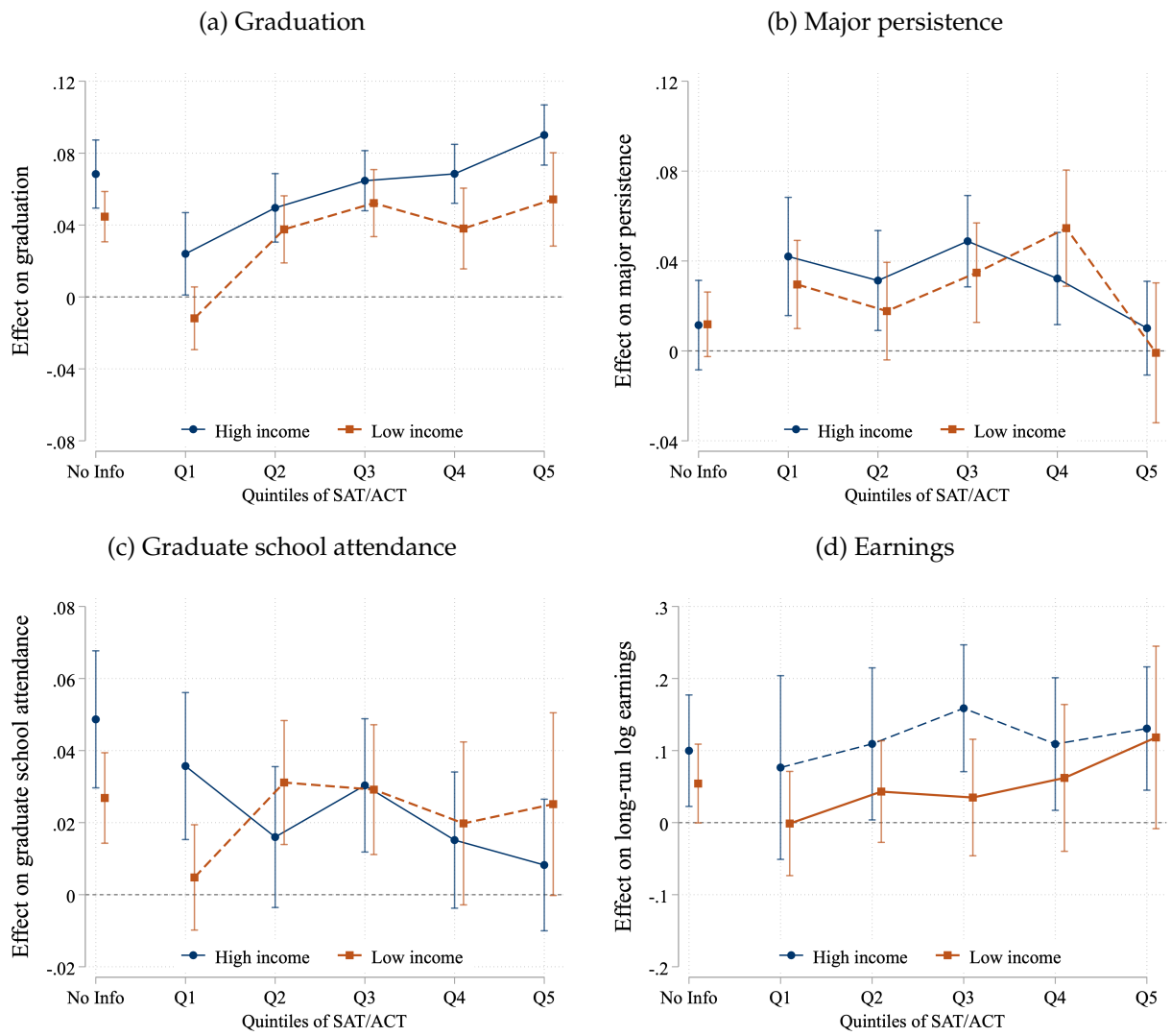
Notes: 5th and 90th percentile of the citation lag distribution in each field, used to calculate the frontier knowledge proximity for the syllabi in each field.

