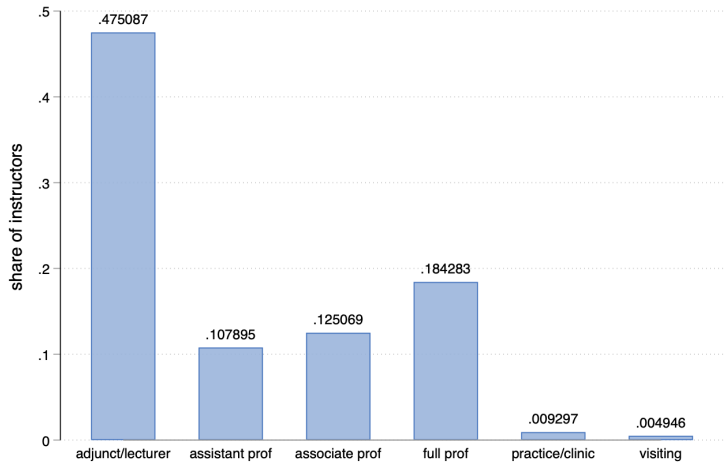
Appendix

For online publication only

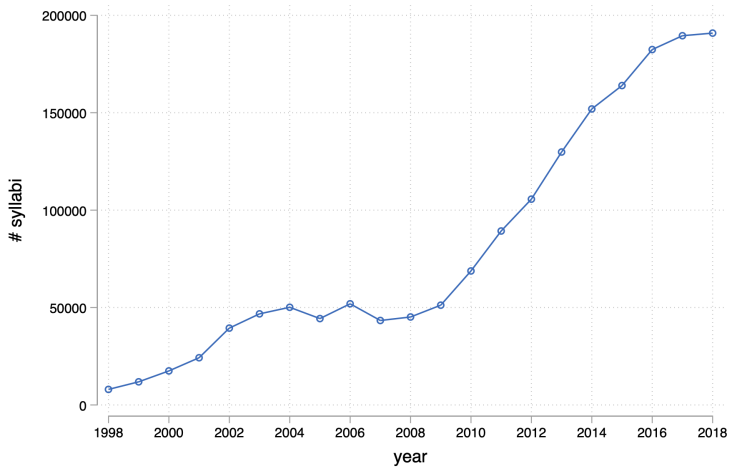
Appendix A Additional Tables and Figures

Figure AI: Distribution of Instructor Job Titles



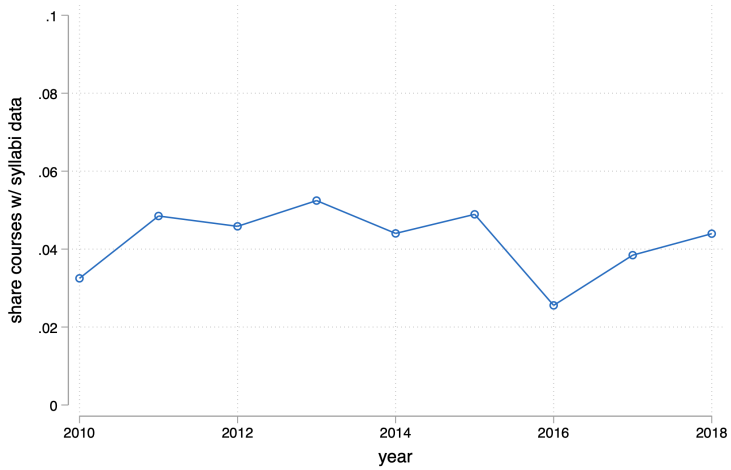
Note: Share of syllabi instructors by job title. The sample is restricted to 32,090 instructors in public institutions for whom title information is available.

