

Frontier Knowledge in College and Student Success*

Barbara Biasi[†] Song Ma[‡]

March 24, 2026

Abstract

We study whether exposure to frontier knowledge in college helps students succeed. Applying text analysis to 459,410 course syllabi and 107 million academic publications, we develop a measure of “frontier knowledge proximity” and show that course content varies substantially, even within the same institution, driven primarily by instructor heterogeneity. Linking syllabi to individual student records and exploiting course updates unobserved to students at enrollment, we find that frontier exposure helps students graduate on time, raises GPA, leads students to attend graduate school, and raises earnings. The distributional pattern, however, is asymmetric: frontier knowledge is most equalizing at the degree-completion margin, where disadvantaged students benefit most, but its returns to graduate school attendance and earnings rise with students’ pre-college resources. A mediation analysis confirms that frontier exposure operates through improved attainment for the disadvantaged, while generating additional labor-market premia for the advantaged—revealing complementarity between frontier content and pre-college resources.

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

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

*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, David Autor, Joe Altonji, Raj Chetty, Ben Sprung-Keyser, Zhengren Zhu, Seth Zimmerman, and seminar participants at various institutions and conferences 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

Frontier knowledge—the cutting edge of current research—is a well-established engine of economic growth (Romer, 1986). Its production generates the new ideas that fuel innovation (Jones, 2009; Iaria et al., 2018); its diffusion to workers and entrepreneurs transforms those ideas into economic value (Goldin and Katz, 2010; Acemoglu and Autor, 2011). Research universities sit at the center of this process. Through their faculties, they produce a large share of frontier research (Andrews, 2023). Through their courses, they have the potential to transmit that knowledge to hundreds of thousands of students each year. Yet whether this second function is fulfilled, and whether it matters for student outcomes, has remained an open empirical question, largely because course-level data on what is actually taught do not exist in standard administrative records.

The answer to these questions is *ex ante* unclear for two reasons. First, the design of course content is largely delegated to instructors, and even within the same institution and major, syllabi can differ sharply in how quickly they incorporate new research.¹ Second, the effects of frontier knowledge exposure on students are not obvious: bringing courses closer to the frontier may increase engagement and learning by making material more relevant and by building skills valued in modern labor markets, but it may also crowd out foundational knowledge or raise difficulty, potentially harming students on fragile completion margins. Answering these questions requires measuring course content at scale, linking it to students’ course-taking histories and outcomes, and addressing students’ selection of courses when estimating causal effects.

This paper addresses these challenges with newly collected data and a novel empirical strategy. Applying text analysis to 459,410 syllabi from seven major public universities in Texas and to a corpus of over 107 million academic publications, we construct each course’s *frontier knowledge proximity*, the degree to which a course’s syllabus aligns with current scholarly research.² We link these syllabi to individual administrative records covering more than 135,000 students and exploit an institutional feature of Texas course registration for causal identification: students enrolling in a course can only observe the *previous* term’s syllabus, not the updated version. Changes in a course’s frontier proximity from one term to the next are therefore unobserved by students at the time of enrollment. Conditional on lagged proximity and instructor identity, these changes provide plausibly quasi-random variation in frontier exposure, which we use to estimate causal effects on educational

¹An 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).

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

attainment, academic performance, graduate-school enrollment, and labor-market earnings.

We develop an empirical strategy to isolate variations in students' course exposure to frontier knowledge that are due to random variation in their ability. Our strategy exploits an institutional feature of Texas course registration: students enrolling in a course are randomly assigned to different sections of the course, which we use to estimate causal effects on educational attainment, academic performance, school enrollment, and labor – market earnings.

We define frontier knowledge proximity as the ratio of a syllabus's average textual similarity to recent research articles to its average similarity to older research articles. To compute similarities, we represent each document—a syllabus or a publication—as a weighted term-frequency vector using a backward-looking inverse-document-frequency scheme (Kelly et al., 2021) that up-weights terms currently rare yet distinctive of a document's content and down-weights common terms. Cosine similarities between syllabi and publications are then averaged over recent and older publication vintages, with the boundary between “recent” and “old” defined by field-specific citation-lag percentiles to account for cross-field variation in the pace of knowledge production. By construction, a higher proximity means a course covers knowledge that is *new relative to what was known previously*—not merely that it covers any particular topic. Because our metric is a ratio, stylistic differences across syllabi (length, verbosity) tend to cancel, and the weighting scheme ensures that widely-taught foundational material is not mechanically penalized. Three validation checks confirm the measure captures what it is designed to capture: syllabi citing more recent references exhibit higher proximity; proximity increases monotonically with course level, from lower-undergraduate to graduate; and a simulation that progressively replaces “older” content terms with “newer” ones raises proximity monotonically.

Applying this metric to our Texas syllabi, we first document the extent of variation in frontier knowledge coverage across courses. The differences are large. Moving a syllabus from the 25th to the 75th percentile of the proximity distribution is equivalent to replacing roughly 30% of its content with newer knowledge. Strikingly, instructors and courses explain the largest shares of this variation (21% and 26%, respectively, from a Shapley-Owen variance decomposition), while differences across schools explain only 0.3% and differences across fields explain 5%. Even within the same institution, field, and course level, students face substantially different exposures to frontier knowledge depending on which courses they enroll in—a pattern that motivates the central causal question of the paper.

To answer that question, we exploit the Texas course-registration process: when students register, they can observe the prior term's syllabus and the instructor's name, but not the updated

content. Conditional on lagged course proximity and instructor identity, within-course updates in proximity are therefore unobserved by students at enrollment, rendering them conditionally exogenous. We implement this strategy by augmenting an outcome equation with controls for each student's average lagged course proximity and with instructor fixed effects, absorbing school-by-major-by-cohort fixed effects to account for the menu of courses available in each program and cohort. The identifying variation is thus within-program, within-instructor changes in course proximity that students could not have anticipated. This variation corresponds roughly to one third of the overall variation in proximity across all courses. Placebo tests using proximity measured two years after a student's enrollment yield coefficients indistinguishable from zero, supporting the conditional-independence assumption underlying our design.

We find that exposure to frontier knowledge improves all measured dimensions of educational attainment. Among undergraduate students, a one-standard-deviation (SD) increase in average course proximity raises the probability of completing a bachelor's degree by 1.7%, reduces time-to-degree by 0.4 years, raises GPA by 0.04 points, and increases the probability of graduate school enrollment by 16%. The GPA effects do not merely reflect more lenient grading: a dynamic specification shows that greater proximity in earlier courses predicts better grades in subsequent courses, indicating genuine cumulative skill accumulation. Among graduate students, the same increase raises Master's graduation rates and reduces time-to-degree for both Master's and PhD students by roughly 0.2 years.

The effects on educational outcomes are largest when frontier exposure occurs in the *first year* of college. A one-SD increase in first-year proximity substantially increases degree completion, more than in any subsequent year, consistent with early exposure to cutting-edge material shaping students' effort and motivation throughout their trajectory. Effects on degree completion and time-to-degree are driven by exposure to one or two very high-proximity courses within a standard curriculum. Effects on graduate school enrollment, by contrast, appear driven by an overall shift in the curriculum.

Across the ability distribution, frontier knowledge generates broad learning gains but translates into different outcome margins. GPA effects are roughly uniform across ability quartiles, consistent with frontier content improving academic performance for all students. Yet converting those gains into threshold attainment outcomes is ability-complementary. Students in the bottom ability quartile see no significant effect on degree completion, but benefit from faster degree progression, improved GPA, and higher graduate school attendance. Students in the top ability quartile, who can

more readily absorb frontier material, experience significant gains on all margins, including degree completion. Transfer students with no reported test scores—predominantly community-college entrants for whom degree completion is often precarious—display the largest graduation response of all, consistent with frontier exposure disrupting non-completion inertia on fragile margins, while showing no effect on graduate school attendance, suggesting that other constraints beyond academic preparedness bind for this group.

Heterogeneity by family income tells a parallel story. Students from the bottom income quartile benefit most on the graduation margin (2.1 pp, versus 1.2 pp at the top), but see no significant increase in graduate school attendance. Graduate school effects instead rise monotonically with family income, reaching 4.8 pp in the top quartile. GPA gains are positive across all income groups and increase with income, as do time-to-degree reductions. Frontier knowledge thus promotes attainment most vigorously where it is most at risk, while access to higher-threshold outcomes—particularly graduate education—continues to depend on the resources that more-advantaged students bring with them.

We next investigate students' labor market performance. Exposure to frontier knowledge significantly increases post-graduation earnings. A one-SD increase in course proximity raises undergraduate earnings by 2.8% on average over the first six years after graduation, with the effect growing over time (2.5% in years 1–3, 3.6% in years 4–6). Effects are larger for graduate students (5.3% overall). At the undergraduate level, the gains are concentrated at the bottom of the distribution: frontier exposure reduces the probability of bottom-quartile earnings by 2 percentage points, while leaving top-quartile odds unchanged. Graduate exposure, by contrast, shifts students across the full distribution. Critically, earnings gains arise almost entirely within industries rather than through sorting into higher-paying sectors.

Earnings returns to frontier exposure are substantially larger for higher-ability and higher-family-income students. A one-SD increase in proximity raises undergraduate earnings by 1.8% for students in the bottom ability quartile and by 10% for those in the top; returns also rise with family income, from 2.4% at the bottom quartile to roughly 5% at the top.

A mediation analysis reveals the mechanism behind this earnings gradient. For lower-income and lower-ability students, controlling for educational outcomes (graduation, GPA, time-to-degree, and graduate school enrollment) explains nearly all of the earnings effect: their wage gains are entirely driven by improved attainment. For higher-income and higher-ability students, a substantial earnings premium survives even after conditioning on those outcomes, consistent with frontier-

acquired skills commanding a direct return in the labor market that credentials alone cannot capture. Frontier knowledge exposure thus narrows socio-economic gaps in degree completion while leaving (and potentially widening) gaps in labor-market earnings, reflecting a fundamental complementarity between frontier content and pre-college resources.

In sum, our results document substantial heterogeneity in the presence of frontier knowledge across university courses and show that these differences carry meaningful consequences for student outcomes. The educational gains from frontier exposure are broadly distributed and particularly pronounced among students on fragile completion margins, whereas earnings gains are larger for students with stronger pre-college resources—a pattern that reveals both the equity potential and the equity limits of frontier knowledge diffusion.

Related literature This paper contributes to several strands of the literature. Most closely related is a body of research on heterogeneity in human capital production, focusing primarily on 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 programs (Hemelt et al., 2023; Li et al., 2021). We depart from this literature in three ways: we examine proximity to the knowledge frontier—a previously unmeasured dimension of course content—rather than field of study or institutional type; we document heterogeneity within the same institution and major rather than across them; and we develop a research design that exploits the information structure of Texas course registration to isolate exogenous variation in the content individual students encounter.

Our findings also speak to the longstanding debate over whether higher education creates skills or merely signals pre-existing ability (Becker, 1964; Spence, 1973; Arrow, 1973). The signaling hypothesis predicts that earnings gains from education should operate through credential-based sorting into better-paying jobs rather than genuine productivity increases. Two features of our results bear directly on this question. First, the earnings gains we document arise almost entirely within industries rather than through inter-industry mobility, consistent with genuine skill formation rather than credential-driven sorting. Second, our mediation analysis shows that for more-advantaged students a substantial earnings premium survives conditioning on all educational attainment outcomes—graduation, GPA, time-to-degree, and graduate enrollment—pointing to skills that command direct market returns beyond what credentials alone can explain. These patterns complement recent evidence that the human capital channel dominates measured returns to schooling (Aryal et al., 2022) and that curriculum content shapes wages beyond the credential it confers

(Arteaga, 2018). We extend both lines of inquiry by showing that it is not merely the level but the specific content of education that matters: courses closer to the knowledge frontier generate earnings gains that cannot be reduced to degree receipt.

Our work also relates to the literature on 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 attending a given institution on outcomes such as earnings. While valuable for ranking schools by the returns they generate, these estimates are difficult to translate into policy guidance: they treat institutional quality as a black box, offering no information on which dimensions of instruction drive positive returns, and they are backward-looking by construction—becoming available only years after the relevant instruction has occurred. By linking student outcomes directly to observed course content, our approach can identify which features of a program help students most, information that can inform curriculum design without delay.

Finally, 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 role in disseminating frontier knowledge, particularly in STEM fields (Baumol, 2005; Toivanen and Väänänen, 2016; Bianchi and Giorelli, 2019; Akcigit et al., 2025). We extend this work by locating the source of frontier exposure within observed course content rather than attributing it to institutions or fields as a whole, and by tracing how that exposure shapes educational attainment and labor-market outcomes across the socioeconomic and ability distribution.

2 Data

Our empirical analysis combines several distinct data sources: course syllabi, academic publications, individual-level student demographics, academic records, and earnings. We provide additional details on the construction of the final dataset in [Appendix B](#).

2.1 Course Syllabi

Our syllabi database covers the majority of courses taught at seven major public universities in Texas: Stephen F. Austin State University (starting in 2009), Sam Houston State University (2011),

³Moser 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.

Texas A&M University (2013), University of Houston-Clear Lake (2010), 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 459,415 documents corresponding to 27,872 courses taught through 2022.⁴ These syllabi represent approximately 52% of all courses offered at these institutions during our analysis period. We converted the downloaded PDF documents into machine-readable text, cleaned the extracted text, and parsed each syllabus into sections to isolate course content (see Appendix B.1.1).

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 undergraduate, or graduate (see Appendix B.1.3). Texas courses are assigned to a field following the National Center for Education Statistics’ Classification of Instructional Programs (CIP). For most analyses, we aggregate fields into four broad macro-fields: STEM, Humanities, Social Sciences, and Business (Appendix Table B19).

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.”⁵ These sections typically include course structure, main concepts, timelines, and reading materials.

Reference list We extract bibliographic information of required and recommended readings listed on each syllabus (textbooks, articles, and academic papers) using OpenAI’s GPT-4.1 mini. We successfully compile reference information for 80.4% of our syllabi sample; we extract at least one publication year for 42.0% of those syllabi.

Sample description Our syllabi and courses data are described in panel (a) of Table 1. The average undergraduate syllabus contains 393 unique knowledge terms (i.e., those belonging to a pre-defined dictionary aimed at capturing a document’s academic content, defined in greater detail in Section 3), while the average graduate syllabus contains 399 unique terms. Syllabi pertain to 18,994 undergraduate and 8,878 graduate courses. Most syllabi are from STEM fields (41% at the undergraduate

⁴We 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 the term of instruction.

⁵The full list of section titles used to identify each section is shown in Appendix Table B18.

and 42% at the graduate level); among undergraduate courses, 74% are upper-level (i.e., generally taken during or after the 3rd year).

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

| Panel a) Syllabi and courses | | | | |
|---|----------------|--------|-----------|--------|
| | Undergraduates | | Graduates | |
| | Mean | SD | Mean | SD |
| <i>Syllabi</i> | | | | |
| Unique knowledge words | 393.08 | 284.41 | 399.31 | 312.13 |
| Frontier knowledge proximity | 105.15 | 7.57 | 106.69 | 7.28 |
| N syllabi | 406,969 | | 52,446 | |
| <i>Courses</i> | | | | |
| Business | 0.067 | 0.251 | 0.090 | 0.286 |
| STEM | 0.409 | 0.492 | 0.428 | 0.495 |
| Social Science | 0.179 | 0.383 | 0.325 | 0.468 |
| Humanities | 0.299 | 0.458 | 0.134 | 0.341 |
| Upper level | 0.739 | 0.439 | | |
| # sections | 4.10 | 10.93 | 1.71 | 2.11 |
| # instructors | 2.62 | 4.45 | 1.48 | 1.67 |
| Frontier knowledge proximity | 106.04 | 6.65 | 107.16 | 7.24 |
| # courses | 18,994 | | 8,878 | |
| N (course * year) | 100,746 | | 30,781 | |
| Panel b) Students | | | | |
| | Undergraduates | | Graduates | |
| Variable | Mean | SD | Mean | SD |
| Female | 0.542 | 0.498 | 0.569 | 0.495 |
| Family income < \$40K | 0.251 | 0.434 | 0.261 | 0.439 |
| Family income > \$90K | 0.363 | 0.481 | 0.309 | 0.462 |
| <i>Field</i> | | | | |
| Business | 0.172 | 0.377 | 0.218 | 0.413 |
| STEM | 0.446 | 0.497 | 0.288 | 0.453 |
| Social Sciences | 0.209 | 0.406 | 0.378 | 0.485 |
| Humanities | 0.164 | 0.370 | 0.096 | 0.294 |
| <i>Courses</i> | | | | |
| # courses | 44.893 | 31.190 | 8.576 | 8.191 |
| Frontier knowledge proximity (in course-level SD) | -0.038 | 0.529 | 0.262 | 0.720 |
| <i>Outcomes</i> | | | | |
| GPA | 2.815 | 0.770 | 3.237 | 0.667 |
| Graduated | 0.838 | 0.369 | 0.593 | 0.491 |
| Time to degree | 5.881 | 1.515 | 2.478 | 1.640 |
| Enrolled in grad school | 0.126 | 0.332 | | |
| Quarterly earnings (\$1,000) | 15.83 | 16.57 | 20.48 | 19.43 |
| N (students) | 135,666 | | 31,482 | |

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

2.2 Academic Publications

We construct a database of peer-reviewed articles using OpenAlex, an openly licensed (CC0) catalog of global research. Maintained by OurResearch, OpenAlex indexes over 240 million scholarly works with approximately 50,000 new works added daily, drawing from multiple sources such as Crossref, PubMed, institutional repositories (e.g., arXiv), and the legacy Microsoft Academic Graph. The platform provides comprehensive metadata on works, authors, institutions, topics, and their interconnections. Our final dataset includes approximately 107 million articles with detailed information on titles, abstracts, keywords, authors, and author affiliations (see Appendix B.2 for additional detail on these data).

2.3 Student records

We link syllabi to individual-level administrative records on students enrolled at our seven public universities, provided by the Texas Education Research Center (ERC). Records from the Texas Higher Education Coordinating Board (THECB) provide demographic data (gender and race), parental income (from FAFSA), SAT/ACT scores, university enrollment and degree completion records (starting year, type of degree, initially declared major, graduation outcome, and graduation year), and full academic transcripts. Records from the Texas Workforce Commission (TWC) provide 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 section identifiers, 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.