Figure A2: Frontier Knowledge Proximity and Students' Outcomes, by Macro-Field



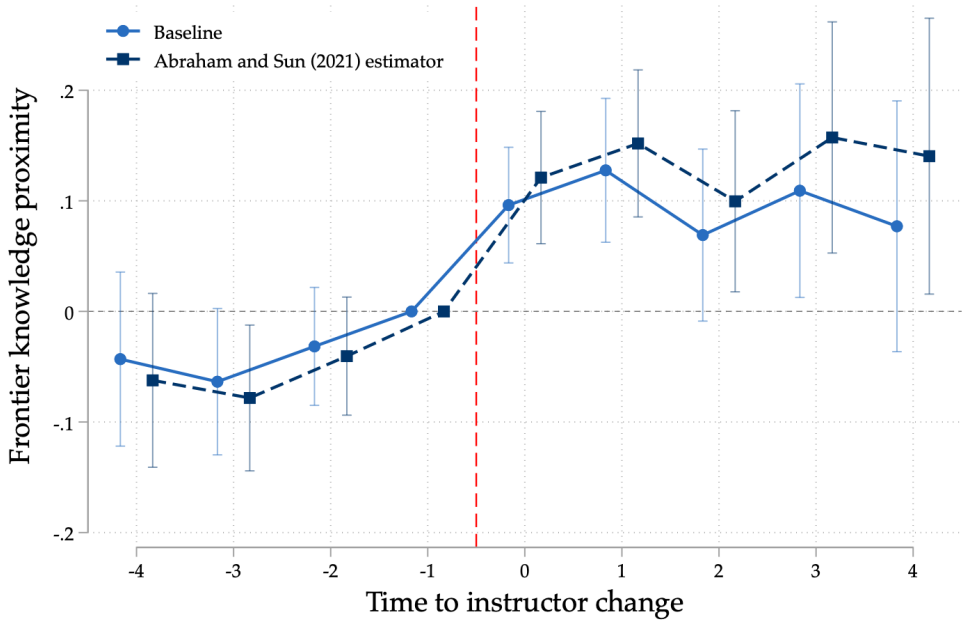
Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students graduating with the same major declared upon enrollment (panel (b)), an indicator undergraduate students attending graduate school within Texas (panel (c)), and average log quarterly earnings 1-3 years and 4-6 years after each student's predicted graduation year (panel (d)). Each coefficient is an estimate of β interacted with indicators for macro-fields, obtained using a measure of average proximity that is residualized using instructor fixed effects and then standardized. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. In panel (b), the sample is restricted to students who graduate; in panel (c), it is restricted to undergraduate students. Robust standard errors in parentheses.

Figure A3: Frontier Knowledge Proximity and Students' Outcomes, by Ability and Income