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



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

Figure AIII: Share of Catalog Courses in the Syllabi Sample

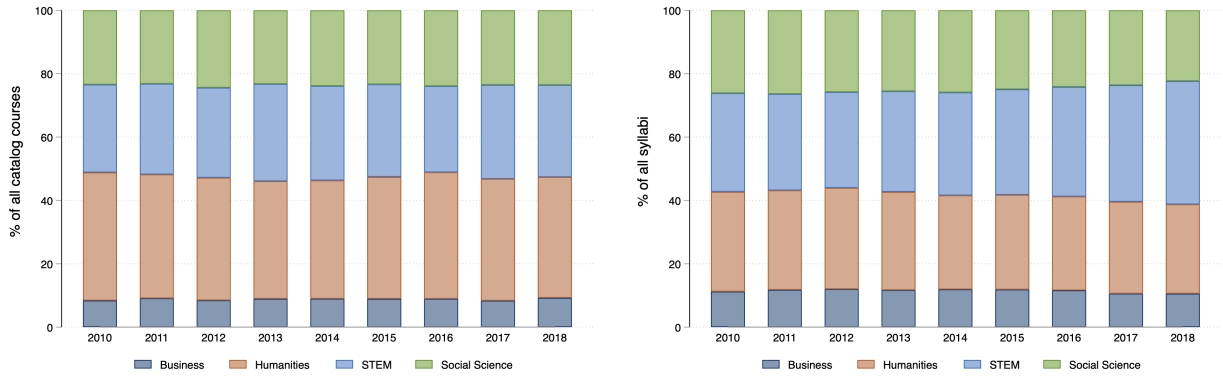


Note: Share of courses from full course catalogs whose syllabi are included in the syllabi sample.

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

(a) Course Catalog

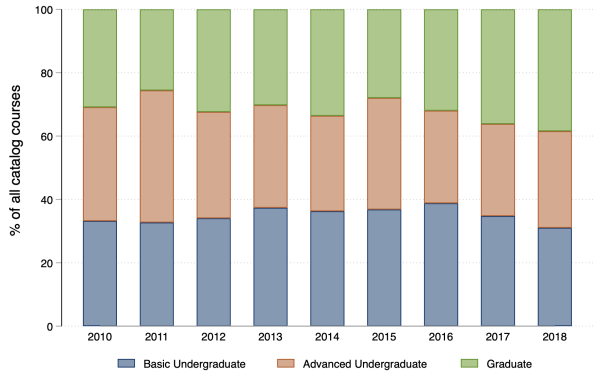
(b) Syllabi Sample



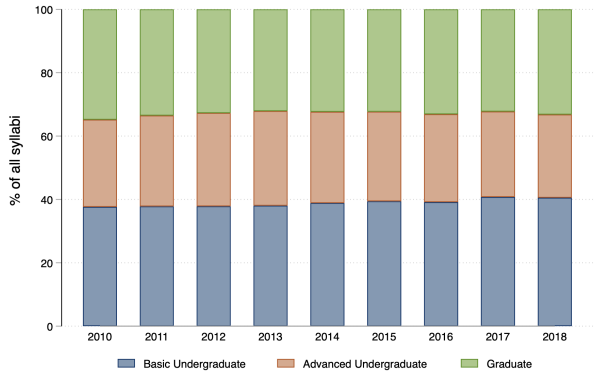
Note: Composition across macro fields, for all courses included in a sample of school catalogs (panel (a)) and for courses included in the syllabi sample (panel (b)).