In our linked data, about 400,000 students with non-missing syllabus information are enrolled at the universities in our sample. Our final analysis sample covers 135,666 full-time students for whom we observe earnings greater than \$1,000 for at least one quarter between one and six years after their predicted graduation year (defined as six years after degree start for undergraduates and three years after start for graduate students; we consider quarterly earnings values below \$1,000 as missing).⁶ We assign each student a degree-specific start year corresponding to the first year in which the student appears in the enrollment records as working towards that degree. We also assign students a major based on the two-digit CIP code declared upon enrollment.

⁶Our results on educational outcomes remain robust if we include all students for whom we have syllabi information, regardless of whether we observe their earnings.

Panel (b) of Table 1 describes our student sample. Approximately 54% of all students are female, 25% have family income below \$40K, and 30-36% above \$90K. Most undergraduate students (45%) are in STEM fields, while most graduates (38%) are in the Social Sciences. On average, undergraduate students enroll in 45 courses and 84% of them graduate, with a mean time-to-degree of 5.9 years. 13% enroll in a graduate program at a Texas public university. Graduate students enroll in 8.6 courses on average and 59% of them graduate, with a mean time-to-degree of 2.5. One to six years after predicted graduation, undergraduate students earn \$15,830 per quarter; graduate students earn \$20,480. Additional details on the student data are in [Appendix D](#).

3 Measuring Frontier Knowledge Proximity

We use the text of syllabi and academic publications to construct a course-level measure of frontier knowledge proximity. Our goal is to capture the relative similarity between a syllabus and different vintages of research knowledge: a syllabus has higher proximity to the frontier if it is more similar to recent research than to older research. We now detail the steps to construct this measure and present a series of checks to validate it using our data. Additional details on the construction of the measure are provided in [Appendix C](#).

3.1 Similarity Between Syllabi and Academic Publications

Our procedure largely follows [Biasi and Ma \(2022\)](#). We start by representing each document d (a syllabus or an article) as a term-frequency vector \mathbf{TF}_d of length $|W|$, where W is a list of all the terms we consider. Each element TF_{dw} of \mathbf{TF}_d is 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 on academic knowledge, we construct W as the list of all unique terms used as keywords in our academic publications sample.⁷

Adjusting for Term Relevance A standard issue with term frequency vectors is that terms commonly used across *all* documents mechanically receive higher weight, regardless of their ability to capture the content of each specific document. For example, the term “Programming” appears frequently in Computer Science syllabi but is less informative than “Natural Language Processing.” Similarly, “Animals” is common in Biology syllabi but less informative than “CRISPR” (a

⁷Our results are robust to considering all terms with an English Wikipedia webpage as of 2019.

gene-editing technology).⁸ The text analysis literature addresses this issue with term-frequency-inverse-document-frequency (TFIDF) weighting (Kelly et al., 2021), implemented by multiplying each element in a word vector by the inverse of the term’s frequency across all documents:

$$TFIDF_{dw} \equiv TF_{dw} \times IDF_w, \quad (1)$$

where IDF_w is the inverse ratio of the count of all documents containing word w , $\sum_{n \in D} \mathbb{1}(c_{nw} > 0)$, and the count of all documents, $|D|$:

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

Accounting for Changes in Term Relevance Over Time Since our goal is to capture novel content in course syllabi, an unappealing feature of traditional TFIDF weights is that they ignore temporal changes in term relevance. Terms that became popular only recently receive low weights even in older documents due to their current widespread use. Consider course CS229 at Stanford University, taught by Andrew Ng in the early 2000s as one of the first courses entirely focused on machine learning. The term “machine learning” has become very popular in later years, so its frequency across all documents is high and its IDF_w is low. Pooling documents from different years would thus assign a low $TFIDF_{dw}$ to “machine learning” in this course’s syllabus, failing to recognize its novelty in the early 2000s.

We address this issue by adjusting the TFIDF weighting scheme to incorporate a backward-looking IDF measure, $BIDF_{tw}$, based only on documents published prior to year t :

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

where D_t is the set of documents published *prior to* t . Using $BIDF_{tw}$, we construct a term-frequency-backward-inverse-document-frequency vector \mathbf{TFBIDF}_d , with elements:

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

where $t(d)$ denotes the publication year of document d . In our example, “machine learning” would have a higher $BIDF$ in the early 2000s than in 2018 (when the term was widely used), appropriately

⁸CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, a family of DNA sequences found in prokaryotic organisms and the basis for gene-editing technology.

assigning a higher *TFBIDF* to course CS229 and better capturing its novelty.

Textual Similarity Between Documents Equipped with weighted vectors \mathbf{TFBIDF}_d , we calculate the textual similarity between pairs of documents d and d' as the cosine similarity between their weighted word vectors (for simplicity, we denote \mathbf{TFBIDF}_d as \mathbf{V}_d):

$$\rho_{d,d'} = \frac{\mathbf{V}_d \cdot \mathbf{V}_{d'}}{\|\mathbf{V}_d\| \cdot \|\mathbf{V}_{d'}\|} \quad (3)$$

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 $[0, 1]$. 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 Frontier Knowledge Proximity

To capture the similarity between each syllabus d and different vintages of research, we calculate the average similarity of d with all articles published in a three-year window centered 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. This measure is the inverse of the *education-innovation gap*, which we first proposed in [Biasi and Ma \(2022\)](#).⁹

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

$$p_d \equiv 100 \times \left(\frac{S_d^\tau}{S_d^{\tau'}} \right). \quad (4)$$

A syllabus taught in year t thus has higher proximity if it is more similar to recent research than to older research.

In setting the parameters τ and τ' , which define new and old knowledge vintages, we account for field-specific rates of knowledge production. In fields where knowledge production is fast, both τ and τ' are small; in fields where it is slower, τ' (and possibly τ) are large. We define τ and τ' as the 5th and 90th percentiles of the distribution of citation ages within each field across all articles in that field. The median recent knowledge vintage (5th percentile) is approximately 2 years, while

⁹Our main analysis uses three-year intervals; results are robust to one-year or two-year intervals.

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 appealing properties. First, as a ratio, it is less sensitive to stylistic differences across syllabi (such as length or verbosity) that could otherwise affect similarity scores. Two courses covering the same materials could have different similarities to research publications if one syllabus is more detailed or uses more academic terms. Using a ratio attenuates such variations since both the numerator and denominator are subject to the same style differences. Second, by employing the *TFBIDF* weighting scheme, the measure appropriately reduces the influence of common foundational terms. This ensures that syllabi covering widely taught, fundamental, but old concepts (i.e., the “classics” of a field, such as *Ordinary Least Squares* in applied microeconomics) are not mechanically favored or penalized by raw term prevalence.

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

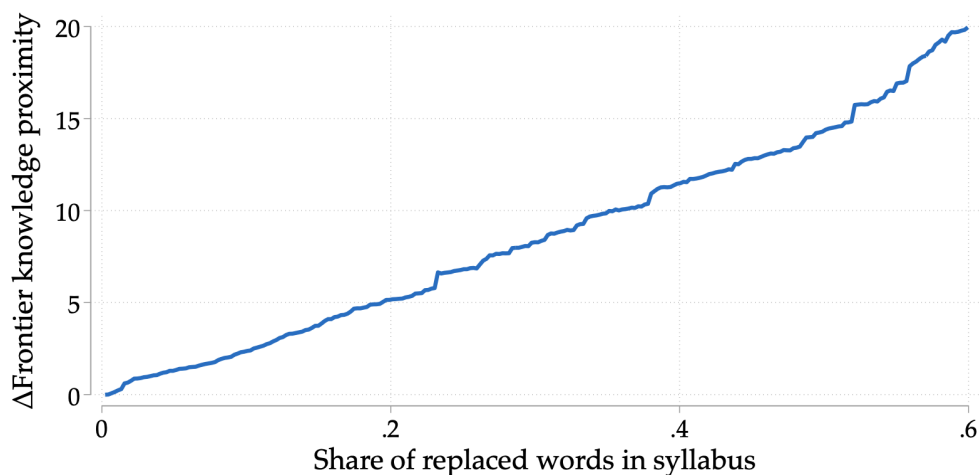
Reference age We verify that frontier proximity correlates negatively with the average age of a syllabus’s references, calculated as the average difference between the syllabus year and the publication years of cited sources (Figure A2, panel (a)). The correlation is modest in magnitude (-0.01). This weak relationship is unsurprising, as many syllabi primarily reference textbooks whose publication dates might not accurately reflect course content. While reference age is easy to compute, our text-based measure better captures actual course content and is available even for syllabi that lack explicit references or rely primarily on a single textbook.

Differences by course level We confirm that proximity varies as expected across course levels. Graduate courses and upper undergraduate courses have higher proximity scores than lower undergraduate courses (Figure A2, panel (b)), consistent with the expectation that more advanced courses incorporate more frontier knowledge.

Simulation exercise We use a simulation to demonstrate that our measure can detect incremental changes in a syllabus’s coverage of knowledge vintages. In each syllabus, we progressively replace terms that appear more frequently in older articles (“old terms”) with terms that appear more frequently in recent articles (“new terms”), then recalculate proximity at each incremental replacement. To ensure a monotonic increase, we use a recursive structure: to replace n terms, we sample from the top $2n$ terms with the highest “oldness” scores, ensuring the set of n terms contains all terms from the $n - 1$ step. We define old terms 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. We define new terms symmetrically as either (a) in the top 5% in terms of frequency in the new publication corpus, or (b) in the new corpus but not in the old corpus.

Figure 1: Updates in Syllabi Content and Changes In Frontier Knowledge Proximity



Note: This figure illustrates the median change in proximity when we manually replace “old” knowledge words with “new” knowledge words in a random sample of 100,000 syllabi.

Proximity increases monotonically as we replace more old terms with new ones. Figure 1 plots the median change in proximity across all syllabi for a given number of replaced terms. A one-SD increase in frontier knowledge proximity (approximately 7.5) is equivalent to replacing 25% of the old content in the median syllabus with new content.

3.3 Alternative Measure: N-Gram Based Frontier Knowledge Proximity

As an additional robustness check, we construct an alternative proximity measure using n-grams rather than dictionary terms. Specifically, we tokenize syllabus content and publication text into n-grams (unigrams and bigrams, and in additional checks trigrams), compute term frequencies, apply the same backward-looking inverse-document-frequency weighting, and calculate cosine similarities exactly as in equations (3) and (4). We then build the same recent-versus-old similarity ratio at the course level.

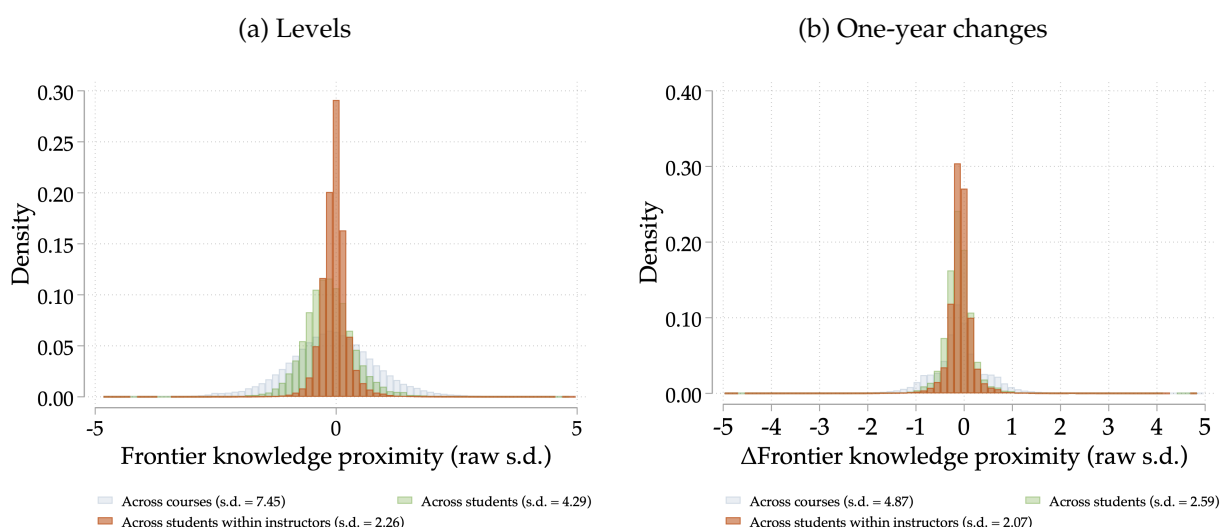
This approach relaxes dependence on a pre-specified keyword dictionary and allows the measure to capture frontier concepts expressed as short phrases. In our data, the n-gram-based measure is highly correlated with the baseline measure (correlation = 0.35) and yields qualitatively similar

reduced-form patterns and regression estimates for both educational and earnings outcomes (Appendix Table A15 and A16).

3.4 Differences in Frontier Knowledge Proximity Across Courses

We begin our analysis by examining the variation in frontier knowledge proximity across courses in our samples. On average, this measure equals 105.2, with a standard deviation of 7.55, a 25th percentile of 101.2, and a 75th percentile of 109.1. Figure 2 (panel a) shows the distribution of frontier knowledge proximity across courses.

Figure 2: Distribution of Frontier Knowledge Proximity



Note: Distribution of frontier knowledge proximity measures, in levels (panel a) and in changes from the previous year (panel b).

To give these numbers an economic interpretation, we rely on the simulation in Figure 1. 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.

3.4.1 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.¹⁰

The results of this decomposition exercise, shown in Appendix Table A1, indicate that differences across instructors and courses explain the largest portion of the overall variation in frontier knowledge proximity, respectively 21% and 26%. By contrast, field effects explain only 5%, and school effects contribute just 0.3%.

4 The Effects of Frontier Knowledge on Students: Research Design

Our descriptive analysis highlights significant differences in frontier knowledge proximity across courses. Do these differences matter for students’ outcomes? We begin by proposing an empirical strategy to isolate the causal effect of frontier knowledge exposure when students can select courses. The next section presents our results.

4.1 Empirical Model

The core empirical model we want to estimate is

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

where y_i denotes student i ’s outcome (e.g., average earnings 1-6 years post graduation) and p_i is the average frontier knowledge proximity of all courses on student i ’s transcript. The vector X_i includes individual characteristics that may affect outcomes, such as gender, race, family income, and prior academic achievement. The term θ_{smc} denotes school s -by-major m -by-starting cohort c fixed effects.

¹⁰The decomposition involves three steps: (i) estimate OLS models of proximity on all possible combinations of factors and record their adjusted R^2 ; (ii) re-estimate these models excluding one factor at a time and record the decrease in adjusted R^2 ; (iii) 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. For factor j , the partial- R^2 is calculated as $R_j^2 = \sum_{T \subseteq V \setminus \{j\}} \frac{|T|!(K-|T|-1)!}{K!} [R^2(T \cup \{j\}) - R^2(T)]$, where $R^2(S)$ is the adjusted R^2 of a regression of 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.

Our parameter of interest is β , which captures the causal effect of average course proximity p_i on student outcomes within groups of students who enter the same program in the same year. The main challenge in estimating β is that p_i is a function of students' course choices, which may be correlated with unobserved characteristics that also affect outcomes. For example, higher ability students may select courses with higher frontier knowledge proximity and earn more after graduation. This implies that $\mathbb{E}(p_i \epsilon_i) \neq 0$ and leads to bias in OLS estimates of β .

4.2 Identification Strategy

To overcome this challenge, we exploit an institutional feature that generates variation in p_i unobservable to students at the time of course selection. In the schools in our sample, students registering for a course do not have access to the latest updated syllabus. Instead, they can only see the syllabus from the prior term and the name of the instructor (see Appendix Table A2 for details). Consequently, conditional on the instructor, changes in a course's frontier knowledge proximity (Δp_i) are unobserved by students and therefore quasi-random from their perspective.

To formalize this argument, consider a student i choosing between two courses, A and B , during term t . Let $p_{A,t}$ and $p_{B,t}$ denote their frontier knowledge proximity and γ_A and γ_B be indicators for their instructors. The student's realized proximity, p_i , equals $p_{A,t}$ if the student enrolls in course A and $p_{B,t}$ if she enrolls in course B . Define the change in proximity for course k as $\Delta_{k,t} \equiv p_{k,t} - p_{k,t-1}$ for all $k \in \{A, B\}$. We can express the realized proximity as

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

where α_i represents student i 's unobserved preference for high-proximity courses and ν_i includes other unobserved factors that may affect course selection. In words, students' course choices depend only on α_i , ν_i , $p_{A,t-1}$, $p_{B,t-1}$, γ_A , and γ_B , whereas course content is determined by $p_{A,t-1}$, $p_{B,t-1}$, γ_A , γ_B , Δ_A , and Δ_B . Endogeneity arises if ν_i and α_i are correlated with the unobserved component of the outcome (denoted by ϵ_i in equation (5)).

Our identification argument relies on the *conditional* independence of p_i and ϵ_i given $p_{A,t-1}$, $p_{B,t-1}$, γ_A , and γ_B , the variables driving a student's course choice. Conditional on these values, any variation in actual proximity p_i is independent of α_i and ν_i and thus exogenous. In this simple example with two courses, we can thus consistently estimate β by augmenting the model in equation (5) with controls for these variables:

$$y_i = \beta p_i + \gamma X_i + \delta_{APA,t-1} + \delta_{BPB,t-1} + \theta_{s(i)m(i)c(i)} + \gamma_A + \gamma_B + \epsilon_i.$$

4.2.1 Implementation

In reality, students choose multiple courses from a large set of potential courses; the average undergraduate student in our data reports about 40 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 \mathcal{C}_i and the identity of i 's instructors. Assuming that all students in the same program and entering cohort face the same choice set, we can approximate \mathcal{C}_i with the set of all courses that students in the same school, major, and cohort ever took. The set \mathcal{C}_i is large: for the average student in our data, it contains about 40 courses. Directly including the lagged proximity of all courses in a student's choice set is thus infeasible. Instead, we summarize it with two aggregate measures: $p_{i,-1}$, the average lagged proximity of courses that the student takes, and p_{-1} , the average lagged proximity of all courses in the student's choice set, which in our empirical model is subsumed by the school-by-major-by-cohort fixed effects.

We can thus implement our identification strategy by augmenting equation (5) with controls for $p_{i,-1}$ and fixed effects for the instructors of i 's courses. Since instructor fixed effects are not mutually exclusive, we stack course-level observations pertaining to each student; our resulting dataset is thus at the student-course level. The estimating equation becomes:

$$y_i = \beta p_i + \delta p_{i,-1} + \gamma X_i + \gamma_{k(i)} + \theta_{s(i)m(i)c(i)} + \epsilon_{ik}, \quad (6)$$

where γ_k is a fixed effect for the instructor of course k . To account for variation in the number of courses taken across students, we weight observations by one divided by the number of courses each student takes, and we cluster standard errors at the student level.

