How have the skills college graduates developed changed over time? Evidence from 3 million course descriptions

[Click here for latest version]

Jacob Light

Working Paper
December 2020

Abstract

Skill-biased technological change has advantaged highly educated workers relative to workers with less education. The link between education and skill complementarity with technology is not obvious: differences across institutions and fields of study yield substantial variation in students' skill development, and incentives for institutions to emphasize new skills are theoretically ambiguous. I provide novel evidence of institution responses to changing skill demand using a unique dataset of course descriptions scraped from a diverse sample of colleges and universities in the United States. Course descriptions are a snapshot of a university's offerings and provide insight into its instructional priorities and resources. I find that college courses increasingly emphasized cognitive and social skill development as demand for workers with strong cognitive and social skills grew in the labor market. Between 2012-2020, courses intensive in cognitive or social skills each grew by 11% as a share of total course offerings. The paper provides some of the first analysis of within-institution course supply responses to changes in student and labor market demand.

Introduction

Skill-biased technological change is among the most disruptive forces in the American economy over the last 40 years. Technology has become a substitute for routine labor in many applications, leading to a polarization of the labor force whereby workers with skills complementary to technology experienced rising earnings and steady employment, while workers with skills substitutable by technology experienced disruption and higher rates of unemployment (Autor et al. (2003); Autor and Dorn (2013)). The evidence of technology-driven labor force disruption is evident today, as firms increasingly seek job candidates with higher levels of education, experience, and skills (Hershbein and Kahn (2018)). Beyond labor demand, technology-driven disruption is also evident in wages: the returns to a college degree have grown dramatically since the 1980s, as high-skilled workers disproportionately reap the benefits of changing production technology (Goldin and Katz (2007)).

Workers have responded to these changes in the type of labor demanded by firms, in part, by participating in higher education. During the same period of technology-driven workforce disruption, college attendance rates increased substantially (Bound et al. (2010)). Moreover, there is strong evidence that college enrollment peaks during economic downturns (Long (2004)). Higher education has been both a response to and insurance against change. However, the channel through which a college degree equips workers with the skills they need for jobs in a changing economy is not obvious. Colleges and universities instruct students in disciplines directly impacted by technological change, such as STEM, but also in disciplines less connected to the technologies driving skill-biased technological change, including some social science and humanities fields. Whether and how colleges and universities have an objective or prerogative to respond across disciplines to changing skill demand are open questions.

In this paper, I study how colleges respond to changing skill demand. The paper uses a new data source, a collection of course catalogs scraped from a large sample of colleges and universities in the United States, to study the skill content of college degrees and document changes in skill offerings across institutions, majors, and time. Course catalogs contain rich detail of the skills and learning objectives prioritized by students and institutions at fixed points in time. I extract course descriptions from these catalogs to construct a dataset that tracks changes in skill offerings over time and differences across institutions.

The purpose of this paper is to validate the use of catalog data as a measure of students' skill development. First, I examine how a school's stock of courses evolves over time. I study changes along two margins: an intensive margin, where existing courses are updated to reflect new content or objectives, and an extensive margin, where new content is introduced in entirely new courses. I find that courses are relatively stable along the intensive margin: among the set of courses offered continuously between 2012-2020, the average course description was updated only once during the eight year period. Along the extensive margin, I document substantial growth in course offerings over the between 2012-2020. The growth is most striking in STEM and Computer Science fields, consistent with contemporaneous growth in degree completions in these fields (NCES (2019)).

Second, I measure the skills a student develops in college using the course description data. Motivated by Deming and Kahn (2018), I score course descriptions for exposure to skills in six skill domains associated with higher earnings: social, cognitive, reading, writing, basic computer, and programming skills. I consider a course to be "intensive" in a skill domain if the course description contains keywords or phrases linked to the skill. I demonstrate that these skills associated with high earnings are common in college courses and appear with increasing frequency over time. This is particularly the case for cognitive and social skills.

The results in this paper suggest that colleges and universities respond to evolving skill demand in the labor market. The paper proceeds as follows: first, I introduce the course catalog dataset; second, I summarize the process of course evolution over time; third, I validate the use of course descriptions to link college courses to skills and demonstrate that cognitive and social skills are increasingly emphasized in college courses; finally, I conclude by discussing the policy implications of my descriptive findings.

Data and Descriptive Analysis

In this paper, I introduce a novel dataset constructed from course descriptions scraped from course catalogs for a sample of non-profit, degree-granting Title IV colleges and universities in the United States. Colleges and universities publish course descriptions in course catalogs released annually or biannually. I extract information from course descriptions published by colleges and universities to characterize the skills students develop through their college classes and track changes in these skills over time. The dataset currently contains a diverse and growing sample of nearly 150 institutions.