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

(a) Course Catalog

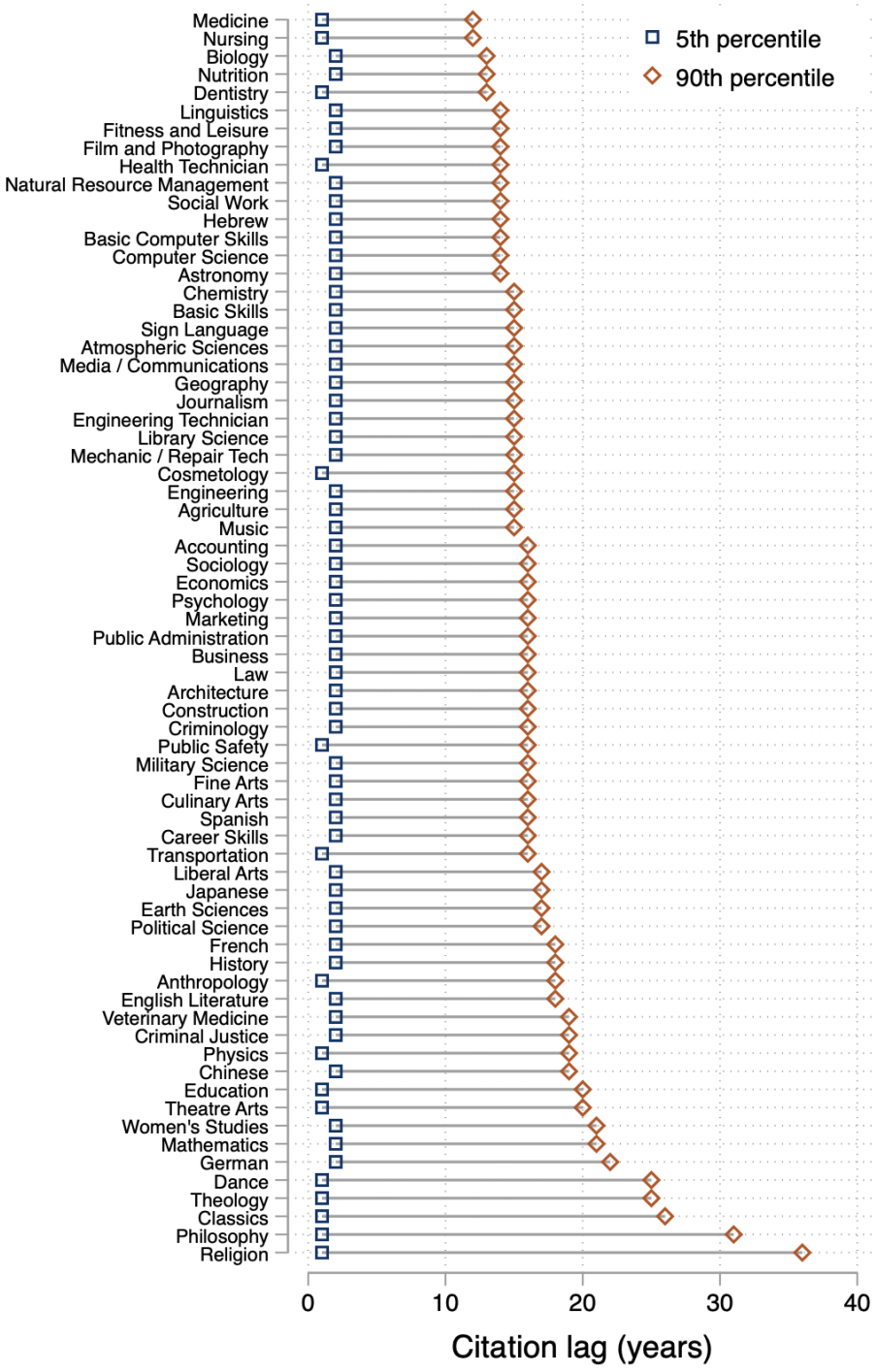


(b) Syllabi Sample



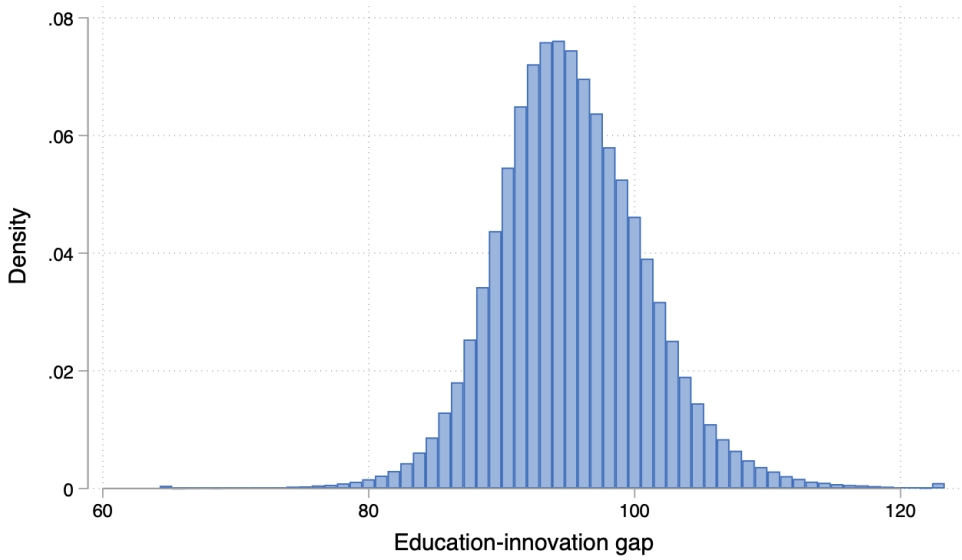
Note: Composition across course levels, for all courses included in a sample of school catalogs (panel (a)) and for courses included in the syllabi sample (panel (b)).