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

Identifying variation Our research design uses variation in course proximity among students who enrolled in the same program in the same year *and* were taught by the same instructors. Since instructors explain most of the variation in proximity across courses (Table A1), one may wonder whether we have enough residual variation to identify causal effects. Figure 2 contrasts the distri-

bution of our raw proximity measure across all courses (in blue), across all students (thus giving more weight to courses with a larger class size, in green), and across students within school-major-cohort and instructor cells. The standard deviation is equal to 7.45 across courses, 4.29 across all students, and 2.26 across students within cells (panel (a)). The standard deviation of the proximity change across students within instructors is 2.07 (panel (b)). The simulation exercise in Figure 1 implies that the latter corresponds to a 10% change in the content of the average course taken by each student.

4.2.2 Identification assumptions

Our identification strategy rests on three main assumptions.

The first assumption is that $\Delta_{k,t}$ must be exogenous. This implies that students cannot predict syllabus updates when choosing courses. Even if students cannot see updated course syllabi at the time of enrollment (a possibility we rule out in Appendix A2), 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. To address this issue, we explicitly control for instructors in our main equation. 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 account for this possibility, we assign courses to students based on enrollment status on the first day of the term, regardless of whether the course was completed. Our estimates of β can thus be interpreted as intent-to-treat (ITT) effects of course proximity on outcomes. In our data, 28% of all students drop courses in a given year, with an average of one course dropped per student.

The second assumption is that course design is independent of student characteristics, 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 because syllabi are generally compiled before instructors meet their students. In support of it, in Appendix Table A3 we regress absolute course proximity ($p_{k,t}$, column 1) and the change in course proximity from the previous term ($\Delta_{k,t}$, column 3) on a host of characteristics of all students taking the course. While these characteristics appear correlated with $p_{k,t}$ — a clear indicator of endogenous course selection — they do not predict $\Delta_{k,t}$: the p-value of an F-test of joint significance is 0.69.

The third assumption, essential for interpreting our estimates as the effects of frontier knowledge exposure, is that changes in a course's proximity do not occur at the same time as changes in

other dimensions of content or instruction. This assumption could be violated if, when updating frontier content, instructors also change course design (for example, shifting evaluation from exams to reports or presentations) in ways that directly affect outcomes. It would also fail if course updates occur at specific points of instructors' careers when effort or teaching effectiveness is also changing. To assess these possibilities, we regress $p_{k,t}$ and $\Delta_{k,t}$ on syllabus characteristics (an indicator for whether the course requires an exam, and the shares of total assessment assigned to homework and reports) and instructor characteristics (the total number of instructors, whether any instructor is new to the course, and average instructor experience). Syllabus characteristics do not jointly predict proximity changes (F-statistic p-value = 0.23), whereas instructor characteristics do (F-statistic p-value = 0.0002). Our baseline specifications control for instructor fixed effects, which absorb time-invariant instructor heterogeneity. In robustness tests, we show that our results remain stable when we control semi-parametrically for instructor experience (Appendix Tables [A13](#) and [A14](#)).

5 Frontier Knowledge, Educational Attainment, and Performance

Using this empirical strategy, we now study the effects of exposure to frontier knowledge on educational outcomes during and after college. We focus on degree completion, time-to-degree, academic performance, and graduate school enrollment.

The effects of frontier knowledge on these outcomes are ambiguous *ex ante*. On the one hand, frontier knowledge may increase the perceived returns to effort and strengthen incentives to persist by making coursework appear more relevant to current research and real-world applications. It may also shorten time-to-degree by revealing the structure, demands, and potential payoffs of advanced coursework earlier in students' college careers, allowing them to better align effort and course choices. Lastly, exposure to frontier knowledge may directly enhance skill acquisition by deepening conceptual understanding, strengthening problem-solving abilities, and improving the capacity to apply knowledge to new contexts. On the other hand, frontier knowledge may negatively affect students by crowding out useful foundational knowledge or by making material too difficult. While we do not directly observe all these mechanisms, the empirical patterns we document provide suggestive evidence that frontier knowledge improves outcomes by enhancing motivation and promoting skill acquisition.

5.1 Average Effects of Frontier Knowledge on Educational Outcomes

5.1.1 Undergraduate Students

For the average undergraduate student, exposure to frontier knowledge in college has unambiguously positive effects on degree attainment. Panel (a) of Table 2 shows estimates of β in equation (6), with various educational outcomes as the dependent variables. A one-SD increase in average course proximity raises the probability of completing a bachelor's degree by 1.5 pp, or 1.7% relative to a mean of 90% (Table 2, panel a, column 1). It also helps students graduate on time by reducing time-to-degree by 0.40 years, or 7% (Table 2, panel a, column 2). In addition, frontier knowledge slightly improves students' performance during college, as captured by their GPA. A one-SD increase in proximity raises GPA by 0.04 points (Table 2, panel a, column 3).

A possible concern with these results is that, because frontier knowledge tends to be higher for more advanced courses, it may mechanically be higher for students who progress towards the latest stages of their degree, creating a spurious relationship with outcomes such as completion and time-to-degree. However, our identification strategy isolates the effects of changes in frontier knowledge (rather than levels) on outcomes. Furthermore, our estimates are largely unchanged if we restrict attention to the subsample of students who reach year 3 of their program (Appendix Table A4).

The positive effect of frontier knowledge on GPA could be driven by students learning more, or by courses with higher proximity being graded more leniently. To better isolate the effects on learning, we exploit the time dimension of students' transcripts and study whether exposure to frontier knowledge in earlier courses affects performance in subsequent courses. We modify our estimating equation as follows:

$$y_{i,k} = \tilde{\beta}p_{i,t(k)} + \delta p_{i,-1,t(k)} + \gamma X_i + \gamma_{k(i)} + \theta_{s(i)m(i)c(i)} + \epsilon_{ik}, \quad (7)$$

where $y_{i,k}$ is student i 's grade in course k , $t(i,k)$ is the year in which i takes course k , $p_{i,t}$ is the average proximity of all courses taken by student i prior to t , and $p_{i,-1,t}$ is its lagged counterpart. Everything else is as before.

Estimates for undergraduate students indicate that a one-SD increase in the average frontier knowledge proximity of prior courses raises grades in later courses by 0.011 points, or about 4% of the increase needed to move from a B to a B+; the estimate, however, is noisy (Table 2, panel a, column 4, p-value = 0.15).¹¹

¹¹See <https://catalog.utexas.edu/general-information/academic-policies-and-procedures/>

Exposure to frontier knowledge in college also makes students significantly more likely to attend graduate school. A one-SD increase in proximity raises the probability of graduate school attendance by 1.7 pp, or 16% relative to a mean rate of 0.11 (Table 2, panel a, column 5).

Taken together, these results indicate that exposure to frontier knowledge helps students not only to complete their degrees at higher rates, but also to do so more quickly and with modest improvements in academic achievement. The dynamic specification suggests that these performance gains are not solely mechanical artifacts of grading practices, but are at least partially consistent with cumulative learning. Furthermore, increased graduate school attendance indicates that frontier exposure may shift students' expectations about the returns to further education or increase their preparedness for advanced study.

5.1.2 Graduate Students

Frontier knowledge also benefits graduate students, particularly by shortening time-to-degree. For Master's students, a one-SD increase in proximity increases graduation rates by 2.0 pp (or 3% of the mean, equal to 0.73; Table 2, panel (b), column 1). It also reduces time-to-degree by 0.22 years (column 2), but it does not affect GPA (column 3). For PhD students, the same increase in proximity does not affect the likelihood of graduating (column 4) but reduces time-to-degree by 0.25 years (column 5) and raises GPA by 0.04 points (column 6). Overall, these results indicate faster progression in both programs, with GPA gains concentrated among PhD students.

computation-of-the-grade-point-average/ for information on the conversion of letter grades into grade points at UT Austin.

Table 2: Educational Effects of Frontier Knowledge Exposure

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|----------------------|---------------------|--------------------|----------------------|-------------------|
| | Graduates | Time-to degree | GPA | Future grade rank | Attends grad school | |
| | (1) | (2) | (3) | (4) | (5) | |
| Proximity (sd) | 0.015*** (0.004) | -0.399*** (0.019) | 0.036*** (0.009) | 0.005** (0.002) | 0.017*** (0.004) | |
| Mean dep. var. | 0.905 | 5.781 | 2.877 | 0.506 | 0.105 | |
| R ² | 0.294 | 0.552 | 0.292 | 0.102 | 0.223 | |
| N (student * course) | 5,993,879 | 5,443,026 | 5,989,673 | 2,079,811 | 5,993,879 | |
| N clusters (students) | 126,157 | 107,216 | 125,447 | 95,044 | 126,157 | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.020** (0.008) | -0.223*** (0.027) | 0.015 (0.012) | 0.009 (0.010) | -0.252*** (0.077) | 0.040* (0.022) |
| Mean dep. var. | 0.735 | 2.543 | 3.381 | 0.291 | 6.874 | 2.765 |
| R ² | 0.420 | 0.660 | 0.450 | 0.700 | 0.801 | 0.740 |
| N (student * course) | 199,970 | 182,734 | 199,467 | 47,380 | 17,672 | 47,238 |
| N clusters (students) | 22,312 | 19,616 | 22,132 | 4,343 | 2,203 | 4,313 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), the course grade (column 4 in panel (a)), and an indicator for enrollment in a graduate program within Texas (column 5 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations; in column 5 of panel (a), this variable is calculated as the average over all courses taken by the student prior to the focal course. All specifications control for average lagged proximity, instructor and school-major-cohort fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

5.2 Which Courses Drive the Returns to Frontier Exposure?

Our baseline estimates capture the effects of an increase in the average frontier knowledge proximity of all courses taken by each student. This model implicitly assumes that students benefit equally from all the courses they take. In reality, frontier knowledge may yield larger benefits if experienced at particular points in a student's college career. Additionally, a student may benefit from just one or two high-proximity capstones or advanced electives, even within a largely standard curriculum. To assess these two possibilities, we perform two complementary tests.

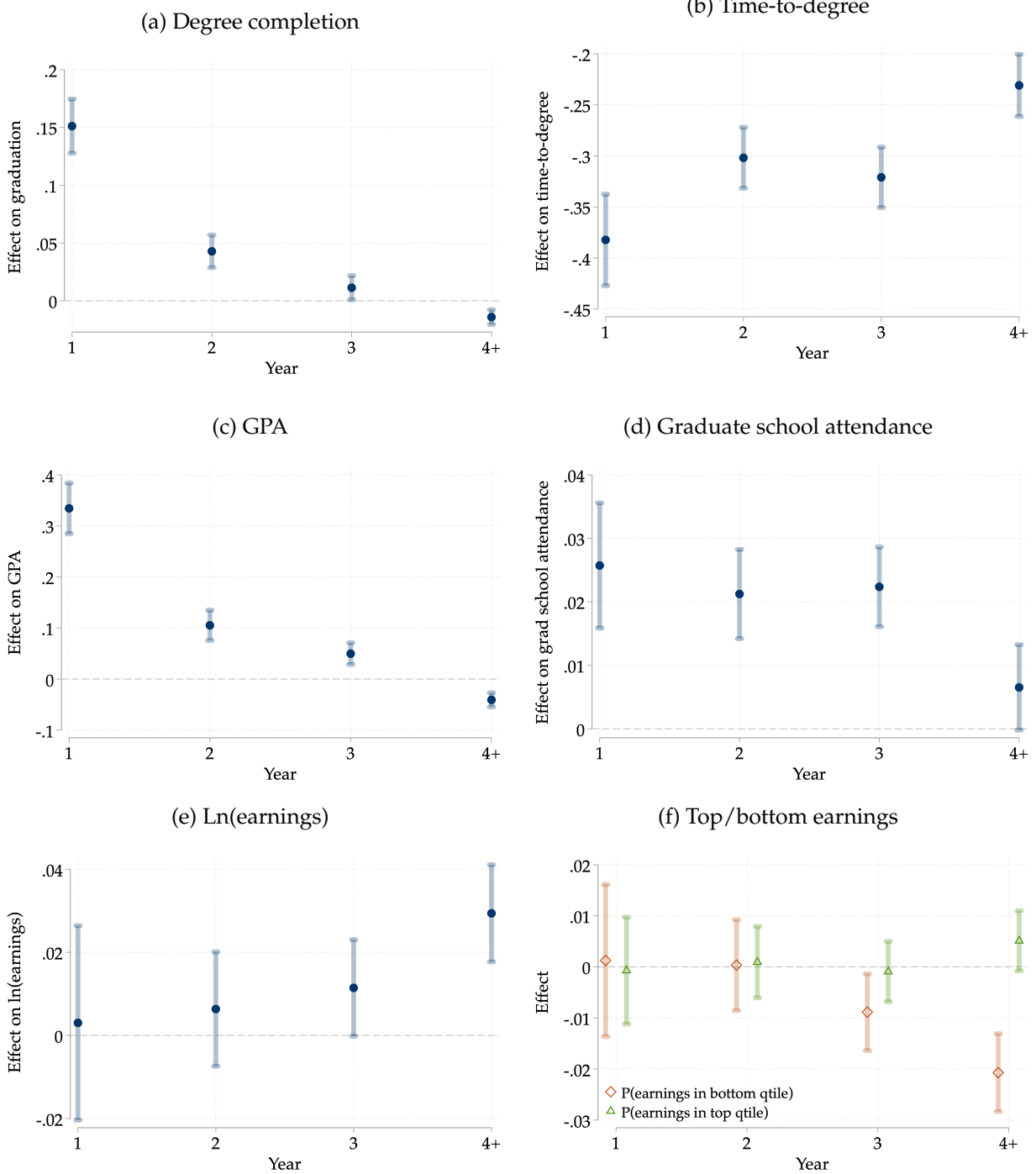
The first test investigates whether the effect of frontier knowledge on undergraduate students differs across the various stages of the college journey. To implement it, we allow both the variables p_i and $p_{i,-1}$ and the parameters β and δ in equation (6) to vary by year of instruction (pooling together all years after the third).

These estimates indicate that frontier knowledge has the largest positive effects when exposure happens early. Effects attenuate over time and, for some outcomes, become small or slightly negative. For example, a one-SD increase in proximity increases degree completion significantly more (by 15 pp) when experienced during the first year, whereas it has a smaller positive effect during the second and third year and a small negative effect during the last year (Figure 3, panel (a)). GPA effects display a similar pattern, with a much larger estimate for the first year (0.33 points) and a small negative effect in the later years (panel (b)). The effects of frontier knowledge on time-to-degree are negative and significant across years of instruction, but again larger in the first years (panel (c)), and so are those on graduate school attendance rates (panel (d)). These results indicate that the benefits of frontier knowledge are strongest in the early stages of an undergraduate student's trajectory.

The second test examines whether the positive effects of average frontier knowledge exposure come primarily from shifts in the entire curriculum or from the presence of one or two very high-proximity courses. To implement it, we calculate both the median proximity and the 90th-percentile proximity across all courses in a student's transcript and estimate the separate effects of these variables on educational outcomes within equation (6). This specification allows us to distinguish between broad exposure to frontier content and exposure driven by a small number of high-proximity courses.

Estimates from this specification on the sample of undergraduate students confirm that the effects of frontier knowledge exposure on degree completion come primarily from "star" courses. A one-SD increase in frontier knowledge proximity of the 90th-percentile courses in each undergraduate student's transcript raises degree attainment rates by 5.1 pp, or 5.7% of the mean (Table A7, column 1). By contrast, the effect of **median** proximity on graduation is negative at -3.0 pp: holding the presence of a high-proximity course fixed, a uniformly frontier-heavy curriculum actually reduces the probability of completion, suggesting that the benefit to graduation operates through concentrated peak experiences rather than broad exposure and that broad exposure may even harm students. We observe a similar pattern **for** GPA (column 3): 90th-percentile proximity raises GPA by 0.047 points, while median proximity has no significant effect. Results are similar (but magnitudes are smaller) for graduate students (columns 5 and 6).

Figure 3: Educational Effects of Frontier Knowledge Proximity, by Year



Note: OLS estimates; one observation is a student-course. The dependent variable is an indicator for students who graduate from their program (panel (a)), time-to-degree (panel (b)), GPA (panel (c)), an indicator for students ever enrolling in a graduate program within public schools in Texas (panel (d)), the natural logarithm of earnings 1 to 6 years after a student’s expected graduation year (panel (e)), and indicators for earnings in the top and bottom quartiles of a student’s cohort (column (f)). Each coefficient is an estimate of β in equation (6), where we allow proximity to vary by year of instruction. The sample is restricted to undergraduate students. Observations are weighted by one divided by the number of courses taken by each student. Standard errors in parentheses are clustered at the student level.

Table 3: Educational Effects of Frontier Knowledge Exposure - Median vs Top Courses

| Panel (a) Undergraduate students | | | | | | |
|---|----------------------|----------------------|---------------------|---------------------|----------------------|--------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Median proximity (sd) | -0.030*** (0.004) | -0.399*** (0.016) | -0.009 (0.008) | 0.018*** (0.003) | | |
| 90th pctile proximity (sd) | 0.051*** (0.003) | -0.027*** (0.010) | 0.047*** (0.005) | -0.001 (0.002) | | |
| Mean dep. var. | 0.905 | 5.781 | 2.877 | 0.105 | | |
| R ² | 0.302 | 0.551 | 0.293 | 0.223 | | |
| N (student * course) | 5,993,868 | 5,443,021 | 5,989,662 | 5,993,868 | | |
| N clusters (students) | 126,157 | 107,216 | 125,447 | 126,157 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Median proximity (sd) | -0.001 (0.008) | -0.129*** (0.026) | 0.001 (0.011) | 0.002 (0.010) | -0.418*** (0.078) | 0.010 (0.020) |
| 90th pctile proximity (sd) | 0.027*** (0.005) | -0.034** (0.016) | 0.016** (0.008) | 0.011 (0.009) | 0.129** (0.052) | 0.034** (0.016) |
| Mean dep. var. | 0.735 | 2.543 | 3.381 | 0.291 | 6.877 | 2.765 |
| R ² | 0.422 | 0.659 | 0.450 | 0.700 | 0.806 | 0.740 |
| N (student * course) | 199,995 | 182,775 | 199,487 | 47,395 | 17,677 | 47,255 |
| N clusters (students) | 22,309 | 19,612 | 22,128 | 4,343 | 2,198 | 4,313 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). *Median proximity* is the median frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. *90th pctile proximity* is the proximity of the 90th percentile course in each student's transcript, also measured in course-level standard deviations. All specifications control for average lagged proximity, instructor and school-major-cohort fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Examining effects on time-to-degree and the probability of attending graduate school tells a different story. For these outcomes, the strongest effect comes from median proximity, with top proximity having a smaller or insignificant effect. For example, a one-SD increase in median proximity, holding top proximity fixed, reduces time-to-degree by 0.4 years (column 2) and increases the chances of enrolling in graduate school by 1.8 pp (column 4). By contrast, an increase in top proximity holding median proximity fixed has a smaller effect on time-to-degree (-0.03 years; column 2)

and no effect on the probability of graduate school attendance (column 4). Overall, these estimates suggest that different outcomes load on different margins of frontier exposure.