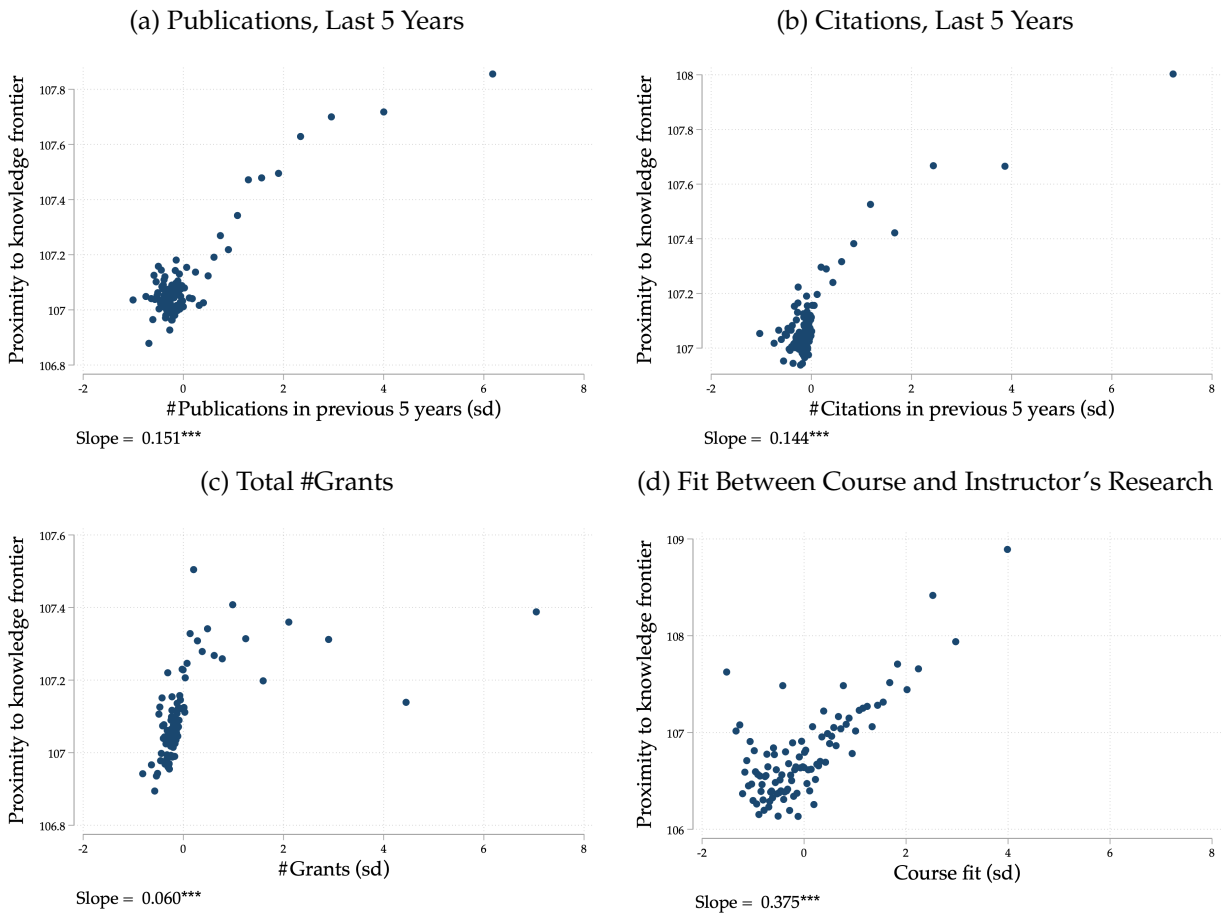
Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students graduating with the same major declared upon enrollment (panel (b)), an indicator undergraduate students attending graduate school within Texas (panel (c)), and average log quarterly earnings 1-3 years and 4-6 years after each student's predicted graduation year (panel (d)). Each coefficient is an estimate of β interacted with indicators for quintiles of SAT/ACT scores, obtained using a measure of average proximity that is residualized using instructor fixed effects and then standardized. "N/A" refers to students without a test score. "High-income" ("low-income") includes students with parental income above (below) \$80,000. Macro-fields are assigned based on major intentions indicated by the student upon enrollment. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. In panel (b), the sample is restricted to students who graduate; in panel (c), it is restricted to undergraduate students. Robust standard errors in parentheses.

Figure A4: Event Study of Frontier Knowledge Proximity Around an Instructor Change: Baseline and Abraham and Sun (2021) Estimator



Notes: Estimates of the confidence intervals of δ_k in equation (7), 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). We focus on courses taught by at most one instructor per year. Standard errors are clustered at the instructor level.

Figure A5: Frontier Knowledge Proximity and Instructors' Instructors' Characteristics



Notes: Binned scatterplots of the frontier knowledge proximity (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 highest proximity 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 proximity on the corresponding standardized variable, controlling for field-by-course level-by-year effects and clustering standard errors at the course level.

Table A1: Timeline of Course Enrollment, Add/Drop Decisions, and Syllabi Visibility - Texas Sample

School	Registration Period	Date Checked	Syllabi Visible?	Add/Drop Deadline
UT Austin	Dec 16 - Dec 20	Dec 16	No	Jan 30
Texas A&M	Nov 18	Nov 18	No	Jan 10
UT Dallas	Oct 21	Nov 20	No	Jan 26
West Texas A&M	Nov 6- Nov 11	Nov 20	Very few visible	Jan 28
Sam Houston State	Oct 31-Nov 14	Nov 19	No	Jan 18
Stephen F Austin	Oct 30 - Nov 4	Nov 19	No	Jan 24
U Houston Clear Lake	Nov 20	Nov 20	No	Feb 1

Note: The table summarizes the timeline of course enrollment, add/drop decisions, and syllabi visibility for the Winter/Spring semester of 2025 at the seven universities in our sample.