Course catalogs are a useful data source for a number of reasons. Catalogs are a snapshot of a university's offerings and provide insight into its instructional priorities and resources. The wide availability of course catalogs makes it possible to build a large, nationally representative data sample. In addition, the common structure of entries in course catalogs ensures that the extracted data are comparable across time and institutions. An entry typically includes a unique course ID and descriptive summary of the course content; course descriptions may also list pre-requisite courses, cross-listed courses, and the name of the course instructor. Finally, catalogs expose networks of courses (e.g. through cross-listing or pre-requisites) useful for demonstrating overlap or polarization of disciplines over time. An example of a course catalog entry is provided in Figure 1.

Analysis in this paper relies on the count of courses offered and the text descriptions of courses in the catalogs. Course counts and text descriptions alone are noisy measures of students' classroom experiences. Course counts may be inaccurate to the extent that the dataset does not observe how frequently a course is offered during the year, if at all; course descriptions may not update as frequently as course content updates. In this paper, I emphasize inference from changes in course counts and descriptions over time. Emphasis on the trend rather than levels should reduce some noise inherent in the data source.

The institutions sampled for this paper are reflective of the diversity of colleges and universities in the United States. The sample includes two- and four-year institutions, public and private institutions, highly selective and less selective institutions, and institutions from nearly every state. Thus, the sample is meant to be generally representative of the typical college or university in the United States. Table 1 compares characteristics of institutions in my sample to the broader population of colleges and universities. The table suggests that the sample resembles the typical college or university demographically and in various measures of selectivity. The typical institution in my sample is slightly larger than the average American college. Figure 2 summarizes coverage of the dataset over time and various institution characteristics. Coverage of the dataset is densest in recent years (2012-onward). In time series analyses throughout this paper, I restrict the sample to a panel of institutions that appear in my dataset in each year from 2012-13 to 2019-20.

Course Evolution

An institution can adjust course offerings to meet student demand along intensive and extensive margins. Institutions adjust along the intensive margin by modifying the content of an existing course, while adjustment along the extensive margin involves creating a new course to meet student demand. Figures 3 and 4 measure course growth along intensive and extensive margins for colleges in my sample.

I measure intensive margin responses by calculating the survival rate of course descriptions for courses offered in 2012-13. In each year, a previously-offered course can transition into one of three states: a course is discontinued when it is not offered in any future years; a course changes description if an institution continues to offer the course but the description does not match the course description in 2012-13; a course is unchanged if the course is offered and the description is identical to the 2012-13 description. Figure 3 plots the evolution of courses over time. The figure suggests that existing courses evolve slowly. Conditional on not being discontinued, the typical course is updated once during the eight-year period. Survival rates are comparable across broad categories of majors.

I measure extensive margin responses as growth or reduction in the stock of courses offered over time. In Figure 4, I track growth in course offerings over time. The figure counts the number of courses offered in each academic year for the panel of institutions covered 2012-2020 in my dataset, indexed as a share of courses offered in 2012-13. To demonstrate heterogeneity in the growth trends of course offerings, I plot growth rates for courses classified into five department categories: Business/Economics, Education, Social Sciences/Humanities, STEM excluding Computer Science, and Computer Science.

Total course offerings grew between 2012-2020. Secular growth is consistent with recent trends in higher education, such as increasing college enrollment, competition for smaller course sizes, and expanded access to colleges through the Common App Joseph et al. (2012). Growth rates, however, differ across major categories. In particular, course offerings in STEM and Computer Science have grown dramatically over the last 5 years. The growth in Computer Science courses coincides with the introduction of information and data science courses at many colleges, as well as steadily increasing growth in the share of college graduates who major in STEM fields (NCES (1995-2018)).

Taken together, these figures document three new facts. First, courses offered over a long period of time are quite stable. The typical course description for a course continuously offered updates approximately once every eight years. Second, new course offerings have grown in the last ten years. The general upward trend coincides with growing college enrollment and competition for smaller class sizes. Third, Computer Science courses have expanded substantially since 2015-16. These trends provide useful context for analysis of skills embedded in course descriptions as described below.

Skill Content

I next use the course catalog data to show that college students are developing skills associated with higher earnings. Deming and Kahn (2018) define 10 skill domains broadly applicable across occupations and show that local markets where these skills are in greater demand are markets where average earnings are higher. A natural question arising from their work is where students develop these skills associated with high earnings. Among Deming and Kahn's skill domains, five skills stand out as skills that are learned or can be practiced in the classroom: cognitive, social, writing, computer (basic), and programming.[1] Students develop "cognitive" skills in courses that involve research, problem solving, and critical thinking. Students develop "social" skills in courses with group projects or where students present their work to the class. I track writing skills, and add reading skills, as two of the most basic and broadly applicable skills a student develops in college. The final two skill domains - basic and advanced computer skills - represent skill domains of increasing technological importance Beblavỳ et al. (2016). I distinguish basic computer skills, such as computer literacy or using simple software, from advanced computing skills, which involve knowledge of specific programming languages and applications of computer science. Following Deming and Kahn (2018), I summarize keywords and phrases associated with each skill domain in Table 2. I characterize a course as "intensive" in a skill domain if its description contains at least one of the keywords associated with the skill.