Taken together, these results point to two mechanisms through which exposure to frontier knowledge helps students. The first is a mechanism consistent with increased motivation and engagement stemming from a few high-impact experiences (i.e., one or two high-proximity courses), which raise graduation odds and GPA. The negative coefficient on median proximity in these regressions reinforces this interpretation: a uniformly frontier-heavy curriculum without a standout experience to anchor student engagement raises the stakes of coursework without providing the motivational payoff, and can reduce the probability of completion. The second is a mechanism consistent with broader and more sustained skill accumulation and familiarity with research-oriented content, driven by higher median proximity across all courses. This is the margin that matters for faster degree progression and advancement to graduate school: advancing efficiently through the degree and qualifying for graduate study requires the entire curriculum to be oriented toward the frontier, not just a single exceptional course.

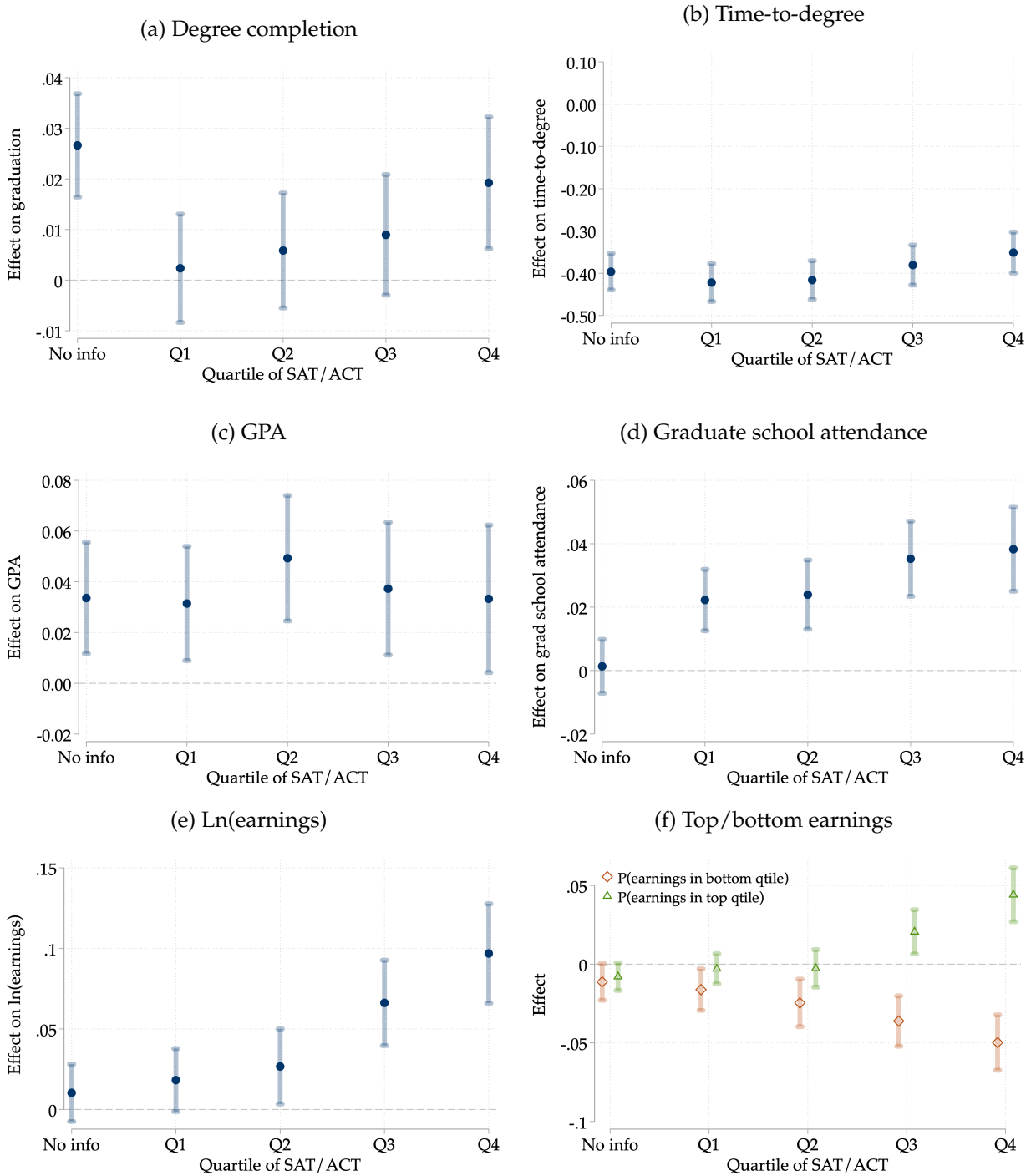
5.3 Differences by Baseline Ability

Our results so far show positive effects of frontier knowledge exposure for the average student. To understand whether these results apply to all students and to shed light on mechanisms, we now study whether these effects differ across students with different characteristics. We begin with baseline ability, by allowing the parameter β in equation (6) to vary flexibly by quartiles of the ACT/SAT score distribution. Approximately 27% of all undergraduate students in our sample do not have a reported score. This group is mainly composed of transfer students from community colleges, who are typically not required to submit a test score upon transfer. We include these students in our analysis as a separate “No Info” category.¹² This grouping lets us compare students with observed test-based ability while retaining a policy-relevant transfer population.

We find that exposure to frontier knowledge benefits students across the entire ability distribution, but along different margins. For students in the bottom quartile of the ability distribution, frontier knowledge has no effects on the odds of graduating (the point estimate of a one-SD increase in proximity is 0.002, with a p-value of 0.66; Figure 4, panel (a)). However, frontier knowledge reduces time-to-degree by 0.42 years, increases GPA by 0.034 points, and raises the likelihood of

¹²All 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: Educational Effects of Frontier Knowledge Proximity, by Baseline Ability



Note: OLS estimates; one observation is a student-course. The dependent variable is an indicator for students who graduate from their program (panel (a)), time-to-degree (panel (b)), GPA (panel (c)), an indicator for students ever enrolling in a graduate program within public schools in Texas (panel (d)), the natural logarithm of earnings 1 to 6 years after a student's expected graduation year (panel (e)), and indicators for earnings in the top and bottom quartiles of a student's cohort (column (f)). Each coefficient is an estimate of β in equation (6), allowed to vary by quartiles of the SAT/ACT scores. "No Info" refers to students without a test score. The sample is restricted to undergraduate students. Observations are weighted by one divided by the number of courses taken by each student. Standard errors in parentheses are clustered at the student level.

attending graduate school by 2.2 pp (Figure 4, panels (b)-(d)). By contrast, students in the top quartile of the ability distribution see significant improvements in degree attainment rates, while also experiencing increases in GPA and graduate school attendance rates. For students in the “No Info” group, a one-SD increase in frontier proximity has the largest effect on degree completion (2.7 pp), and also increases GPA and reduces time-to-degree, but has no detectable effect on the probability of attending graduate school. Notably, while degree attainment effects and graduate school attendance effects increase with ability, effects on time-to-degree become less negative (though remain significant) and GPA effects are roughly constant across the distribution.

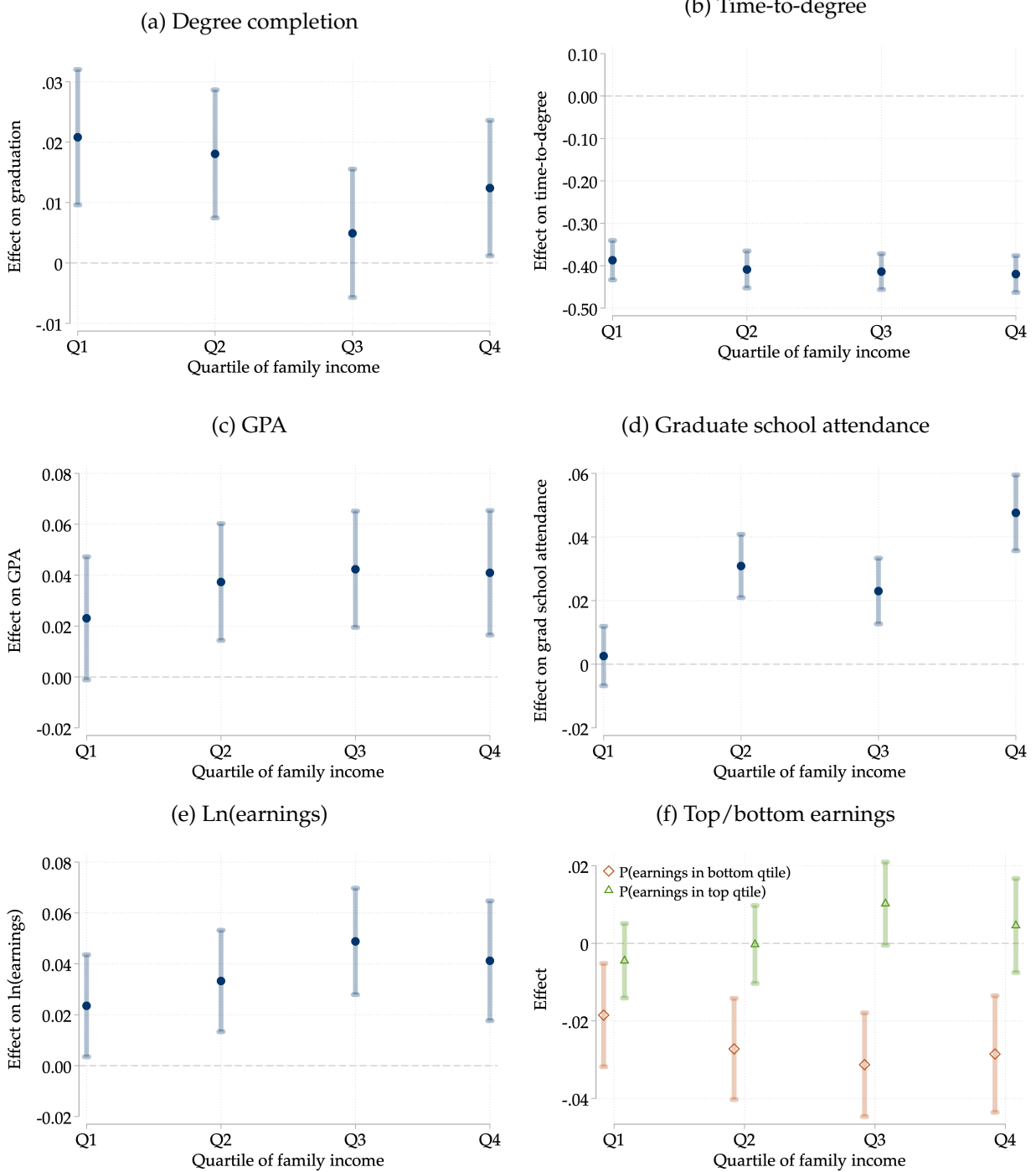
This pattern is consistent with broad learning gains, but heterogeneous conversion into longer-run attainment outcomes. The approximately uniform GPA effects suggest that frontier exposure improves academic performance across the distribution, although translating those performance gains into high-threshold outcomes is ability-complementary: higher-ability students are more likely to convert frontier exposure into degree completion and graduate-school progression. At the same time, the uniformly negative effects on time-to-degree suggest that frontier content improves academic progression broadly, especially for lower-ability students, even when it does not immediately raise completion rates. The strong completion response in the “No Info” group further indicates that frontier exposure can be particularly valuable on fragile completion margins, consistent with a role for motivation, academic focus, and course-to-course momentum among transfer students. By contrast, the absence of an effect on graduate-school attendance in this group suggests that progression to graduate education also depends on other constraints (e.g., information, advising, financing, or application readiness) that frontier exposure alone does not fully relax.

5.4 Differences by Family Income

The effects of frontier knowledge exposure may also vary by socioeconomic background. For example, lower-income students may be closer to fragile completion margins, while higher-income students may have more advising, financial flexibility, and research opportunities to convert the same exposure into stronger academic and postgraduate outcomes. Testing for this type of heterogeneity thus allows us to study both distributional implications and mechanisms. To do so, we allow the parameter β in equation (6) to vary by quartiles of baseline family income (from FAFSA records). This exercise asks whether frontier knowledge exposure is equally productive across students with different pre-college resources.

As with ability, our results indicate that frontier knowledge helps students across the whole family income distribution, but along different margins. Undergraduate students in the lower quartile

Figure 5: Educational Effects of Frontier Knowledge Proximity, by Family Income



Note: OLS estimates; one observation is a student-course. The dependent variable is an indicator for students who graduate from their program (panel (a)), time-to-degree (panel (b)), GPA (panel (c)), an indicator for students ever enrolling in a graduate program within public schools in Texas (panel (d)), the natural logarithm of earnings 1 to 6 years after a student's expected graduation year (panel (e)), and indicators for earnings in the top and bottom quartiles of a student's cohort (column (f)). Each coefficient is an estimate of β in equation (6), allowed to vary by quartiles of the family income distribution. The sample is restricted to undergraduate students. Observations are weighted by one divided by the number of courses taken by each student. Standard errors in parentheses are clustered at the student level.

of the distribution see the largest increase in graduation rates from a one-SD increase in frontier knowledge proximity, equal to 2.1 pp (compared with 1.2 pp in the top quartile; Figure 5, panel (a)). They also experience a 0.023-point increase in GPA (panel (b)) and a 0.39-year reduction in time-to-degree (panel (c)). However, they do not benefit in terms of graduate school attendance (panel (d)). The effects of frontier knowledge on GPA are positive in all quartiles and increase monotonically with income (from 0.023 to 0.041 points). Similarly, the effects on time-to-degree are negative in all quartiles and become slightly more negative with income (from -0.39 to -0.42 years), implying faster progression at higher income levels. The effects on graduate school attendance are only significant above the bottom income quartile and rise roughly monotonically with income, reaching 4.8 pp in the top quartile.

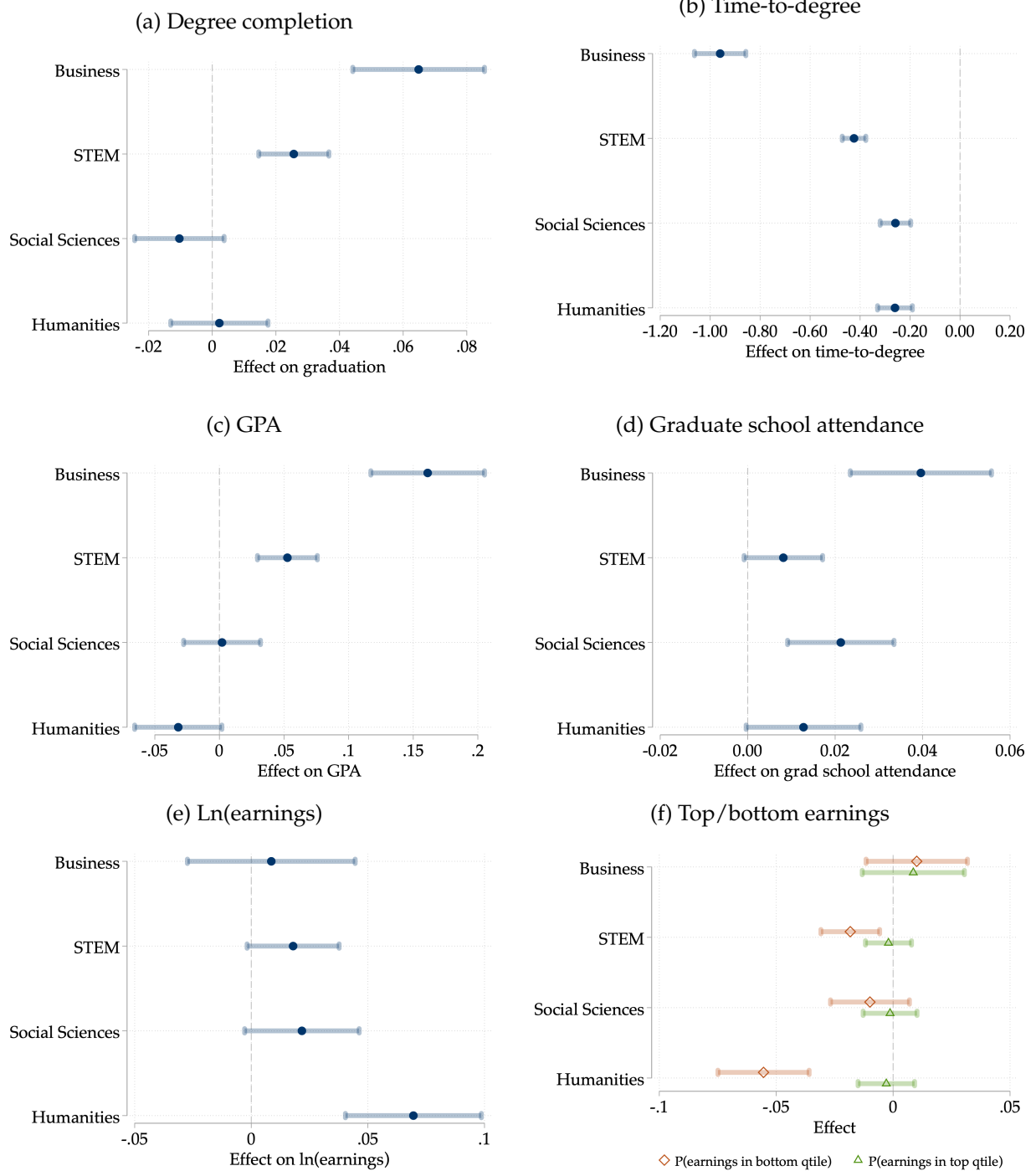
Taken together, these results suggest that frontier exposure appears to improve degree completion most where completion margins are fragile, especially among low-income students. However, for outcomes that require sustained investment and conversion of academic gains into longer-run educational choices (higher GPA, faster progression, graduate-school entry), returns rise with family income. This pattern is consistent with complementarity between frontier knowledge and family resources: exposure to advanced content is valuable for all groups, but students with greater resources are better able to translate that exposure into high-return trajectories beyond basic degree attainment. From an inequality perspective, the estimates imply partial equalization at the degree-completion margin, but persistent, and potentially widening, gaps in higher-order outcomes such as GPA gains and graduate-school transition.