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



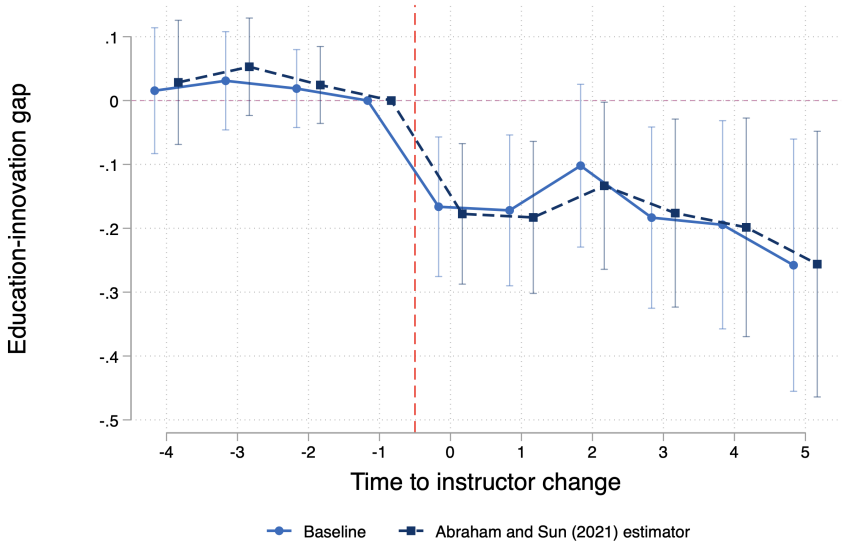
Notes: 5th and 90th percentile of the citation lag distribution in each field, used to calculate the education-innovation gap for the syllabi in each field.

Figure AVII: Education-Innovation Gap: Distribution



Notes: Histogram of the education-innovation gap.

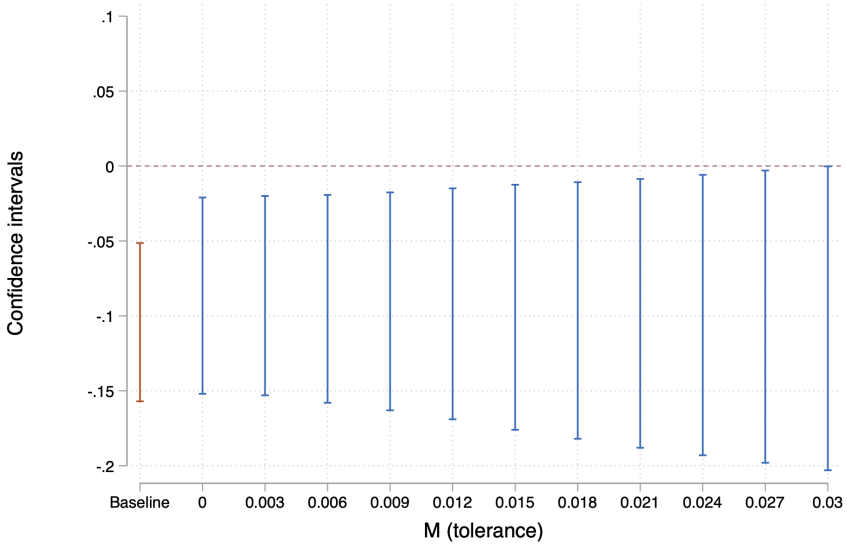
Figure AVIII: Event Study of The Gap Around an Instructor Change: Baseline and Abraham and Sun (2021) Estimator