Table A2: Predicting Changes in Course Proximity Using Student Observables

	Univariate regression			Multivariate regression		
Dep. Var: ΔProximity	Coeff.	Std. Error	P-value	Coeff.	Std. Error	P-value
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.025	0.015	0.109	-0.024	0.016	0.149
Black	-0.045	0.028	1.114	-0.047	0.053	0.373
White	0.003	0.016	0.860	-0.005	0.049	0.920
Asian	0.058	0.028	0.035	0.051	0.049	0.305
Pacific Islander	-0.096	0.122	0.430	-0.111	0.131	0.395
Unknown race	-0.037	0.031	0.229	-0.044	0.060	0.466
Native	-0.071	0.050	1.587	-0.071	0.050	0.152
International	-0.013	0.023	0.581	-0.008	0.057	0.893
SAT/ACT>median	0.011	0.016	0.479	-0.006	0.022	0.780
No SAT/ACT	0.000	0.000	0.461	0.000	0.000	0.396
Income:						
Below $20K	0.010	0.035	0.763	-0.021	0.039	0.577
$20-40K	-0.014	0.028	0.621	-0.002	0.031	0.954
$40-60K	-0.027	0.030	0.370	0.032	0.032	0.319
$60-80K	-0.030	0.032	0.343	-0.025	0.034	0.468
Above $80K	-0.001	0.017	0.935	-0.001	0.023	0.978
N (course-year)					120,204	
F-stat of joint sign.					1.00	
F-stat p-value					0.46	

Note: OLS estimates; one observation is a course-year. The dependent variable is the change in course proximity from the previous year. The independent variables are average characteristics of all students taking the course. Columns 1-3 refer to univariate regressions of each independent variable on the change in proximity; column 4-6 refer to one single multivariate regression. All specifications control for field (as indicated by the course prefix) by-school-by-year fixed effects. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table A3: Frontier Knowledge Proximity and Students' Educational Attainment - Not Controlling for Instructor Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Graduation	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.059***	0.073***	0.101***	0.017***	0.072***	0.075***
	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
Mean of Dep. Var.	0.77	0.78	0.77	0.80	0.75	0.80
Adj. R^2	0.16	0.16	0.18	0.15	0.17	0.16
N	460,362	460,362	397,185	63,177	242,466	215,540
Panel (b): Major persistence	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	-0.026***	-0.020***	-0.027***	0.002	-0.026***	-0.013***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Mean of Dep. Var.	0.56	0.56	0.53	0.75	0.56	0.56
Adj. R^2	0.39	0.39	0.38	0.38	0.43	0.43
N	357,137	357,137	307,604	49,533	193,532	161,276
Panel (c): Grad. school attendance	Undergraduate				Females	Males
Proximity (sd)	0.034***	0.045***			0.048***	0.043***
	(0.002)	(0.003)			(0.003)	(0.005)
Mean of Dep. Var.	0.15	0.15			0.17	0.13
Adj. R^2	0.10	0.10			0.10	0.10
N	397,185	397,185			212,612	182,973
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students who graduate with the same major initially declared upon enrollment (panel (b)) and an indicator for undergraduate students attending graduate school within Texas (panel (c)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, calculated as the average across all courses in the student's transcript for which we observe a syllabus. In column 3 of panels (a) and (b) and in panel (c), the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. In panel (b), the sample is restricted to students who graduate. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below $20K, between $20-40K, between $40-60K, between $60-80K, and above $80K. Columns 2-6 additionally control for the lagged average proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table A4: Frontier Knowledge Proximity and Students' Earnings - Not Controlling for Instructor Fixed Effects

Panel (a): Short-run earnings	(1) Full sample	(2) Full sample	(3) Ugrad	(4) Grad	(5) Females	(6) Males
Proximity (sd)	0.117***	0.076***	0.065***	0.064***	0.071***	0.084***
	(0.005)	(0.006)	(0.008)	(0.010)	(0.009)	(0.010)
Adj. R^2	0.14	0.14	0.10	0.34	0.10	0.16
N	136,112	136,112	116,245	19,867	73,333	60,821
Panel (b): Long-run earnings	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.109***	0.072***	0.070***	0.062***	0.055***	0.092***
	(0.007)	(0.009)	(0.010)	(0.016)	(0.014)	(0.016)
Adj. R^2	0.15	0.15	0.11	0.27	0.11	0.16
N	63,893	63,893	54,295	9,598	34,546	28,078
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. OLS estimates; one observation is a student. The dependent variable is the logarithm of average quarterly earnings 1-3 years after each student's predicted graduation year (panel (a)) and the logarithm of average quarterly earnings 4-6 years after each student's predicted graduation year (panel (b)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, calculated as the average across all courses in the student's transcript for which we observe a syllabus. Columns 2-6 additionally control for the lagged average proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. The sample is restricted to students with observed earnings greater than \$1,000 for at least one quarter post-predicted graduation year. In column 3, the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table A5: Frontier Knowledge Proximity and Students' Educational Attainment - Alternative Controls for Lagged Proximity