5.5 Differences by Field

The returns to frontier knowledge may also vary across macro fields. To test this, we allow the parameter β in equation (6) to vary by field (Business, STEM, Social Sciences, Humanities), and we estimate effects separately for undergraduate and graduate students.

We find substantial variation across fields among undergraduates. A one-SD increase in proximity affects graduation the most in Business (6.5 pp), followed by STEM (2.6 pp), with no significant effects in Social Sciences or Humanities (Figure 6, panel (a)). GPA effects follow a similar pattern: large and positive in Business and STEM, near zero in Social Sciences, and small negative in Humanities (panel (b)). Time-to-degree effects are negative and significant in all fields, with the largest magnitude in Business (panel (c)). Graduate-school attendance effects are also positive and significant in all fields, with the same ranking: Business (3.9 pp), Social Sciences (2.1 pp), Humanities (1.3 pp), and STEM (0.8 pp; panel (d)).

Figure 6: Educational Effects of Frontier Knowledge Proximity, by Field



Note: OLS estimates; one observation is a student-course. The dependent variable is an indicator for students who graduate from their program (panel (a)), time-to-degree (panel (b)), GPA (panel (c)), an indicator for students ever enrolling in a graduate program within public schools in Texas (panel (d)), the natural logarithm of earnings 1 to 6 years after a student's expected graduation year (panel (e)), and indicators for earnings in the top and bottom quartiles of a student's cohort (column (f)). Each coefficient is an estimate of β in equation (6), allowed to vary by macro-field. The sample is restricted to undergraduate students. Observations are weighted by one divided by the number of courses taken by each student. Standard errors in parentheses are clustered at the student level.

Taken together, these estimates point to substantial field heterogeneity in how frontier exposure maps into outcomes. A possible explanation is that the same increase in proximity may generate different measured returns across fields, because pedagogy and evaluation differ by discipline. In methods- and problem-set-intensive fields, effects may appear more directly in GPA and progression; in reading- and writing-intensive fields, effects may be less visible in short-run grades and more visible in downstream choices such as graduate-school attendance.

6 Labor Market Returns to Frontier Knowledge

We now turn to the effects of exposure to frontier knowledge on labor market outcomes. Focusing on earnings in the first six years after predicted graduation, we study how these effects vary over time, across the earnings distribution, across students with different characteristics, and whether they operate through sorting across industries or within each industry. We present estimates of frontier knowledge exposure during undergraduate and graduate studies, without conditioning on program completion or enrollment in additional programs.

6.1 Exposure to Frontier Knowledge Increases Post-College Earnings

Exposure to frontier knowledge leads to significant increases in earnings, for both undergraduate and graduate students. Estimates of equation (6), using the log of average quarterly earnings 1-6 years after predicted graduation (defined as 6 years after degree start for undergraduates and 3 years after start for graduates), indicate that a one-SD increase in frontier knowledge proximity raises earnings by 2.8% for undergraduates (Table 4, panel (a), column 1) and 5.3% for graduates (panel (b), column 1). Effects are larger for men (3.6% increase for undergraduates and 7.4% for graduates) compared to women (2.1% and 2.8%, respectively; Appendix Table A6).

In columns 2-3 of Table 4, we examine the dynamics of these earnings effects by estimating impacts separately for earnings 1-3 years and 4-6 years after predicted graduation. For undergraduates, the effect of frontier knowledge exposure grows with time: a one-SD increase in proximity raises earnings by 2.5% after 1-3 years and 3.6% after 4-6 years (Table 4, panel (a), columns 2-3). We observe the opposite pattern for graduates, with a 5.7% increase after 1-3 years and a 2.8% increase after 4-6 years.

Distributional Effects of Frontier Knowledge Exposure We next examine how frontier knowledge proximity affects the likelihood of being in different parts of the earnings distribution. We do so by re-estimating equation (6) using, as outcomes, indicators of having earnings in the top and bottom quartiles of the cohort-specific earnings distribution.

Table 4: Earnings Effects of Frontier Knowledge Exposure

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|---------------------|---------------------|----------------------|---------------------|-----------------------|
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | |
| Proximity (sd) | 0.028*** (0.008) | 0.025*** (0.009) | 0.035*** (0.012) | -0.020*** (0.005) | 0.001 (0.004) | -0.005 (0.005) |
| R ² | 0.223 | 0.219 | 0.253 | 0.156 | 0.248 | 0.207 |
| N (student * course) | 5,993,879 | 5,296,230 | 1,914,649 | 5,993,879 | 5,993,879 | 5,993,879 |
| N clusters (students) | 126,157 | 108,868 | 47,313 | 126,157 | 126,157 | 126,157 |
| Panel (b) Graduate students | | | | | | |
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | |
| Proximity (sd) | 0.052*** (0.012) | 0.056*** (0.013) | 0.028 (0.020) | -0.022*** (0.007) | 0.019*** (0.007) | 0.011* (0.006) |
| R ² | 0.442 | 0.455 | 0.463 | 0.368 | 0.443 | 0.465 |
| N (student * course) | 248,758 | 216,292 | 97,699 | 248,758 | 248,758 | 248,758 |
| N clusters (students) | 25,663 | 21,924 | 10,860 | 25,663 | 25,663 | 25,663 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, instructor and school-major-cohort fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Exposure to frontier knowledge significantly reduces the likelihood of low earnings for undergraduates and raises the likelihood of high earnings for graduates. A one-SD increase in frontier exposure during undergraduate studies lowers the likelihood of being in the bottom earnings quartile by 2 pp (or roughly 8%; Table 4, panel (a), column 4), while it does not affect the probability of being in the top quartile (column 5, p -value = 0.73). The same increase during graduate studies reduces the risk of low earnings by a smaller 2 pp and significantly raises the odds of high earnings by 2 pp (Table 4, panel (b), columns 4 and 5). This pattern suggests that undergraduate exposure mainly reduces downside risk, whereas graduate exposure shifts students away from the bottom and toward the top of the earnings distribution.

Table 5: Earnings Effects of Frontier Knowledge Exposure - Average vs Top Courses

| Panel (a) Undergraduate students | | | | | |
|---|---------------------|---------------------|-------------------|----------------------|----------------------|
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Median proximity (sd) | 0.026*** (0.007) | 0.021*** (0.008) | 0.020* (0.011) | -0.003 (0.003) | 0.012*** (0.004) |
| 90th pctlile proximity (sd) | -0.001 (0.005) | -0.001 (0.006) | 0.008 (0.008) | -0.003 (0.002) | -0.012*** (0.003) |
| R ² | 0.223 | 0.218 | 0.253 | 0.104 | 0.248 |
| N (student * course) | 5,993,868 | 5,296,215 | 1,914,647 | 11,681,380 | 5,993,868 |
| N clusters (students) | 126,157 | 108,868 | 47,313 | 234,005 | 126,157 |
| Panel (b) Graduate students | | | | | |
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Median proximity (sd) | 0.044*** (0.012) | 0.044*** (0.013) | 0.016 (0.020) | -0.012*** (0.004) | 0.020*** (0.007) |
| 90th pctlile proximity (sd) | 0.001 (0.008) | 0.004 (0.009) | 0.005 (0.014) | 0.003 (0.002) | -0.000 (0.005) |
| R ² | 0.440 | 0.454 | 0.460 | 0.250 | 0.440 |
| N (student * course) | 248,800 | 216,320 | 97,666 | 484,760 | 248,800 |
| N clusters (students) | 25,667 | 21,925 | 10,863 | 50,626 | 25,667 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5). *Median proximity* is the median frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. *90th pctlile proximity* is the proximity of the 90th percentile course in each student's transcript, also measured in course-level standard deviations. All specifications control for lagged proximity (median and 90th percentile), instructor and school-major-cohort fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

The Role of Industry Sorting A possible channel through which frontier knowledge may increase earnings is by raising the likelihood of employment in high-paying industries. To test for this, we calculate the average earnings in each student's industry of first employment and use them as the outcome in equation (6). Estimates are indistinguishable from zero in the undergraduate sample (Table 4, panel (a), column 6), while they are small (1%) and only marginally significant in the graduate sample (panel (b), column 6). These results indicate that earnings gains arise mostly within industry, rather than through sorting into high-paying industries.

6.2 Which Courses Drive the Returns to Frontier Exposure?

To understand which courses drive returns to frontier exposure, we repeat the two tests used above for educational outcomes. First, we allow both p_i and $p_{i,-1}$ and the corresponding coefficients in equation (6) to vary by year of instruction (pooling years after the third). We find that frontier knowledge improves students' earnings the most when experienced later in a student's career. A one-SD increase in proximity raises average earnings by 1.1% in year 3 and 3.0% in years 4 and beyond. The effect is insignificant in years 1 and 2 (Figure 3, panel (e)). Similarly, the same proximity increase reduces the probability of being in the bottom earnings quartile by 1 pp in year 3 and by 2.1 pp in years 4 and beyond, with no effect in the first two years. It also raises the probability of being in the top earnings quartile by 0.5 pp in years 4 and beyond (panel (f)). These estimates indicate that the labor-market returns to frontier exposure are concentrated in later-stage coursework.

Second, we augment equation (6) to include both median proximity and 90th-percentile proximity (together with their lags). We find that the earnings effects of frontier knowledge proximity are largely driven by median proximity, while the effects of 90th-percentile are generally small and have the opposite sign. For example, a one-SD increase in median proximity, holding top proximity fixed, raises earnings by 2.6%, while a one-SD increase in top proximity has no effect (Table 5, column 1). Similar patterns hold for the probability of being in the bottom earnings quartile (columns 4). For the top quartile, we again find a positive and significant effect of median proximity and a negative effect of top proximity (column 5).

Taken together, these causal effects indicate that earnings gains are driven primarily by broad frontier exposure across the curriculum rather than by isolated high-proximity courses. This pattern is consistent with the earnings payoff reflecting cumulative skill accumulation across many courses, rather than exposure to a single "top" course.

6.3 Differences by Baseline Ability

The earnings effects of frontier knowledge exposure display a strong positive gradient with respect to baseline ability. While frontier exposure increases earnings across the entire ability distribution, the magnitude of the effect is larger for students in the top ability quartile (10%) and much smaller for those in the bottom quartile (1.8%; Figure 4, panel (e)). For students with no reported test score, the effect on mean earnings is statistically indistinguishable from zero. Given that this group is largely composed of transfer students, this null is consistent with frontier exposure improving skills but not fully relaxing the labor-market frictions or weaker pre-college signals that may limit

earnings growth.

These patterns again reveal important asymmetries. Exposure to frontier knowledge reduces the probability of low earnings for students across the ability distribution and for those with no reported scores, with the largest reductions occurring among higher-ability students. At the upper end of the distribution, however, increases in the probability of top-quartile earnings are concentrated entirely among students in the top two ability quartiles. For these students, frontier exposure raises the likelihood of top-quartile earnings by 2.1-4.5 pp, while effects for students in the bottom half of the ability distribution and for those with no reported score are small, negative, and insignificant (Figure 4, panel (f)).

These patterns suggest that baseline ability plays a key role in shaping the extent to which frontier-acquired skills translate into high-paying labor market outcomes. While frontier knowledge appears to raise skills broadly, access to high-return positions is concentrated among students with stronger initial preparation, consistent with complementarities between frontier exposure and pre-existing skills. Taken together, the estimates imply compression in downside earnings risk across groups, but divergence in upper-tail gains by baseline ability.

6.4 Differences by Family Income

We next examine how the labor market returns to frontier knowledge vary across undergraduate students' socioeconomic backgrounds. Figure 5 reports estimates of the effect of frontier exposure on earnings outcomes by quartiles of family income, including mean earnings effects (panel (e)) and distributional outcomes (panel (f)).

We find that exposure to frontier knowledge increases earnings across the entire family income distribution, but effects are larger for students from more advantaged backgrounds. A one-SD increase in frontier exposure raises average earnings by 2.4% in the bottom income quartile and by roughly 5% in the top two quartiles (Figure 5, panel (e)). While the effect is positive throughout, the gradient indicates that students from higher-income families are better positioned to translate frontier-acquired skills into earnings gains in the labor market. This pattern implies that, although all income groups gain, average earnings gains are larger at higher family income, consistent with widening between-group gaps in mean post-college earnings.

Distributional results reinforce this pattern. Exposure to frontier knowledge reduces the probability of having earnings in the bottom quartile across all family income groups, with reductions ranging from 1.9 pp in the lowest income quartile to about 3 pp in the top two quartiles. In contrast, increases in the probability of having earnings in the top quartile are concentrated among students

from the top two family income quartiles. For these students, a one-SD increase in frontier exposure raises the likelihood of top-quartile earnings by 1 pp, while effects for students from the bottom two income quartiles are small, negative, and statistically indistinguishable from zero (Figure 5, panel (f)).

Taken together, these results suggest that frontier knowledge reduces downside risk broadly, but that access to upside earnings outcomes depends more strongly on family background. This pattern is consistent with complementarity between frontier exposure and family resources: exposure to advanced content is valuable for all groups, but students with greater resources are better able to translate that exposure into high-return trajectories beyond basic earnings protection. From an inequality perspective, the estimates point to compression at the bottom of the earnings distribution but expansion at the top. Thus, frontier exposure appears equalizing on downside risk yet potentially inequality-increasing in average and upper-tail earnings.

6.5 Differences by Field

We next examine whether labor-market returns to frontier knowledge differ across fields. For undergraduates, returns are largest in Humanities: a one-SD increase in frontier proximity raises earnings by 7.2% (Figure 6, panel (e)), largely due to a reduction in the probability of low earnings (panel (f)). Effects are smaller in Social Sciences (2.2%) and STEM (1.8%), and statistically indistinguishable from zero in Business. For graduate students, the pattern differs: returns are largest in STEM (7.4%), followed by Social Sciences (5.2%); the Business estimate is positive but imprecise (3.3%), and the Humanities estimate is close to zero.

These earnings patterns are informative when read together with the earlier educational-outcome results. In the undergraduate sample, Business and STEM showed the strongest effects on completion/progression, but not the largest earnings returns. Conversely, Humanities shows comparatively weak educational effects but the largest earnings gains. The key implication is that field-level earnings heterogeneity is not a mechanical reflection of field-level heterogeneity in educational outcomes.

6.6 Frontier Knowledge, Educational Attainment, and Earnings: Mediation Analysis

Our results so far indicate that exposure to frontier knowledge in college improves both educational outcomes and earnings. While effects on both sets of outcomes are present across the family-income and ability distributions, the earnings returns to frontier knowledge exposure tend to be larger for middle- and high-income students, whereas the educational attainment and learning gains are more

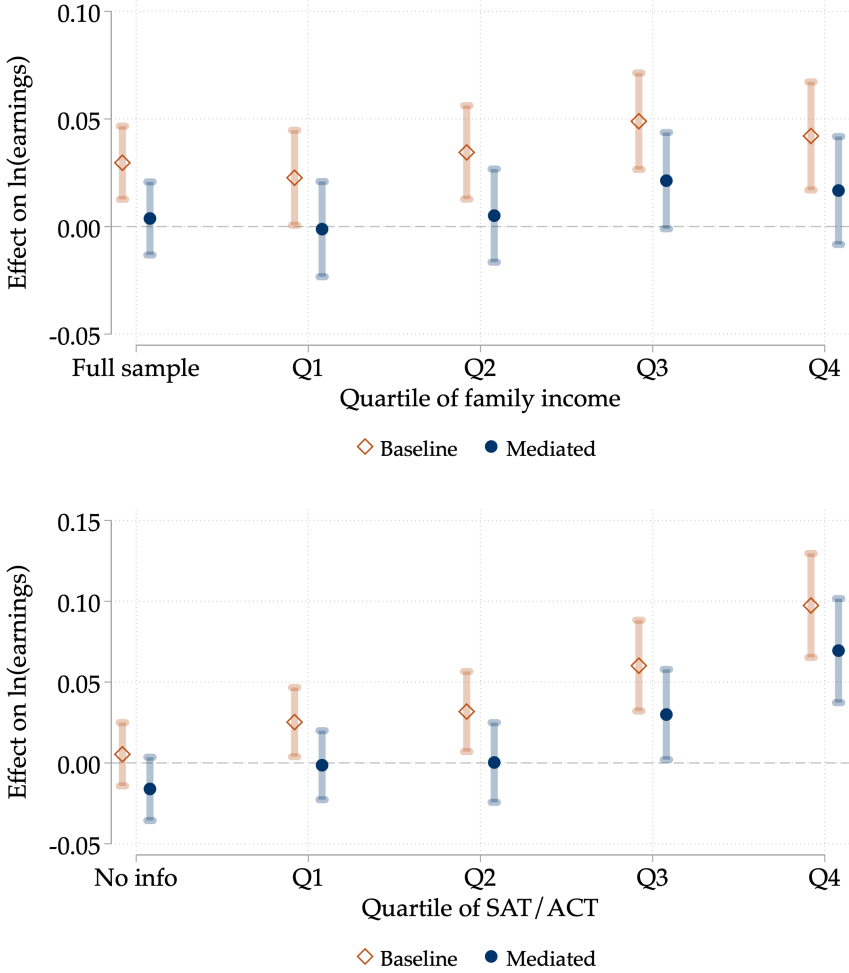
equally distributed or even benefit lower-income students more. This pattern suggests that higher-income students may be able to reap the benefits of frontier knowledge in a way that goes beyond improvements in educational outcomes.

To discipline this interpretation and, more generally, to better understand how much of the earnings gains can be accounted for by improvements in educational attainment and performance, we conduct a descriptive mediation analysis. We contrast our baseline earnings estimates with those obtained from specifications that add controls for undergraduate and graduate degree completion, graduate school enrollment, cumulative GPA, and time to degree. When examining heterogeneous effects by family income, ability, and field, we additionally interact these educational controls with the relevant group indicators.

The results from this exercise are shown in Figure 7. In the full sample, the coefficient on frontier knowledge proximity is substantially attenuated and no longer statistically distinguishable from zero when we control for all educational outcomes. This pattern is consistent with the earnings returns to frontier knowledge exposure being largely accounted for by differences in educational attainment and performance. However, the degree of attenuation differs markedly across groups. For students from families in the first income quartile and for those with baseline ability below the median, the inclusion of educational outcomes explains a large share of the earnings effect of frontier exposure. In these groups, controlling for degree completion, academic performance, and time to degree makes the effects of frontier knowledge on earnings small and statistically insignificant. In contrast, for students from higher-income families and those with higher baseline ability, a non-trivial portion of the earnings effect remains even after conditioning on the full set of educational outcomes.