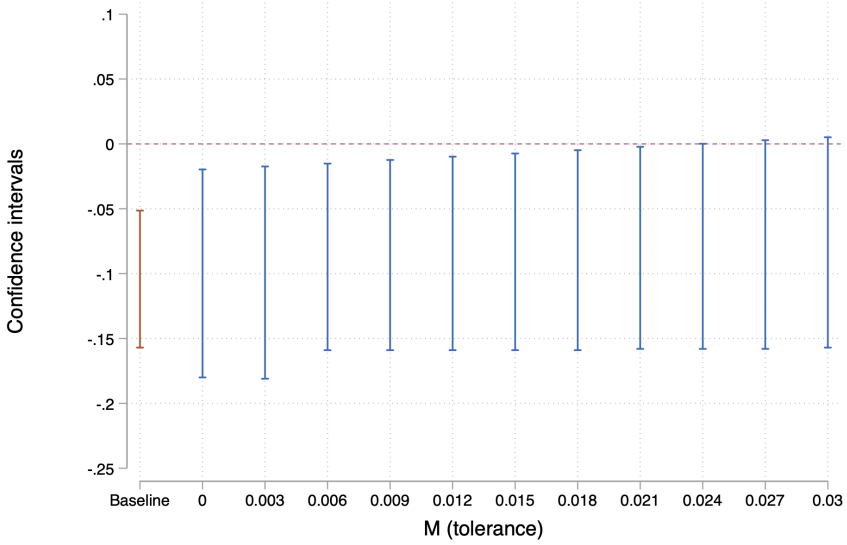
Notes: Estimates of the confidence intervals of δ_k in equation (5), obtained using the baseline approach used in the paper (solid series) and the estimator developed by Sun and Abraham (2021) (dashed series), which accounts for the possibility of heterogeneous treatment effects across cohorts of treated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

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

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



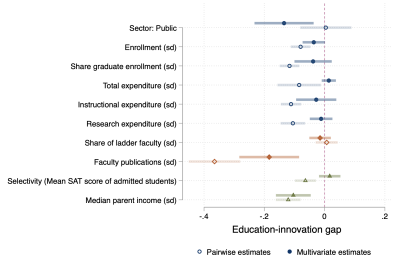
(b) With monotonicity ($\Delta^{SDD}(M)$)



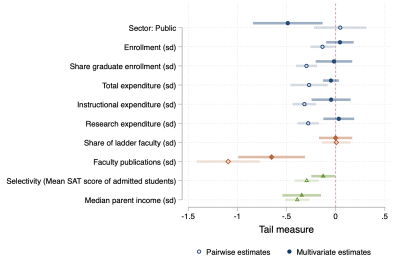
Notes: Sensitivity plots of the confidence intervals of δ_0 in equation (5), constructed following the approach of Rambachan and Roth (2019). The approach tests for violations of the parallel trends assumption and studies their impacts on the point estimates and confidence intervals of interest. Specifically,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