	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Graduation	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.117***	0.076***	0.065***	0.064***	0.071***	0.084***
	(0.005)	(0.006)	(0.008)	(0.010)	(0.009)	(0.010)
Mean of Dep. Var.	0.77	0.78	0.77	0.80	0.75	0.77
Adj. R^2	0.14	0.10	0.10	0.34	0.11	0.16
N	136,112	136,112	116,245	19,867	73,333	60,821
Panel (b): Major persistence	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.109***	0.072***	0.070***	0.062***	0.055***	0.092***
	(0.007)	(0.009)	(0.010)	(0.012)	(0.014)	(0.013)
Mean of Dep. Var.	0.56	0.56	0.53	0.75	0.56	0.56
Adj. R^2	0.15	0.15	0.11	0.27	0.11	0.16
N	63,893	63,893	54,295	9,598	34,546	28,078
Panel (c): Grad. school attendance	Undergraduate				Females	Males
Proximity (sd)	0.034***	0.045***			0.048***	0.043***
	(0.002)	(0.003)			(0.003)	(0.005)
Mean of Dep. Var.	0.15	0.15			0.17	0.13
Adj. R^2	0.10	0.10			0.10	0.10
N	397,185	397,185			212,612	182,973
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is an indicator for students who graduate from their program (panel (a)), an indicator for students who graduate with the same major initially declared upon enrollment (panel (b)) and an indicator for undergraduate students attending graduate school within Texas (panel (c)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, residualized using instructor fixed effects and calculated as the average across all courses in the student's transcript for which we observe a syllabus. In column 3 of panels (a) and (b) and in panel (c), the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. In panel (b), the sample is restricted to students who graduate. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. Columns 2-6 additionally control for the 25th, 50th, and 75th percentiles of the lagged proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table A6: Frontier Knowledge Proximity and Students' Earnings - Alternative Controls for Lagged Proximity

	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Short-run earnings	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.046***	0.030***	0.022***	0.035***	0.028***	0.034***
	(0.007)	(0.007)	(0.009)	(0.011)	(0.010)	(0.011)
Adj. R^2	0.14	0.14	0.09	0.34	0.10	0.16
N	135,459	135,459	115,837	19,622	73,045	60,550
Panel (b): Long-run earnings	Full sample		Ugrad	Grad	Females	Males
Proximity (sd)	0.091***	0.069***	0.075***	0.051***	0.063***	0.089***
	(0.009)	(0.010)	(0.012)	(0.016)	(0.014)	(0.015)
Adj. R^2	0.14	0.15	0.11	0.27	0.11	0.15
N	63,436	63,436	53,975	9,461	34,354	27,917
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is the logarithm of average quarterly earnings 1-3 years after each student's predicted graduation year (panel (a)) and the logarithm of average quarterly earnings 4-6 years after each student's predicted graduation year (panel (b)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, residualized using instructor fixed effects and calculated as the average across all courses in the student's transcript for which we observe a syllabus. The sample is restricted to students with observed earnings greater than \$1,000 for at least one quarter post-predicted graduation year. In column 3, the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. Columns 5 and 6 show estimates for female and male students, respectively. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. Columns 2-6 additionally control for the 25th, 50th, and 75th percentiles of the lagged proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table A7: Frontier Knowledge Proximity and Selection into the Earnings Sample

Panel (a): No earnings	(1) Full sample	(2) Full sample	(3) Ugrad	(4) Grad	(5) Female	(6) Male
Proximity (sd)	0.058***	0.054***	0.080***	0.007	0.057***	0.052***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
Mean Dep. Var.	0.38	0.38	0.38	0.36	0.36	0.40
Adj. R^2	0.51	0.51	0.54	0.32	0.53	0.50
N	459,122	459,122	396,182	62,940	241,911	215,012
Panel (b): No-earnings spell $\geq$ 4 qtrs	Full sample		Ugrad	Grad	Female	Male
Proximity (sd)	-0.003	-0.002	-0.001	-0.005	0.004	-0.009
	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
Mean Dep. Var.	0.11	0.11	0.11	0.16	0.11	0.12
Adj. R^2	0.06	0.06	0.05	0.09	0.06	0.06
N	285,166	285,166	244,937	40,229	153,616	129,410
Controls:						
School-major-year FE	X	X	X	X	X	X
Lagged proximities		X	X	X	X	X
Socio-demographics	X	X	X	X	X	X