While descriptive, these patterns suggest that frontier knowledge exposure translates into earnings gains through different pathways across student groups. For lower-income and lower-ability students, earnings gains appear to be closely linked to improvements in educational progression and performance, consistent with frontier exposure facilitating degree completion and reducing barriers to persistence. For higher-income and higher-ability students, frontier exposure is associated with earnings gains that are less tightly connected to observable educational outcomes, consistent with complementarities between frontier-acquired skills and post-college opportunities.

Figure 7: Educational and Earnings Effects of Frontier Knowledge Exposure: Mediation Analysis



Notes: The figure shows *baseline* estimates of frontier knowledge proximity on the natural logarithm of earnings 1-6 years after predicted graduation, along with the same *mediated* estimates obtained controlling for intermediate educational outcomes (graduation, time-to-degree, GPA, and an indicator for graduate school attendance). Panel (a) shows estimates in the full sample and by quartile of parental income, obtained allowing β in equation (6) to vary across quartiles. Panel (b) shows estimates by quartile of the ability distribution, allowing β in equation (6) to vary across quartiles. Mediated estimates by family income and quartile are obtained by also interacting intermediate educational outcomes by quartiles of the relevant distribution. Standard errors are clustered at the student level.

7 Robustness

In this section, we assess two issues: (i) whether our identifying variation is plausibly conditionally exogenous, and (ii) whether the results depend on the specific way we measure frontier knowledge proximity. We begin with a placebo test that directly targets the identifying assumption.

7.1 Conditional Independence: Placebo Test with Future Updates

Our identification strategy assumes that, conditional on lagged proximity and fixed effects, within-course updates in frontier proximity are unanticipated and unrelated to unobserved student-level shocks that affect outcomes. This assumption would be violated if proximity changes are part of persistent course- or instructor-level trends, rather than unexpected, one-off syllabus updates. If this were true, then proximity measured after a student takes a course could still predict that student's earlier outcomes.

To test this implication, we construct a “placebo” regressor equal to the proximity of the same course two academic years after the student took it.¹³ We then augment equation (6) by including this future proximity measure and its lag, alongside contemporaneous proximity and its lag. Because the placebo regressor is realized after the outcome period, it should not affect outcomes; any predictive power would instead indicate residual confounding or persistent unobserved trends.

Across specifications, the coefficient on future proximity is small and statistically indistinguishable from zero (see Appendix Table A7 for educational outcomes and Appendix Table A8 for earnings). This evidence indicates that our baseline estimates are unlikely to be driven by persistent unobserved trends or anticipatory sorting, and is consistent with the conditional-independence assumption underlying our main results.

7.2 Time-Varying Instructor Effects

A potential threat to our strategy is that changes in a course's frontier knowledge proximity may be correlated with time-varying instructor effects that directly affect student outcomes. For example, consider an instructor who has been teaching the same course, with the same material, for many years and suddenly decides to update the course. Such an update could lead the instructor to start teaching better. If this occurrence is prevalent, it could explain part of the estimates we find. More generally, we want to account for the possibility that an increase in instructor effort could simultaneously lead to greater incorporation of frontier material and improvements in teaching effectiveness.

While we are unable to directly observe teaching ability beyond instructor fixed effects, we probe this threat in two ways: by letting instructor effects vary semi-parametrically with experience, and by excluding large course updates, i.e., dramatic changes in course content that could coincide with large changes in instructor effort.

¹³We use a two-year forward measure, rather than a one-year forward measure, because the lag of the one-year forward measure mechanically coincides with current proximity, our main variable of interest.

To implement the first test, we augment equation (6) by flexibly controlling for instructors' teaching experience, a proxy for time-varying teaching effectiveness. We define teaching experience as the number of years elapsed between the year of course instruction and the year in which the instructor first appears teaching in our data. We then augment equation (6) to include instructor fixed effects interacted with indicators for experience in the following intervals: 0, 1-3, 4-5, and 6 or more. As shown in Appendix Tables A13 and A14, our main estimates are largely unchanged when controlling for instructor experience, suggesting that time-varying instructor ability is unlikely to drive our results.

To conduct the second test, we calculate p_i and $p_{i,-1}$ in equation (6) by excluding courses that experience a major content overhaul, defined as a proximity update larger than one SD (or 25% of the course's content). Our estimates are largely unchanged (Appendix Tables A11 and A12).

Notably, estimates that exclude both instructor fixed effects and experience controls are similar in magnitude to our baseline estimates (Appendix Tables A9 and A10). Although this is a less preferred specification, it further indicates that instructor-specific effects do not appear to be systematically related to changes in course proximity.

7.3 N-gram-Based Measures of Course Proximity

The measure of frontier knowledge proximity used throughout the paper is based on similarities between term vectors constructed using a pre-specified dictionary. As an alternative, we construct n-gram-based versions of the proximity measure, which rely on short phrase overlap rather than only dictionary terms. Our results are qualitatively unchanged (Appendix Tables A15 and A16).

8 Discussion and Conclusion

This paper has documented substantial heterogeneity in the frontier knowledge content of university courses and shown that this heterogeneity carries meaningful consequences for students. Two features of the results stand out. First, the benefits of frontier exposure are broad across educational outcomes—graduation rates, time-to-degree, GPA, and graduate school attendance—and particularly strong for students on fragile completion margins, where early exposure to cutting-edge material appears to shape engagement and persistence across the entire college trajectory. Second, the ability to translate those attainment gains into higher-threshold outcomes—graduate school and labor-market earnings—rises with pre-college resources. A mediation analysis confirms the mechanism: for disadvantaged students, frontier exposure works by improving attainment; for more advantaged students, it additionally generates skills that command a direct return in the

labor market. Frontier knowledge thus reduces socioeconomic gaps in degree completion while leaving—and potentially widening—gaps in earnings, reflecting a fundamental complementarity between curriculum content and the resources students bring to college.

These results carry several policy implications. Since the largest attainment gains accrue to students on fragile completion margins, and since instructor heterogeneity drives most of the variation in frontier proximity, decisions about which faculty teach introductory and gateway courses may matter as much as—or more than—decisions about advanced offerings. In a companion paper (Biasi and Ma, 2022), we use a broader national sample of syllabi to document that research-active instructors—those with more publications, citations, and grants, and whose research closely matches the course topic—teach substantially more frontier knowledge, consistent with teaching and research being complements rather than substitutes. Combined with our findings, this suggests that investments in academic research generate dual returns: new knowledge and, through research-active faculty, more current instruction and benefits for students.

Our findings open avenues for future work. The datasets and methodology developed here and in the companion work can be extended to study other dimensions of course content—problem-set structure, interdisciplinarity, the integration of empirical methods—and to examine outcomes beyond those studied here, including innovation and patent production. Whether the gaps in frontier exposure we document translate into gaps in the creation of new ideas is a consequential next question. We plan to make our syllabi data, proximity measure, and underlying code publicly available 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, 2025, Tapping into talent: Coupling education and innovation policies for economic growth, *Review of Economic Studies* 92, 696–736.
- Altonji, Joseph G, Erica Blom, and Costas Meghir, 2012, Heterogeneity in human capital investments: High school curriculum, college major, and careers, *Annual Review of Economics* 4, 185–223.
- Andrews, Michael 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.

- Arrow, Kenneth J., 1973, Higher education as a filter, *Journal of Public Economics* 2, 193–216.
- Arteaga, Carolina, 2018, The effect of human capital on earnings: Evidence from a reform at Colombia’s top university, *Journal of Public Economics* 157, 212–225.
- Aryal, Gaurab, Manudeep Bhuller, and Fabian Lange, 2022, Signaling and employer learning with instruments, *American Economic Review* 112, 1669–1702.
- Baumol, William J, 2005, Education for innovation: Entrepreneurial breakthroughs versus corporate incremental improvements, *Innovation Policy and the Economy* 5, 33–56.
- Becker, Gary S., 1964, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (Columbia University Press, New York).
- 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 Song Ma, 2022, The education-innovation gap, Technical report, National Bureau of Economic Research.
- 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.
- 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.
- Deming, David J, and Kadeem L Noray, 2020, Earnings dynamics, changing job skills, and stem careers, *Quarterly Journal of Economics* .
- 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).
- Hemelt, Steven W, Brad Hershbein, Shawn M Martin, and Kevin M Stange, 2023, College majors and skills: Evidence from the universe of online job ads, *Labour Economics* 85, 102429.
- 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 Shima, 2021, Major complexity index and college skill production, *Available at SSRN 3791651* .
- Ma, Xuezhe, and Eduard Hovy, 2016, End-to-end sequence labeling via bi-directional lstm-cnns-crf, *arXiv preprint arXiv:1603.01354* .
- Moser, Petra, and Alessandra Voena, 2012, Compulsory licensing: Evidence from the trading with the enemy act, *American Economic Review* 102, 396–427.
- 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* .
- Spence, Michael, 1973, Job market signaling, *Quarterly Journal of Economics* 87, 355–374.
- Toivanen, Otto, and Lotta Väänänen, 2016, Education and invention, *Review of Economics and Statis-*

tics 98, 382–396.

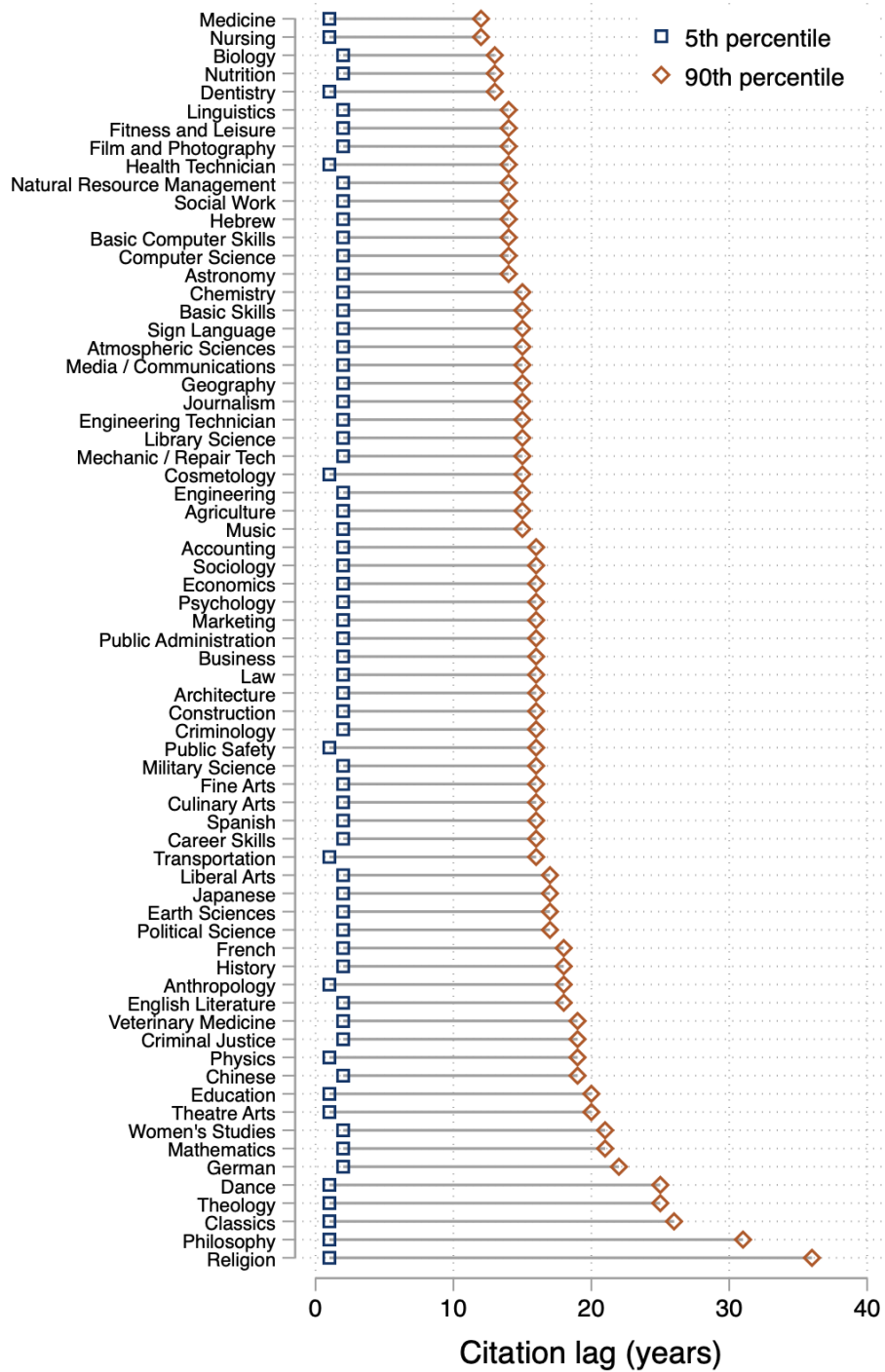
Williams, Heidi L, 2013, Intellectual property rights and innovation: Evidence from the human genome, *Journal of Political Economy* 121, 1–27.

Appendix

For online publication only

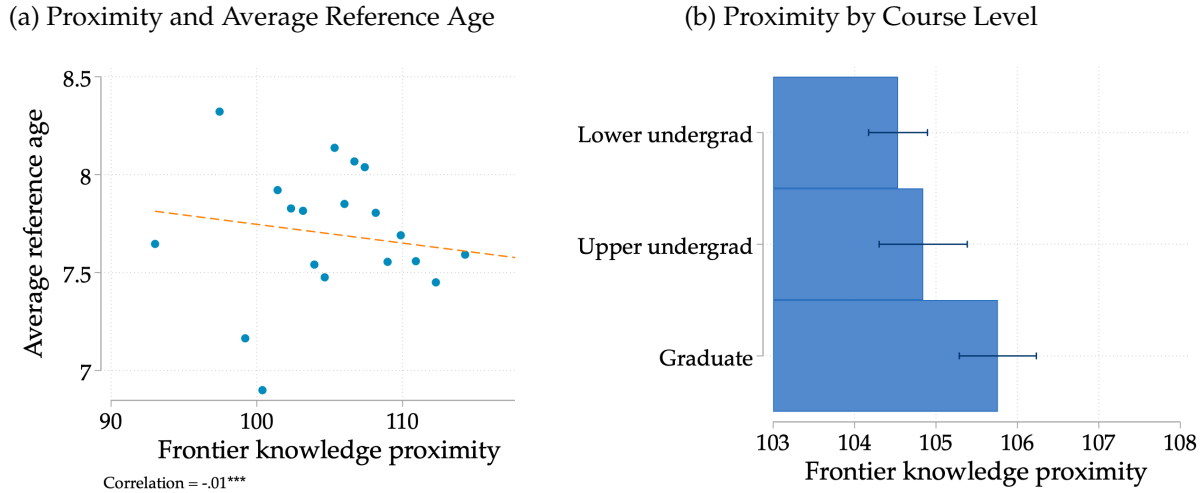
Appendix A Additional Tables and Figures

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: Validating 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), where 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.

Table A1: Decomposing the Variation in the Frontier Knowledge Proximity: Schools, Years, Fields, Courses, and Instructors

| | Partial R^2 | |
|------------|---------------|-------|
| Year | 0.087 | 0.079 |
| Field | 0.132 | 0.046 |
| School | 0.007 | 0.003 |
| Course | . | 0.258 |
| Instructor | . | 0.210 |
| All | 0.226 | 0.597 |

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.

Table A2: 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 A3: Predicting Changes in Course Proximity Using Student Observables

| Variable | Proximity ($p_{k,t}$) | | Change in proximity ($\Delta_{k,t}$) | |
|-------------------------------|-------------------------|----------------------|--|----------------------|
| | (1) | (2) | (3) | (4) |
| <i>Students</i> | | | | |
| Family income = 0 | -0.060 (0.055) | -0.026 (0.063) | -0.002 (0.047) | 0.011 (0.049) |
| Family income in 1st quartile | -0.073* (0.038) | -0.062 (0.044) | 0.013 (0.030) | 0.009 (0.030) |
| Family income in 2nd quartile | -0.072** (0.037) | -0.032 (0.043) | 0.015 (0.030) | 0.029 (0.031) |
| Family income in 3rd quartile | -0.083** (0.037) | -0.033 (0.044) | -0.022 (0.031) | -0.030 (0.033) |
| Family income in 4th quartile | -0.056 (0.036) | 0.004 (0.041) | -0.011 (0.029) | -0.005 (0.030) |
| SAT/ACT in 1st quartile | -0.067** (0.032) | 0.007 (0.036) | 0.002 (0.020) | -0.022 (0.022) |
| SAT/ACT in 2nd quartile | -0.146*** (0.032) | -0.087** (0.037) | 0.039 (0.025) | 0.034 (0.025) |
| SAT/ACT in 3rd quartile | -0.110*** (0.033) | -0.045 (0.038) | 0.033 (0.025) | 0.037 (0.027) |
| SAT/ACT in 4th quartile | -0.095*** (0.036) | -0.068 (0.042) | -0.007 (0.022) | -0.022 (0.024) |
| Share female | -0.004 (0.024) | 0.044 (0.028) | -0.007 (0.015) | -0.004 (0.015) |
| <i>Syllabi</i> | | | | |
| Grade points | | 0.121*** (0.011) | | -0.005 (0.005) |
| Requires exam | | -0.159*** (0.049) | | |
| Δ Requires exam | | | | -0.073 (0.046) |
| Share homework | | 0.065 (0.063) | | |
| Δ Share homework | | | | 0.104 (0.075) |
| Share report | | 0.059* (0.036) | | |
| Δ Share report | | | | 0.002 (0.046) |
| <i>Instructors</i> | | | | |
| Nr instructors = 2 | | -0.024** (0.012) | | -0.023*** (0.007) |
| Nr instructors = 3+ | | -0.054*** (0.014) | | 0.029*** (0.009) |
| Experience 2-3 years | | 0.015 (0.016) | | -0.024 (0.017) |
| Experience 4-5 years | | -0.009 (0.018) | | -0.035** (0.017) |
| Experience 5+ years | | -0.033* (0.019) | | -0.046*** (0.017) |
| New instructor | | -0.018* (0.009) | | -0.018** (0.008) |
| N (course * year) | 108,435 | 77,971 | 78,057 | 65,585 |
| F-stat, student variables | 4.30 | 0.15 | 0.69 | 0.32 |
| F-stat, syllabi variables | | 0.00 | | 0.23 |
| F-stat, instructor variables | | 0.00 | | 0.00 |

Note: OLS estimates; one observation is a course-year. The dependent variable is frontier knowledge proximity (columns 1 and 3) and the change in proximity from the previous year (columns 2 and 4). The independent variables are average characteristics of all students taking the course, of the syllabus, and of the course's instructors. *Exam* equals one if the course requires an exam; *Share homework* and *Share report* are the shares of terms in the assignment portion of a syllabus that refer to these activities; and variables starting with Δ denote changes to the variable from the previous year. Instructor experience is calculated as the time elapsed since the instructor's first publication in OpenAlex. All specifications control for field (as indicated by the course prefix) by-school-by-year fixed effects. Standard errors in parentheses are clustered at the course level.