Figure AX: School Characteristics and Alternative Measures of Course Novelty

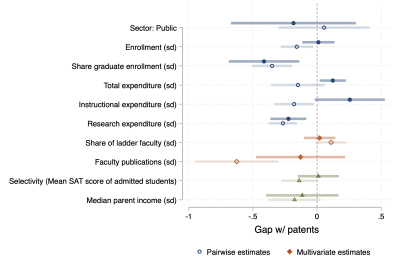
(a) % non-covered new knowledge



(b) Tail measure



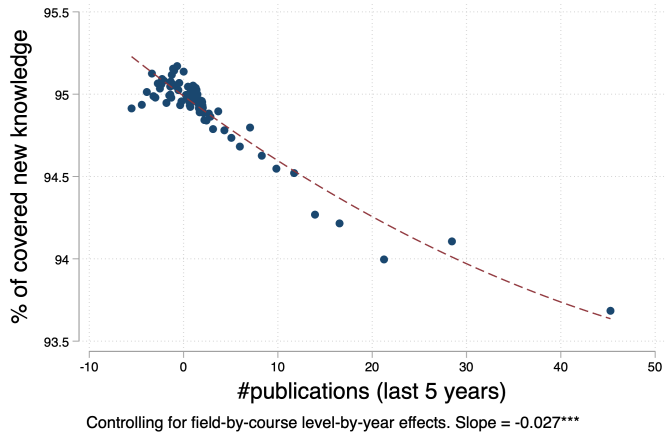
(c) Gap with patents



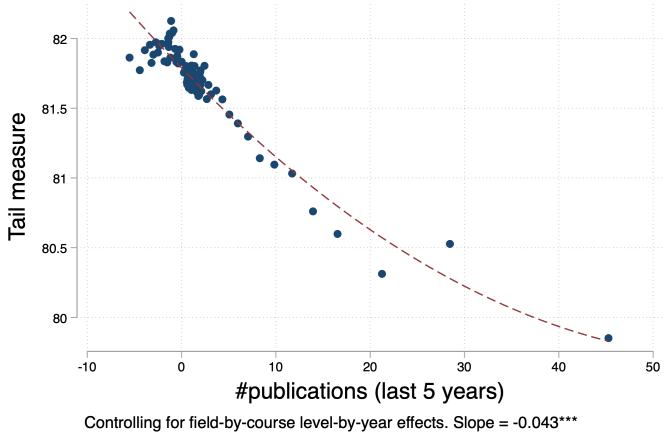
Notes: Point estimates and 95-percent confidence intervals of coefficient β in equation (7), using three alternative measures of course novelty: a measure of non-covered new knowledge, defined as one minus the share of all new words contained by each syllabus (where new words are knowledge words that are (a) in the top five percent of the word frequency among articles published between $t-3$ and $t-1$ or (b) used in articles published between $t-3$ and $t-1$ but not in those published between $t-15$ and $t-13$, panel a); a "tail measure," calculated for each syllabus by (a) randomly selecting 100 subsamples containing 20 percent of the syllabus's words, (b) calculating the gap for each subsample, and (c) selecting the 5th percentile of the corresponding distribution (panel b); and the education-innovation gap calculated using the text of all patents as a benchmark, instead of academic articles (panel c). Each coefficient is estimated from a separate regression, with the exception of selectivity tiers (Ivy Plus/Elite, Highly Selective, Selective) which are jointly estimated. Endowment, expenditure, and share minority information refers to the year 2018 and is taken from IPEDS. Estimates are obtained by pooling syllabi data for the years 1998 to 2018. Standard errors are clustered at the school level.