Note: OLS estimates; one observation is a student. The dependent variable is an indicator for never appearing in the earnings sample with earnings greater than \$1,000 (panel (a)) and an indicator for having a no-earnings spell longer than four consecutive quarters, preceded and followed by quarters with earnings greater than \$1,000 (panel (b)). The variable *Proximity* is the standardized frontier knowledge proximity experienced by each student, residualized using instructor fixed effects and calculated as the average across all courses in the student's transcript for which we observe a syllabus. The sample is restricted to students with observed earnings greater than \$1,000 for at least one quarter post-predicted graduation year. In column 3, the sample is restricted to undergraduate students. In column 4, the sample is restricted to graduate students. Columns 5 and 6 show estimates for female and male students, respectively. Panel (b) is restricted to students who appear in the earnings sample at least once. All specifications control for school-by-major-by entry cohort fixed effects and for indicators for gender, race/ethnicity, quintiles of SAT/ACT scores, and for having parental income below \$20K, between \$20-40K, between \$40-60K, between \$60-80K, and above \$80K. Columns 2-6 additionally control for the lagged average proximity of courses taken and not taken by the student, calculated using the proximity of each course in the previous year. Robust standard errors in parentheses. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Appendix B Dataset Construction

B.1 Syllabi

We gathered our sample of syllabi from two sources. The first are the websites of seven public universities in Texas, which make these documents available for download (Appendix Table BVIII). We downloaded these documents in October-December 2023.The second is the records of the Open Syllabus Project (OSP),[25] containing 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 OSP 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;

[25] https://opensyllabus.org

(iii) Taught between 1998 and 2018;

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

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

Table BVIII: Texas Syllabi: General Information

Institution	Years available	Link
Sam Houston State U	2012-24	https://samweb.shsu.edu/faci10wp/
Stephen F. Austin State U	2010-23	https://orion.sfasu.edu/courseinformation/
Texas A&M U	2014-24	https://howdy.tamu.edu/uPortal/p/public-class-search-ui.ctf1/max/render.uP#
U of Texas at Austin	2011-23	https://utdirect.utexas.edu/apps/student/coursedocs/nlogon/?year=&semester=&department=GEO&course_number=420K&course_title=&unique=&instructor_first=&instructor_last=&course_type=In+Residence&search=Search
U of Texas at Dallas	2011-23	https://coursebook.utdallas.edu/
U of Houston–Clear Lake	2010-24	https://saprd.my.uh.edu/psc/saprd/EMPLOYEE/HRMS/c/UHS_SS_CUSTOM.UHS_HB2504_DISPLAY.GBL?institution_name=UHCL&
West Texas A&M U	2012-24	https://syllabus.wtamu.edu/syllabi/

Note: List of public universities in Texas included in our sample, with years of syllabi availability and links to webpages containing the syllabi. We downloaded syllabi in October-December 2023.

Course catalog data To complement the OSP 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.

Table BIX: Open Syllabus Project Syllabi: Summary Statistics

	# 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).

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

For the OSP sample, the full text of a syllabus is contained in the variable `text` of the OSP database. For the Texas sample, we extracted the text from the PDF files downloaded from university websites. To transform text into usable content, we (i) clean it by removing html language (which occasionally gets left over from web scraping) and 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[26] to convert Unicode text into ASCII text. This includes

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

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 BX 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 BVI, we use part of a syllabus as an example to present our process step by step.

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 BVI, 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 BVI, we focus on highlighted sections, such as "Course Objective," "Prerequisites," and "Text".

Table BX: 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.

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).[27] 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[28] 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*[29]; (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 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[30] (Sanh et al., 2019), accessed via the transformers library.[31] The out-of-sample prediction precision is 85.04%.