* ≤ 0.1 , ** ≤ 0.05 , *** ≤ 0.01 .

Table A4: Educational Effects of Frontier Knowledge Exposure - Sample of Undergraduate Students Who Reach Year 3

| | Graduates (1) | Time-to degree (2) | GPA (3) | Future grade rank (4) | Attends grad school (5) |
|-----------------------|---------------------|--------------------------|---------------------|-----------------------------|-------------------------------|
| Proximity (sd) | 0.032*** (0.004) | -0.409*** (0.019) | 0.062*** (0.009) | 0.005** (0.002) | 0.021*** (0.004) |
| Mean dep. var. | 0.930 | 5.777 | 2.920 | 0.506 | 0.108 |
| R ² | 0.237 | 0.552 | 0.254 | 0.102 | 0.224 |
| N (student * course) | 5,782,490 | 5,399,343 | 5,780,286 | 2,079,811 | 5,782,490 |
| N clusters (students) | 116,714 | 105,115 | 116,223 | 95,044 | 116,714 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1), time-to-degree in years (column 2), GPA (column 3), and an indicator for enrollment in a graduate program within Texas (column 4). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations; in column 5 of panel (a), this variable is calculated as the average over all courses taken by the student prior to the focal course. All specifications control for average lagged proximity, instructor and school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Estimates are shown for undergraduate students who reach year 3. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A5: Educational Effects of Frontier Knowledge Exposure, by Gender

| Panel (a) Undergraduate women | | | | | | |
|--------------------------------------|--------------------|----------------------|---------------------|--------------------|----------------------|---------------------|
| | Graduates | Time-to degree | GPA | Future grade rank | Attends grad school | |
| | (1) | (2) | (3) | (4) | (5) | |
| Proximity (sd) | 0.011** (0.006) | -0.361*** (0.025) | 0.009 (0.012) | 0.006** (0.003) | 0.017*** (0.006) | |
| Mean dep. var. | 0.920 | 5.670 | 2.960 | 0.522 | 0.125 | |
| R ² | 0.329 | 0.586 | 0.329 | 0.134 | 0.246 | |
| N (student * course) | 3,090,731 | 2,852,567 | 3,088,669 | 1,033,190 | 3,090,731 | |
| N clusters (students) | 64,243 | 56,007 | 63,867 | 48,834 | 64,243 | |
| Panel (b) Graduate women | | | | | | |
| | Graduates | MA | GPA | Graduates | PhD | GPA |
| | (1) | Time-to degree | (3) | (4) | Time-to degree | (6) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.020 (0.012) | -0.355*** (0.040) | 0.002 (0.018) | 0.021 (0.018) | -0.694*** (0.169) | -0.010 (0.042) |
| Mean dep. var. | 0.715 | 2.617 | 3.425 | 0.226 | 7.207 | 2.828 |
| R ² | 0.479 | 0.706 | 0.507 | 0.776 | 0.866 | 0.797 |
| N (student * course) | 113,583 | 104,775 | 113,310 | 21,698 | 7,150 | 21,657 |
| N clusters (students) | 11,672 | 10,324 | 11,570 | 1,765 | 814 | 1,757 |
| Panel (c) Undergraduate men | | | | | | |
| | Graduates | Time-to degree | GPA | Future grade rank | Attends grad school | |
| | (1) | (2) | (3) | (4) | (5) | |
| Proximity (sd) | 0.014* (0.007) | -0.488*** (0.029) | 0.067*** (0.015) | 0.004 (0.003) | 0.018*** (0.005) | |
| Mean dep. var. | 0.891 | 5.926 | 2.784 | 0.490 | 0.089 | |
| R ² | 0.351 | 0.590 | 0.358 | 0.137 | 0.276 | |
| N (student * course) | 2,622,471 | 2,345,162 | 2,620,748 | 933,047 | 2,622,471 | |
| N clusters (students) | 53,775 | 44,797 | 53,536 | 40,345 | 53,775 | |
| Panel (d) Graduate men | | | | | | |
| | Graduates | MA | GPA | Graduates | PhD | GPA |
| | (1) | Time-to degree | (3) | (4) | Time-to degree | (6) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.013 (0.015) | -0.129*** (0.049) | 0.052** (0.021) | 0.006 (0.015) | -0.383** (0.158) | 0.087*** (0.032) |
| Mean dep. var. | 0.754 | 2.504 | 3.308 | 0.264 | 6.897 | 2.712 |
| R ² | 0.514 | 0.728 | 0.525 | 0.786 | 0.739 | 0.807 |
| N (student * course) | 69,322 | 62,458 | 69,135 | 19,113 | 6,665 | 19,066 |
| N clusters (students) | 8,319 | 7,236 | 8,255 | 1,796 | 859 | 1,785 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panels (b) and (d)), time-to-degree in years (column 2 and column 5 in panels (b) and (d)), GPA (column 3 and column 6 in panels (b) and (d)), and an indicator for enrollment in a graduate program within Texas (column 4 in panels (a) and (c)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, instructor and school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate women, panel (b) for graduate women, panel (c) for undergraduate men, and panel (d) for graduate men. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A6: Earnings Effects of Frontier Knowledge Exposure, by Gender

| Panel (a) Undergraduate women | | | | | | |
|--------------------------------------|---------------------|---------------------|--------------------|----------------------|---------------------|-----------------------|
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | (6) |
| Proximity (sd) | 0.020* (0.011) | 0.023* (0.012) | 0.027* (0.016) | -0.018** (0.008) | 0.002 (0.005) | -0.006 (0.007) |
| R ² | 0.232 | 0.235 | 0.277 | 0.187 | 0.266 | 0.232 |
| N (student * course) | 2,023,801 | 1,769,027 | 708,257 | 2,023,801 | 2,023,801 | 2,023,801 |
| N clusters (students) | 59,811 | 51,736 | 22,439 | 59,811 | 59,811 | 59,811 |
| Panel (b) Graduate women | | | | | | |
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | (6) |
| Proximity (sd) | 0.028 (0.017) | 0.037* (0.019) | 0.017 (0.027) | -0.013 (0.011) | 0.005 (0.010) | 0.003 (0.010) |
| R ² | 0.459 | 0.481 | 0.464 | 0.413 | 0.457 | 0.508 |
| N (student * course) | 136,423 | 118,798 | 54,928 | 136,423 | 136,423 | 136,423 |
| N clusters (students) | 13,104 | 11,231 | 5,680 | 13,104 | 13,104 | 13,104 |
| Panel (c) Undergraduate men | | | | | | |
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | (6) |
| Proximity (sd) | 0.036** (0.014) | 0.030* (0.016) | 0.044** (0.020) | -0.026*** (0.009) | 0.006 (0.008) | -0.004 (0.010) |
| R ² | 0.282 | 0.286 | 0.333 | 0.228 | 0.267 | 0.268 |
| N (student * course) | 1,774,823 | 1,530,905 | 649,071 | 1,774,823 | 1,774,823 | 1,774,823 |
| N clusters (students) | 49,215 | 41,971 | 19,352 | 49,215 | 49,215 | 49,215 |
| Panel (d) Graduate men | | | | | | |
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | (6) |
| Proximity (sd) | 0.072*** (0.021) | 0.084*** (0.025) | 0.014 (0.034) | -0.031*** (0.010) | 0.046*** (0.013) | 0.022* (0.011) |
| R ² | 0.538 | 0.563 | 0.558 | 0.511 | 0.495 | 0.537 |
| N (student * course) | 89,625 | 77,347 | 33,302 | 89,625 | 89,625 | 89,625 |
| N clusters (students) | 9,728 | 8,231 | 4,041 | 9,728 | 9,728 | 9,728 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, instructor and school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate women, panel (b) for graduate women, panel (c) for undergraduate men, and panel (d) for graduate men. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A7: Educational Effects of Frontier Knowledge Exposure - Placebo Test

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|----------------------|---------------------|---------------------|----------------------|------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Proximity (sd) | -0.010 (0.006) | -0.385*** (0.024) | -0.001 (0.013) | 0.014*** (0.005) | | |
| Future proximity (sd) | 0.050*** (0.004) | 0.147*** (0.018) | 0.096*** (0.009) | -0.006* (0.004) | | |
| Mean dep. var. | 0.905 | 5.780 | 2.877 | 0.105 | | |
| R ² | 0.295 | 0.557 | 0.297 | 0.222 | | |
| N (student * course) | 5,970,541 | 5,423,736 | 5,966,693 | 5,970,541 | | |
| N clusters (students) | 122,771 | 104,659 | 122,166 | 122,771 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.025** (0.011) | -0.232*** (0.035) | 0.009 (0.016) | 0.019 (0.013) | -0.343*** (0.100) | 0.030 (0.023) |
| Future proximity (sd) | -0.004 (0.008) | 0.092*** (0.027) | -0.013 (0.011) | -0.008 (0.013) | 0.101 (0.086) | 0.026 (0.022) |
| Mean dep. var. | 0.742 | 2.560 | 3.382 | 0.289 | 6.867 | 2.753 |
| R ² | 0.428 | 0.677 | 0.455 | 0.715 | 0.821 | 0.754 |
| N (student * course) | 187,152 | 171,365 | 186,794 | 45,620 | 16,716 | 45,498 |
| N clusters (students) | 19,111 | 16,864 | 18,992 | 3,870 | 1,969 | 3,847 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. The variable *Future proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations two years after the student takes the course. All specifications control for average lagged proximity, instructor and school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A8: Earnings Effects of Frontier Knowledge Exposure - Placebo Test

| Panel (a) Undergraduate students | | | | | |
|---|---------------------|---------------------|--------------------|----------------------|---------------------|
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.023** (0.011) | 0.016 (0.012) | 0.033** (0.016) | -0.021*** (0.007) | -0.007 (0.005) |
| Future proximity (sd) | -0.002 (0.008) | 0.006 (0.009) | 0.001 (0.013) | -0.003 (0.005) | -0.002 (0.004) |
| R ² | 0.224 | 0.220 | 0.254 | 0.156 | 0.249 |
| N (student * course) | 5,970,541 | 5,277,433 | 1,903,248 | 5,970,541 | 5,970,541 |
| N clusters (students) | 122,771 | 106,029 | 45,868 | 122,771 | 122,771 |
| Panel (b) Graduate students | | | | | |
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.059*** (0.015) | 0.054*** (0.017) | 0.046* (0.025) | -0.021** (0.009) | 0.030*** (0.009) |
| Future proximity (sd) | 0.013 (0.012) | 0.018 (0.013) | -0.031 (0.021) | -0.003 (0.007) | -0.003 (0.007) |
| R ² | 0.441 | 0.455 | 0.466 | 0.376 | 0.437 |
| N (student * course) | 234,157 | 204,452 | 90,078 | 234,157 | 234,157 |
| N clusters (students) | 22,007 | 18,996 | 8,981 | 22,007 | 22,007 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. The variable *Future proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations two years after the student takes the course. All specifications control for average lagged proximity, instructor and school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A9: Educational Effects of Frontier Knowledge Exposure - No Instructor Fixed Effects

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|----------------------|---------------------|---------------------|----------------------|--------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Proximity (sd) | 0.032*** (0.004) | -0.409*** (0.020) | 0.067*** (0.009) | 0.022*** (0.004) | | |
| Mean dep. var. | 0.917 | 5.791 | 2.930 | 0.124 | | |
| R ² | 0.157 | 0.472 | 0.211 | 0.201 | | |
| N (student * course) | 201,558 | 185,963 | 201,094 | 201,558 | | |
| N clusters (students) | 115,273 | 103,878 | 114,810 | 115,273 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.021** (0.009) | -0.199*** (0.030) | 0.024* (0.012) | 0.007 (0.012) | -0.242*** (0.085) | 0.058** (0.024) |
| Mean dep. var. | 0.716 | 2.368 | 3.343 | 0.391 | 6.606 | 2.793 |
| R ² | 0.334 | 0.511 | 0.334 | 0.592 | 0.630 | 0.647 |
| N (student * course) | 21,215 | 18,554 | 21,047 | 3,875 | 1,829 | 3,847 |
| N clusters (students) | 21,215 | 18,554 | 21,047 | 3,872 | 1,827 | 3,844 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A10: Earnings Effects of Frontier Knowledge Exposure - No Instructor Fixed Effects

| Panel (a) Undergraduate students | | | | | |
|---|---------------------|---------------------|---------------------|----------------------|---------------------|
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.043*** (0.009) | 0.040*** (0.010) | 0.045*** (0.013) | -0.033*** (0.006) | 0.001 (0.004) |
| R ² | 0.203 | 0.199 | 0.216 | 0.132 | 0.237 |
| N (students) | 201,558 | 174,428 | 73,984 | 201,558 | 201,558 |
| Panel (b) Graduate students | | | | | |
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.058*** (0.013) | 0.062*** (0.014) | 0.028 (0.021) | -0.025*** (0.007) | 0.024*** (0.007) |
| R ² | 0.379 | 0.385 | 0.366 | 0.289 | 0.387 |
| N (students) | 25,625 | 21,879 | 10,250 | 25,625 | 25,625 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A11: Educational Effects of Frontier Knowledge Exposure - Excluding Major Course Overhauls

| Panel (a) Undergraduate students | | | | | | |
|---|-------------------|----------------------|---------------------|---------------------|----------------------|--------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Proximity (sd) | -0.001 (0.005) | -0.388*** (0.022) | 0.029*** (0.011) | 0.010** (0.004) | | |
| Mean dep. var. | 0.905 | 5.780 | 2.877 | 0.105 | | |
| R ² | 0.293 | 0.553 | 0.293 | 0.223 | | |
| N (student * course) | 5,872,931 | 5,332,061 | 5,866,780 | 5,872,931 | | |
| N clusters (students) | 125,692 | 106,844 | 124,988 | 125,692 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.012 (0.010) | -0.223*** (0.033) | 0.007 (0.014) | 0.000 (0.014) | -0.362*** (0.097) | 0.057** (0.025) |
| Mean dep. var. | 0.735 | 2.541 | 3.380 | 0.292 | 6.854 | 2.758 |
| R ² | 0.425 | 0.666 | 0.452 | 0.696 | 0.802 | 0.750 |
| N (student * course) | 192,549 | 175,910 | 192,081 | 45,322 | 16,935 | 45,182 |
| N clusters (students) | 21,806 | 19,165 | 21,635 | 4,077 | 2,139 | 4,047 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations and constructed excluding courses with an update of more than one standard deviation. All specifications control for average lagged proximity, instructor fixed effects, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A12: Earnings Effects of Frontier Knowledge Exposure - Excluding Major Course Overhauls

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|---------------------|--------------------|----------------------|------------------|-----------------------|
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | |
| Proximity (sd) | 0.020** (0.009) | 0.012 (0.010) | 0.034** (0.014) | -0.009 (0.006) | 0.005 (0.005) | -0.010 (0.006) |
| R ² | 0.223 | 0.219 | 0.253 | 0.156 | 0.248 | 0.206 |
| N (student * course) | 5,872,931 | 5,188,854 | 1,876,808 | 5,872,931 | 5,872,931 | 5,872,931 |
| N clusters (students) | 125,692 | 108,493 | 47,050 | 125,692 | 125,692 | 125,692 |
| Panel (b) Graduate students | | | | | | |
| | ln(earnings) | | | Prob. earnings in... | | ln(industry earnings) |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) | |
| Proximity (sd) | 0.048*** (0.014) | 0.045*** (0.016) | 0.033 (0.023) | -0.027*** (0.008) | 0.010 (0.008) | 0.013* (0.008) |
| R ² | 0.443 | 0.454 | 0.473 | 0.372 | 0.447 | 0.468 |
| N (student * course) | 239,245 | 207,912 | 94,456 | 239,245 | 239,245 | 239,245 |
| N clusters (students) | 24,911 | 21,317 | 10,455 | 24,911 | 24,911 | 24,911 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations and constructed excluding courses with an update of more than one standard deviation. All specifications control for average lagged proximity, instructor fixed effects, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A13: Educational Effects of Frontier Knowledge Exposure - Controlling for Time-Varying Instructor Effects