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

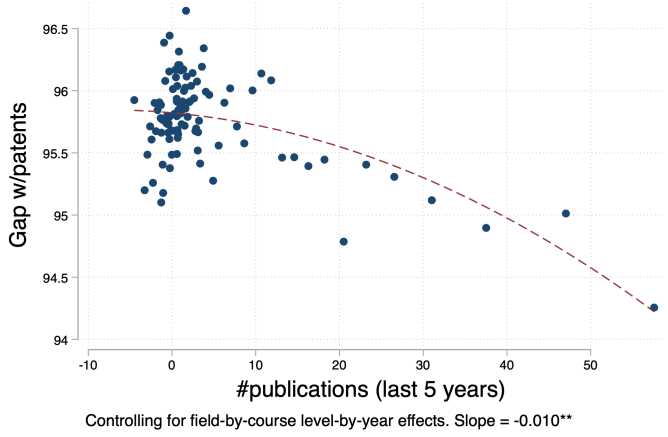
(a) % non-covered new knowledge



(b) Tail measure

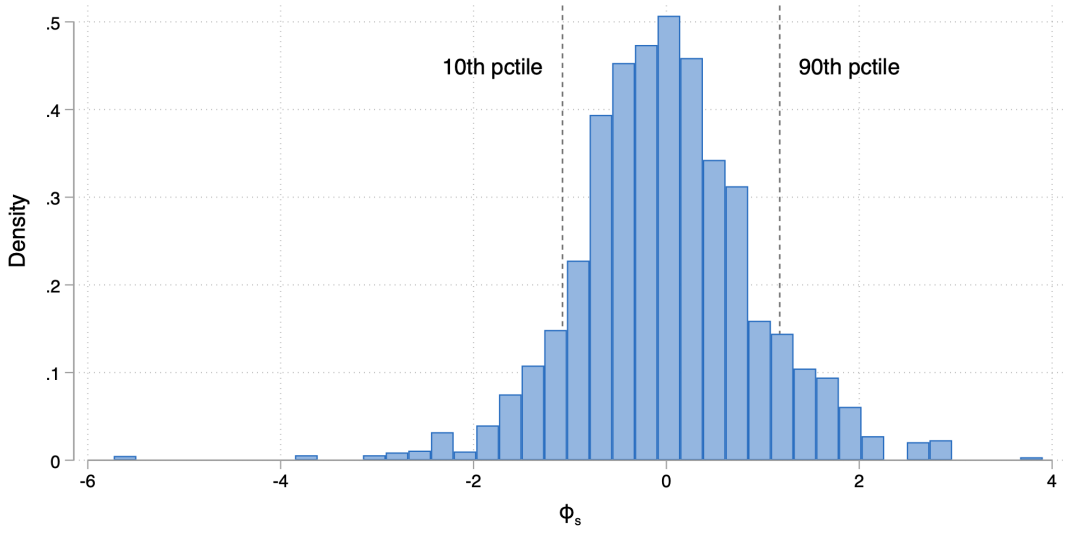


(c) Gap with patents



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

Figure AXII: Distribution of School-Level Gap



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

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

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

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

Table AII: Alternative Measures of Novelty and Student Outcomes

		Income (College Scorecard)			Income (Chetty et al., 2020)				
	Grad rate	Mean	$P_y \leq 33$ pctile	Median	Mean	P(top 20%)	P(top 10%)	P(top 5%)	P(top 20%| $P_y \leq 20$ pctile)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel (a): Share of non-covered new knowledge, no controls									
Gap (sd)	-0.0401***	-0.0528***	-0.0580***	-0.0434***	-0.0675***	-0.0300***	-0.0265***	-0.0197***	-0.0273***
	(0.0079)	(0.0102)	(0.0111)	(0.0086)	(0.0122)	(0.0063)	(0.0047)	(0.0034)	(0.0061)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (b): Share of non-covered new knowledge, with controls									
Gap (sd)	-0.0032	-0.0043	-0.0028	-0.0027	-0.0117**	-0.0053*	-0.0040*	-0.0027*	-0.0005
	(0.0032)	(0.0047)	(0.0059)	(0.0048)	(0.0046)	(0.0031)	(0.0023)	(0.0015)	(0.0034)
Mean dep. var.	0.5816					0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					
Panel (c): Tail measure, no controls									
Gap (sd)	-0.0537***	-0.0644***	-0.0703***	-0.0572***	-0.0886***	-0.0389***	-0.0336***	-0.0248***	-0.0372***
	(0.0083)	(0.0097)	(0.0114)	(0.0087)	(0.0119)	(0.0057)	(0.0046)	(0.0034)	(0.0057)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763
# schools	761	760	734	760					
Panel (d): Tail measure, with controls									
Gap (sd)	-0.0020	-0.0094**	-0.0140***	-0.0107**	-0.0172***	-0.0101***	-0.0081***	-0.0052***	-0.0095***
	(0.0034)	(0.0043)	(0.0053)	(0.0046)	(0.0048)	(0.0028)	(0.0021)	(0.0013)	(0.0031)
Mean dep. var.	0.5816					0.3710	0.2100	0.1159	0.2957
N	11471	1996	1843	1996	718	718	718	718	718
# schools	733	727	701	727					
Panel (e): Gap w/patents, no controls									
Gap (sd)	-0.0232***	-0.0323***	-0.0434***	-0.0282***	-0.0404***	-0.0144**	-0.0140**	-0.0120***	-0.0146**
	(0.0068)	(0.0116)	(0.0122)	(0.0099)	(0.0138)	(0.0067)	(0.0059)	(0.0042)	(0.0064)
Mean dep. var.	0.5692					0.3694	0.2082	0.1143	0.2945
N	15683	3793	3566	3793	763	763	763	763	763

(Continued)

Table AII. Continued

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

Note: OLS estimates of the coefficient δ in equation (9). The variable *Gap (sd)* is a school-level alternative measure of the education-innovation gap (estimated as $\theta_{s(i)}$ in equation (8)), standardized to have mean zero and variance one. In panels (a) and (b), the alternative measure is the share of non-covered new knowledge; in panels (c) and (d) it is a "tail measure;" and in panels (e) and (f) it is the education-innovation gap calculated using the text of all patents as a benchmark for frontier knowledge. The dependent variables are graduation rates (from IPEDS, years 1998-2018, column 1); the log of mean student incomes from the College Scorecard, for all students (column 2) and for students with parental income in the bottom tercile (column 3); the log of median income from the College Scorecard (column 4); the log of mean income for students who graduated between 2002 and 2004 (from Chetty et al. (2020), column 5); the probability that students have incomes in the top 20, 10, and 5 percent of the national distribution (from Chetty et al. (2020), columns 6-8); and the probability that students with parental income in the bottom quintile reach the top quintile during adulthood (column 9). Columns 1-4 in panels a and b control for year effects. All columns in panels b, d, f, and h control for sector (private or public), selectivity tiers, and an interaction between selectivity tiers and an indicator for R1 institutions according to the Carnegie classification; student-to-faculty ratio and the share of ladder faculty; total expenditure, research expenditure, instructional expenditure, and salary instructional expenditure per student; the share of undergraduate and graduate enrollment and the share of white and minority students; an indicator for institutions with admission share equal to 100, median SAT and ACT scores of admitted students in 2006, and indicators for schools not using either SAT or ACT in admission; the share of students with majors in Arts and Humanities, Business, Health, Public and Social Service, Social Sciences, STEM, and multi-disciplinary fields; and the natural logarithm of parental income. Bootstrapped standard errors in parentheses are clustered at the school level. $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Appendix B Dataset Construction