[27] BiLSTM-CNNs-CRF model for named entity recognition (NER), Ma and Hovy (2016).
[28] https://spacy.io/
[29] https://github.com/philipperemy/name-dataset
[30] https://arxiv.org/abs/1910.01108
[31] https://huggingface.co/transformers/index.html

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
- `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; courses in the Texas data are categorized using the National Center for Education Statistics' Classification of Instructional Programs (CIP).[32]. For some of our analyses, we group fields into macro-fields. The grouping is illustrated in Table BXI.

[32] See https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=56 for additional details

Table BXI: Categorization of Course (Macro-)Fields

Macro-field	OSP Fields	Texas Fields
Business	Business, Accounting, Marketing, Public Administration	Business Management
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	Comm. Technologies, English, Gender Studies , Human Sciences Journalism, Liberal Arts Library Science, Linguistics , Philosophy
Science	Mathematics, Biology, Chemistry, Physics, Earth Sciences, Astronomy, Atmospheric Sciences, Dentistry, Medicine, Nutrition, Nursing, Veterinary Medicine, Natural Resource Management	Biology, Health Professions, Mathematics, Multidisciplinary Studies, Natural Resources, Physical Sciences
Engineering	Computer Science, Engineering Architecture, Agriculture Basic Computer Skills, Engineering Technician, Transportation	Agricultural Science, Architecture, Engineering, Engineering Technicians, IT, Science Tech
Social Sciences	Psychology, Political Science, Economics, Law, Social Work, Geography, Education, Linguistics, Sociology Education , Criminology	Education, History, Legal Studies, Psychology, Public Administration, Social Sciences
Other	Fitness and Leisure, Basic Skills, Mechanic / Repair Tech, Cosmetology, Culinary Arts, Health Technician, Public Safety, Career Skills, Construction, Military Science	Fitness/Parks/Rec, Law Enforcement Visual/Performing Arts

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

Figure BVI: 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 our measure of frontier knowledge proximity, 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.[33] 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.[34] 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:[35]

- `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.

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

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

[35] 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).[36] 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 OSP 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.

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

B.4 National Science Foundation and National Institute of Health Grants

We collect information on grants awarded by the National Science Foundation (NSF)[37] and the National Institutes of Health (NIH)[38] 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.4.1 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).[39] This method sometimes leads us to match a NSF/NIH 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.

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

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

[39] 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.

B.4.2 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[40], 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.

(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.

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

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.

Appendix C Calculating Frontier Knowledge Proximity: 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 frontier knowledge proximity.

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.[41] 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.[42]

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.

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

[42] 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.

Appendix D Texas ERC Data: Sample and Variable Definitions

B.1 Sample Definition

We construct our analysis sample as follows. We start by considering all students ever enrolled in one of our seven institutions during the years for which we have syllabi data: Stephen F. Austin State University, Sam Houston State University, Texas A&M University, University of Houston-Clear Lake, University of Texas at Austin, University of Texas at Dallas, and West Texas A&M University. We further restrict attention to students for whom we observe at least one course syllabus, based on the student's transcript. We consider students working towards undergraduate and graduate programs, in all fields.

B.2 Variable Definitions

Academic Program Entry Year Enrollment records report, for each student and year, the degree year the student is registered for (e.g., sophomore). We thus assign each student a program entry year by considering the year when they first appear in the data and the earliest observable information on degree year (for example, if a student first appears in the data in 2011 as a sophomore, we assign 2010 as the program entry year).

Major We define majors using 4-digit CIP codes. Since graduation major information is only available for students who graduate, we use information on a student's declared major upon enrollment, available for 98.6% of all students.

Gender, Race, Family Income We control for gender with an indication for females. We control for indicators for race and ethnicity (Black, Asian, Native American, Pacific Islander, and unknown) and for whether the student is international. Lastly, we control for indicators for family income below $20K, between $20-40K, between $40-60K, between $60-80K, and above $80K.

SAT Test Scores 64% of all students in our sample report a SAT/ACT score. We control for quintiles of the standardized test score distribution within our sample and for an indicator for this variable being missing.

Graduate school attendance We define a student as attending graduate school if we see them enrolling in a graduate program in a Texas public university.

Earnings We consider earnings information for all quarters in which this variable is greater than $1,000. To calculate total earnings over a time span since graduation (for example 1-3 years since graduation), we consider a predicted graduation year equal to t+5 for undergraduates and t+1 for graduate students.