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|----------------------|---------------------|---------------------|---------------------|------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Proximity (sd) | 0.020*** (0.004) | -0.265*** (0.017) | 0.047*** (0.010) | 0.013*** (0.004) | | |
| Mean dep. var. | 0.905 | 5.781 | 2.877 | 0.105 | | |
| R ² | 0.307 | 0.599 | 0.303 | 0.229 | | |
| N (student * course) | 5,992,407 | 5,441,757 | 5,988,196 | 5,992,407 | | |
| N clusters (students) | 126,150 | 107,213 | 125,440 | 126,150 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.017** (0.009) | -0.163*** (0.024) | 0.019 (0.012) | 0.009 (0.011) | -0.170** (0.075) | 0.032 (0.021) |
| Mean dep. var. | 0.735 | 2.539 | 3.383 | 0.285 | 6.855 | 2.760 |
| R ² | 0.438 | 0.709 | 0.471 | 0.719 | 0.825 | 0.751 |
| N (student * course) | 197,458 | 180,367 | 196,957 | 46,171 | 16,770 | 46,035 |
| N clusters (students) | 22,237 | 19,550 | 22,058 | 4,282 | 2,164 | 4,253 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, instructor fixed effects interacted with indicators for instructor experience equal to 0, 1-3, 4-5, and above 5, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A14: Earnings Effects of Frontier Knowledge Exposure - Controlling for Time-Varying Instructor Effects

| Panel (a) Undergraduate students | | | | | |
|---|---------------------|---------------------|---------------------|----------------------|---------------------|
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.025*** (0.008) | 0.024*** (0.009) | 0.034*** (0.012) | -0.020*** (0.005) | 0.000 (0.004) |
| R ² | 0.229 | 0.225 | 0.261 | 0.161 | 0.253 |
| N (student * course) | 5,992,407 | 5,294,724 | 1,912,523 | 5,992,407 | 5,992,407 |
| N clusters (students) | 126,150 | 108,861 | 47,292 | 126,150 | 126,150 |
| Panel (b) Graduate students | | | | | |
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.054*** (0.012) | 0.058*** (0.013) | 0.024 (0.020) | -0.023*** (0.007) | 0.020*** (0.007) |
| R ² | 0.456 | 0.472 | 0.472 | 0.387 | 0.456 |
| N (student * course) | 246,108 | 213,679 | 95,749 | 246,108 | 246,108 |
| N clusters (students) | 25,591 | 21,851 | 10,806 | 25,591 | 25,591 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations. All specifications control for average lagged proximity, instructor fixed effects interacted with indicators for instructor experience equal to 0, 1-3, 4-5, and above 5, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A15: Educational Effects of Frontier Knowledge Exposure - Using N-gram Measure of Proximity

| Panel (a) Undergraduate students | | | | | | |
|---|---------------------|----------------------|---------------------|----------------------|-------------------|---------------------|
| | Graduates | Time-to degree | GPA | Attends grad school | | |
| | (1) | (2) | (3) | (4) | | |
| Proximity (sd) | 0.033*** (0.005) | -0.207*** (0.021) | 0.073*** (0.010) | 0.009** (0.004) | | |
| Mean dep. var. | 0.905 | 5.781 | 2.877 | 0.105 | | |
| R ² | 0.294 | 0.542 | 0.292 | 0.222 | | |
| N (student * course) | 5,993,888 | 5,443,039 | 5,989,682 | 5,993,888 | | |
| N clusters (students) | 126,160 | 107,218 | 125,450 | 126,160 | | |
| Panel (b) Graduate students | | | | | | |
| | MA | | | PhD | | |
| | Graduates | Time-to degree | GPA | Graduates | Time-to degree | GPA |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proximity (sd) | 0.016* (0.009) | -0.292*** (0.036) | 0.024* (0.015) | -0.047*** (0.017) | -0.095 (0.104) | 0.079*** (0.030) |
| Mean dep. var. | 0.735 | 2.543 | 3.381 | 0.291 | 6.877 | 2.765 |
| R ² | 0.420 | 0.661 | 0.449 | 0.700 | 0.800 | 0.740 |
| N (student * course) | 199,995 | 182,775 | 199,487 | 47,395 | 17,677 | 47,255 |
| N clusters (students) | 22,309 | 19,612 | 22,128 | 4,343 | 2,198 | 4,313 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is an indicator for whether the student graduated from the program (column 1 and column 4 in panel (b)), time-to-degree in years (column 2 and column 5 in panel (b)), GPA (column 3 and column 6 in panel (b)), and an indicator for enrollment in a graduate program within Texas (column 4 in panel (a)). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations and constructed using, as dictionary, a list of the most frequent n-grams across all syllabi. All specifications control for average lagged proximity, instructor fixed effects, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Table A16: Earnings Effects of Frontier Knowledge Exposure - Using N-gram Measure of Proximity

| Panel (a) Undergraduate students | | | | | |
|---|---------------------|---------------------|--------------------|----------------------|---------------------|
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.029*** (0.008) | 0.028*** (0.009) | 0.029** (0.011) | -0.018*** (0.005) | 0.003 (0.004) |
| R ² | 0.223 | 0.218 | 0.252 | 0.155 | 0.248 |
| N (student * course) | 5,993,888 | 5,296,235 | 1,914,661 | 5,993,888 | 5,993,888 |
| N clusters (students) | 126,160 | 108,871 | 47,314 | 126,160 | 126,160 |
| Panel (b) Graduate students | | | | | |
| | ln(earnings) | | | Prob. earnings in... | |
| | 1-6 years (1) | 1-3 years (2) | 4-6 years (3) | bottom qtile (4) | top qtile (5) |
| Proximity (sd) | 0.057*** (0.014) | 0.054*** (0.016) | 0.035* (0.021) | -0.027*** (0.008) | 0.025*** (0.008) |
| R ² | 0.440 | 0.454 | 0.460 | 0.369 | 0.440 |
| N (student * course) | 248,800 | 216,320 | 97,666 | 248,800 | 248,800 |
| N clusters (students) | 25,667 | 21,925 | 10,863 | 25,667 | 25,667 |

Notes: OLS estimates; one observation is a student-course pair. The dependent variable is the natural logarithm of earnings 1-6 years (column 1), 1-3 years (column 2), and 4-6 years after predicted graduation (column 3); an indicator for earnings in the bottom quartile (column 4) and the top quartile of the graduation cohort-specific earnings distribution (column 5), and the natural logarithm of mean earnings in the first industry of employment post-graduation (column 6). The variable *Proximity* is the average frontier knowledge proximity of all courses taken by each student, measured in course-level standard deviations and constructed using, as dictionary, a list of the most frequent n-grams across all syllabi. All specifications control for average lagged proximity, instructor fixed effects, school-major-year fixed effects, and indicators for race, family income quartile, and SAT/ACT score quartile. Observations are weighted by one divided by the number of courses taken by each student. Panel (a) shows estimates for undergraduate students and panel (b) shows estimates for graduate students. Standard errors in parentheses are clustered at the student level. * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Appendix B Dataset Construction

We gathered our sample of syllabi from the websites of seven public universities in Texas. We now describe our data construction process for this sample of documents.

B.1 Texas Syllabus Data

Our syllabi sample 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), 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. These universities make these documents available for download (see Appendix Table B17), and we downloaded them between October and December 2023.¹⁴ Our sample includes a total of 459,415 documents corresponding to 27,872 courses taught up to 2022. These account for approximately 52% of all courses offered at these institutions during our analysis period.

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.1.1 Extracting A Course’s Content and Cleaning The Raw Text

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

- (i) We first convert all PDF files of syllabi to texts using PyMuPDF.¹⁵
- (ii) We use the Unidecode Python Package¹⁶ to convert Unicode text into ASCII text. This includes legacy code that does not support Unicode, non-Roman names on a US keyboard, and ASCII approximations for symbols and non-Latin alphabets.

¹⁴We 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.

¹⁵<https://pymupdf.readthedocs.io/en/latest/>

¹⁶<https://pypi.org/project/Unidecode/>

Table B17: Texas Syllabi: General Information

| Institution | Years available | Link |
|---------------------------|-----------------|---|
| Sam Houston State U | 2011-23 | https://samweb.shsu.edu/faci10wp/ |
| Stephen F. Austin State U | 2009-24 | https://orion.sfasu.edu/courseinformation/ |
| Texas A&M U | 2013-23 | 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/nlogin/?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 | 2005-24 | 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 | 2013-24 | 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.

- (iii) 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.1.2 Identifying Syllabi Sections

Most syllabi contents are organized as sections, only some of which are relevant for our analysis. The relevant sections include: instructor and course information (such as course code, course level, and title); course description, requirements, and objectives; 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 remove it. This includes the honor code, disability-related policies, classroom laptop and cellphone policies, and other policies.

To parse sections, we developed a supervised algorithm using a set of section-title keywords. The algorithm identifies a section type by searching a set of keywords associated with each category. Table B18 provides section types along with the corresponding keywords. Using these keywords,

the algorithm partitions the text into sections of the syllabus by combining keywords with each syllabus’s formatting rules. In Figure B3, 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 B3, 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 B3, we focus on highlighted sections, such as “Course Objective,” “Prerequisites,” and “Text”.

B.1.3 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).¹⁷ 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 Spacy¹⁸ 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*¹⁹; (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%.

¹⁷BiLSTM-CNNs-CRF model for named entity recognition (NER), Ma and Hovy (2016).

¹⁸<https://spacy.io/>

¹⁹<https://github.com/philipperemy/name-dataset>

Table B18: Section Title Keywords List

| Section type | Keywords |
|--|---|
| <i>Course Description</i> | Syllabi, Syllabus, Title, Description, Method, Instruction, Content, Characteristics, Overview, Tutorial, Intro, Abstract, Methodologies, Summary, Conclusion, Appendix, Guide, Document, Module, Introduction, Approach, Lab, Background |
| <i>Requirements</i> | Requirement, Applicability, Required |
| <i>Objectives</i> | Objectives, Achievement, Outcome, Motivation, Purpose, Statement, Skill, Competency, Performance, Goal |
| <i>Outline</i> | Outline, Schedule, Timeline, Guideline |
| <i>Materials</i> | Text, Material, Resource, Recommend, Reference, Book, Calendar, Textbook, Guidebook |
| <i>Instructor information</i> | Instructor, About, Email, Phone, Contact, Professor, Staff, Faculty, Information |
| <i>Projects, homework, papers, and exams</i> | Personal, Total, Individual, Exercise, Essay, Submission, Assign, Homework, Paper, Final, Examing, Midterm, Term, Semester, Proposal, Application, Demonstration, Program, Task, Report, Practical, 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 |
| <i>Grades</i> | Assessment, Point, Scope, Evaluation, Record, Grading, Composition, Review |
| <i>Other Policies</i> | Academic, Justice, Administration, Rule, Discipline, Disclaimer, Regulation, Standard, Affair, Dishonesty, Plagiarism, Misconduct, Offence, Medical, Absent, Absence, Trip, Religious, Observance, Attendance, Honesty, Origination, Originator, Help, Technology, Attendance, Accessing, Service, Opportunity, Administrative, Accommodation, Support, Policy, Right, Responsibility, Disability, Weather, Integrity, Copyright |
| <i>Notes</i> | Remark, Notice, Additional, Acknowledgement, Absolutely, Absolute, Important, Note, Cannot, Can, Must, Should, Will, Please, No |
| <i>Other Words</i> | Course, Lecture, Catalog, Campus, Community, Class, Classroom, College, University, 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.

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.

Course Field/Discipline Courses in the Texas data are categorized using the National Center for Education Statistics' Classification of Instructional Programs (CIP).²⁰ For some of our analyses, we group fields into macro-fields. The grouping is illustrated in Table B19.

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 BERT²¹ (Sanh et al., 2019), accessed via the transformers library.²² The out-of-sample prediction precision is 85.04%.

²⁰See <https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=56> for additional details

²¹<https://arxiv.org/abs/1910.01108>

²²<https://huggingface.co/transformers/index.html>

Table B19: Categorization of Course (Macro-)Fields **[REMOVE THIS]**

| 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.

B.1.4 References and Recommended Readings in Each Syllabus

To extract reference information (e.g., textbooks, articles, academic papers) from each syllabus, we employ a large language model (LLM)-based extraction pipeline. Specifically, we use OpenAI’s GPT-4.1-mini model²³ to identify and extract all course materials mentioned in each syllabus. This approach allows us to systematically identify the readings and materials assigned in each course, which we subsequently use to analyze patterns in course content and to validate our education-innovation gap measure. Next, we use regular expressions to extract publication years (when available) from the extracted reference strings. In our sample, we successfully extract reference information for 80.4% of non-empty syllabi and obtain at least one publication year for 42.0% of non-empty syllabi; among syllabi with extracted years, we obtain an average of 1.77 publication years per syllabus. The prompt used for extraction is shown below:

Prompt for Reference Extraction

You will be given a syllabus text. Your task is to extract reference information from this text without making any changes to the content. To extract the reference information, follow these steps:

- 1. Look for sections in the syllabus that contain information about course materials, textbooks, or references. Common prompt words to look for include: Materials, Text, Material, Resource, Recommend, Reference, Book, Calendar, Textbook, Guidebook.*
- 2. Identify sections that list specific items like books, articles, or academic papers.*
- 3. Extract only those parts listing the materials—omit any sentences about how or when to access, read, or bring the materials.*
- 4. If multiple relevant sections are found, extract all and merge them into one for the output, separating each source with “|”.*
- 5. If you don’t find any relevant sections containing reference information, respond with “None”.*

Output Generation: Output only the TSV data without any additional text or formatting. Include columns: Reference. Ensure proper TSV formatting with no extra line breaks or spaces. Use “None” to represent missing values.

B.2 Academic Publications

To construct our measure of frontier knowledge proximity, we collect a large sample of academic articles from academic journals. We describe here how this sample is defined, constructed, and

²³<https://platform.openai.com/docs/models/gpt-4.1-mini>

collected.

B.2.1 Collecting Academic Articles

We collect metadata on academic articles from OpenAlex. OpenAlex is an open scholarly knowledge graph maintained by OurResearch that provides large-scale metadata and linkage information on research outputs and related entities. It covers works (e.g., journal articles, conference papers, and preprints) as well as authors, institutions, sources (journals and venues), topics, publishers, and funders, and encodes relationships such as authorship, affiliations, citations, and topical classifications. OpenAlex is openly licensed under CC0 and is accessible via a public API and periodic data snapshots.

We use the following variables:²⁴

- `id`: OpenAlex ID, used as the unique identifier of each article;
- `title`: title of the article;
- `issn`: ISSN of the journal;
- `publication_date`: publication date;
- `abstract_inverted_index`: abstract (in inverted index format);
- `keywords`: author-provided keywords.

From an initial pool of 261,381,162 OpenAlex records, we identified 133,379,066 valid entries and ultimately retained 107,911,700 English-language articles with non-empty abstracts for our analysis.

B.2.2 Data Cleaning

The main information from academic articles that we use in our analysis is the abstract. In OpenAlex, abstracts are stored in the `abstract_inverted_index` field as an inverted index, where each unique word is mapped to the positions it appears in the text. We first reconstruct the full abstract text by reversing this inverted index structure.²⁵

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

²⁴<https://openalex.org>. The full list of variables available through OpenAlex is documented at <https://docs.openalex.org/api-entities/works/work-object>

²⁵For example, an inverted index `"Purpose": [0], "of": [1], "study": [2]` is converted to "Purpose of study".

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 OpenAlex to measure the research productivity of all people listed as instructors in the syllabi. We download these data from OpenAlex. Because databases differ in coverage, deduplication, and citation-matching rules, OpenAlex citation links and citation counts may not exactly coincide with those reported by services such as Web of Science, Scopus, or Google Scholar; we therefore interpret citation-based measures within the OpenAlex definition and document the data snapshot date used in our analyses.

We link instructor records from the text of the syllabi to OpenAlex records using names, a person’s history of institutions, and research fields. In the Texas sample of syllabi, all (513,141) have an instructor record, covering 29,495 unique instructors. Of these instructors, 43.3% (12,771 / 29,495) are matched to OpenAlex profile.

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

We collect information on grants awarded by the National Science Foundation (NSF)²⁶ and the National Institutes of Health (NIH)²⁷ 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 insti-

²⁶<https://www.nsf.gov/awardsearch/download.jsp>

²⁷https://exporter.nih.gov/ExPORTER_Catalog.aspx

tution 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).²⁸ 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.

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 Package²⁹, 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.

²⁸The 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.

²⁹<https://github.com/rlieb/whoswho>

- (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.

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. In the Texas sample, 2.18% ($= 644 / 29,495$) of the instructors are connected with NSF investigators, corresponding to 9,951 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. For the Texas sample, 7.59% ($= 2,240 / 29,495$) of the instructors are connected with NIH investigators, corresponding to 11,721 grants. Our final grant data combine information from NSF and NIH grants. The syllabi sample used in our analysis covers 29,495 instructors, of whom 8.93% ($= 2,633 / 29,495$) have at least one NSF or NIH grant, accounting for 6.24% ($= 32,021 / 513,141$) 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 Text Pre-processing

Before extracting terms or constructing dictionaries, 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.³⁰ 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.³¹

C.2 Dictionary Construction

The core of our analysis relies on a dictionary, i.e., a list of all knowledge terms. We consider two approaches for constructing this dictionary.

Approach 1: Academic Publications Dictionary The first approach is to use the list of all unique words and expressions ever used as keywords in academic publications. We extract these keywords from the data described in Section B.2.

Approach 2: Syllabi-Based Empirical Dictionary In contrast to the pre-defined academic keyword list, our second dictionary is constructed empirically from the syllabus corpus itself. This data-driven approach ensures that our terminology reflects the actual pedagogical vocabulary used in higher education, rather than solely research-oriented jargon.

To construct this dictionary, we begin with the cleaned text described in Section C.1 and apply two additional refinement steps:

³⁰<https://pypi.org/project/Unidecode/>

³¹We 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>.

1. **Context-Specific Filtering:** We remove dates, timestamps, and a curated list of syllabus-specific stop words (e.g., administrative terms like "office hours" or "prerequisites") to reduce noise.
2. **Frequency Thresholding:** We extract all unique N-grams ($N \in [1, 3]$) and impose a minimum document frequency threshold. Only terms appearing in at least 100 distinct syllabi are retained. This threshold is critical for filtering out idiosyncratic terms and preserving only generalizable knowledge concepts.

The resulting dictionary contains 691,769 unique terms, approximately 50% of which intersect with the academic publications dictionary. The term frequency distribution exhibits a characteristic heavy-tailed structure. For instance, while the dictionary is large, the "core" vocabulary is concentrated: only 9,204 terms appear in at least 1% of the syllabi, and 118,099 terms appear in at least 0.1%. Conversely, the vast majority of terms appear infrequently, reflecting the specialized nature of advanced coursework.

C.3 Term Extraction Algorithm

Once we have cleaned the text and established a dictionary, we convert the documents 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 3, 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.

Figure B3: 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 | <u>COMPUTATIONAL METHODS FOR ECONOMIC DYNAMICS</u> | | ECON 561a |
| <u>Course Objectives:</u> | | | |
| <p>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.</p> | | | |
| <u>Contact Information</u> (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 | | | |
| <u>Class Meetings:</u> | | | |
| The course meets on Mondays and Wednesdays from 2:30PM to 3:50PM in a room to be determined. | | | |
| <u>Prerequisites:</u> | | | |
| 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. | | | |
| <u>Texts:</u> | | | |
| 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: | | | |
| <ul style="list-style-type: none"> • 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). | | | |
| <u>Grading:</u> | | | |
| 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. | | | |
| <u>Approximate Schedule of Lectures</u> (Part I) | | | |
| I. INTRODUCTION | | | |
| Lecture 1 Introduction to numerical dynamic programming (built around the stochastic growth model and the Aiyagari (1994) model). General considerations in numerical analysis: convergence, roundoff error, truncation error. Numerical differentiation. | | | |
| Readings: | | | |
| <ul style="list-style-type: none"> • 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: | | | |
| <ul style="list-style-type: none"> • Numerical Recipes: Chapter 9 <p>.....</p> | | | |

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

Appendix D Texas ERC Data: Sample and Variable Definitions

D.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.

D.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 $t+6$ for undergraduates and $t+3$ for graduate students.