B.1 Syllabi

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

We use the following variables from the OSP database:

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

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

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

(ii) Taught in English;

(iii) Taught between 1998 and 2018;

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

[30] https://opensyllabus.org

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

Table BIII: Summary Statistics of Open Syllabus Project

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

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

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

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

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

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

B.1.2 Cleaning The Raw Text

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

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

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

B.1.3 Identifying Syllabi Sections

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

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

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

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

Table BIV: Section Title Keywords List

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

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

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

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

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

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

B.1.4 Extracting Additional Information

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

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

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

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

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

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

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

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

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

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

B.1.5 References and Recommended Readings in Each Syllabus

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

- `MatchID`: The unique identifier of the match

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

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

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

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

B.1.6 Syllabi Field

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

Table BV: Categorization of Course (Macro-)Fields

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

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

Figure BXIII: Dividing A Syllabus Into Sections: An Example

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

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

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

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

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

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

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

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

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

……

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

B.2 Academic Publications

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

B.2.1 List of Top Journals

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

B.2.2 Collecting Academic Articles

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

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

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

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

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

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

B.2.3 Data Cleaning

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

B.3 Research Productivity

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

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

B.4 Patents

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

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

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

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

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

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

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

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

B.5.1 Linking NSF/NIH Institutions to Syllabi Institutions

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

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

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

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

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

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

B.5.2 Linking NSF/NIH Investigators to Instructors

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

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

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

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

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

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

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

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

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

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

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

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

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

B.6 Instructors' Job Titles and Salaries

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

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

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

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

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

Table BVII: Assigning Job Titles

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

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

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

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

C.1 Extracting Knowledge Terms From Each Document

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

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

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

C.2 A Simulation Exercise

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

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

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

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

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

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

To generate each syllabus, we use the following algorithm:

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

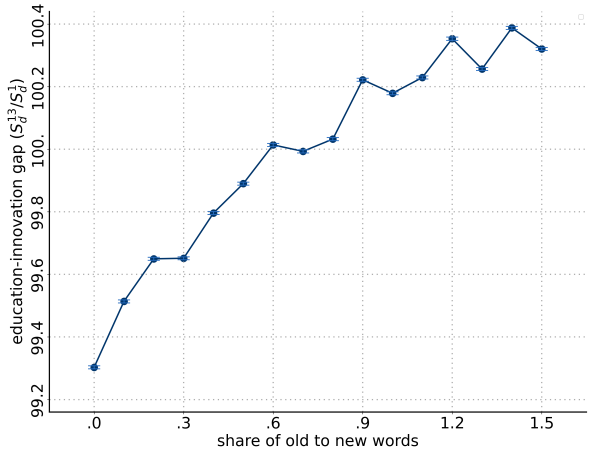
With this algorithm, we generate 10 syllabi for each set of parameters $\{t, L, style, gap\}$. The total number of generated syllabi is thus $= 10 \times 21 \times 46 \times 11 \times 16 = 1,700,160$, which is close to the sample size in our data.

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

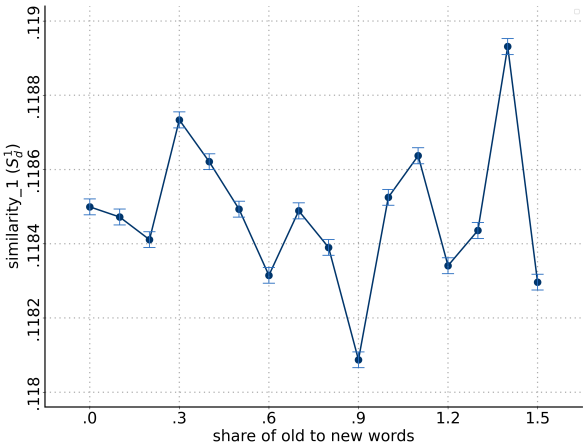
we show the relationship between gap and the cosine similarity between the syllabus and new publications (appeared in $t-3$ to $t-1$), i.e., the denominator of the education-innovation gap. This relationship is much noisier. This is likely to occur because a simple cosine similarity is likely to be affected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

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

(a) gap and The Education-Innovation Gap



(b) gap and Cosine Similarity with New Publications



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