I validate the application of Deming and Kahn's skill domains to measure skills developed in college in two stages. First, I demonstrate that students are widely exposed to these skills associated with high earnings in college classes. Figure 5 summarizes skill exposure across six skill domains using course descriptions from the 2019-20 academic year. The bars show the percent of descriptions for all courses that contain at least one word or phrase matching Deming and Kahn's keyword list.[2] Among the skill domains, college courses are most intensive in cognitive and social skills, the two categories Deming and Kahn highlight as associated with high earnings and in high demand across a diverse set of occupations. I observe references to reading and writing each in approximately 10% of course descriptions. Less frequent are references to computer and programming skills. Table 3 lists the highest- and lowest-scoring department groups in each of the six skill domains. The results align with reasonable expectations: the department most intensive in reading and writing skills

[1]Deming and Kahn's complete list of high-earnings skills includes cognitive, social, character, writing, customer service, project management, people management, financial, general computer skills (basic computing), and specific computer skills (programming and specialized software.)

[2]The keyword search approach is limited by an inability to detect "up-skilling" within a skill domain. For example, a course that replaces an algebra requirement for calculus demands greater skills from students but is equally "cognitive" in a binary skill score. This limitation suggests natural language processing tools as an alternate/preferred skill exposure scoring technique.

is Literature and Languages; the department most intensive in computer and programming skills is Computer Science; the department most intensive in cognitive skills is Mathematics and Statistics.[3]

Second, I show that growing demand for the skills associated with high earnings is generally reflected in growing emphasis on these skills in college classes. Figure 6 plots changes in the share of courses intensive in each of the six skill domains between 2012-2020. As in 5, references to cognitive and social skills appear with the greatest frequency in the course description data. Over time, I document substantial growth in the share of courses intensive in these skills. Over the period 2012-2020, the cognitive and social shares of courses each grew by 11% (2.1 pp. and 1.8 pp., respectively). Figure 6 also documents growth in the share of courses intensive in programming skills, although these courses comprise a smaller share of total courses than courses intensive in cognitive or social skills. The share of courses intensive in reading and writing was largely constant between 2012-20, while the share of courses intensive in basic computer skills declined.

I leverage the diversity of schools in the course catalog data sample to study differences in skill offerings by institution selectivity. I split institutions in the sample into three categories: two-year institutions, less selective four-year institutions, and selective institutions.[4] Figure 7 plots separate skill content trends over time for the three institution categories. The figures are largely consistent with a story of greater exposure to high-earnings skills at more selective institutions. In particular, while the share of courses intensive in cognitive and social skills has grown overall, the growth has been driven primarily by course offerings at four-year institutions, and exposure to these skills is greatest at the most selective institutions. Disaggregating the trend by institution selectivity suggests potential disparities in exposure to in-demand skills which may contribute to earnings disparities among graduates of these

[3] Note that the keywords used to detect skill intensity in the course descriptions are drawn from general skills used in a variety of occupations. This keyword restriction reduces the share of, for example, classes in Computer-Related Fields that are deemed intensive in programming skills, as many advanced Computer Science classes in the data teach specialized skills beyond fundamental computer programming.

[4] I define a four-year institution as "selective" if its 25th percentile composite ACT score is greater than or equal to 24.

institutions.

Taken together, analysis of the course descriptions provides evidence of the growing prominence of cognitive and social skill development in college courses. More selective institutions seem to emphasize skills linked to high earnings. Differences by institution selectivity arise from a combination of differences in the kinds of courses offered and in the skills emphasized in comparable courses offered by institutions of differing selectivity. Determining the sources of growing skill intensity and more refined text analyses are left for future work.

Conclusion

Recent waves of technological change have advantaged worker with skills complements to technology. Among the skills most important for high-skilled workers are cognitive and social skills. In this paper, I show that cognitive and social skills are increasingly emphasized in college courses. Using course catalog data, I study the skills students develop in college and track variation across institutions, majors, and time. This paper introduces new stylized facts about course offerings in a large sample of colleges and universities across the United States. Course offerings have expanded over the last decade, with substantial growth in STEM and computer science courses. The cognitive skills of increasing importance in the labor market are emphasized increasingly in college courses, particularly at selective four-year institutions.

This paper leaves questions of the incentives institutions face to update course offerings and estimating elasticities of course content to labor market changes for future research. The course supply side of higher education is an important, yet under-studied, component of a student's human capital investment. The course catalog dataset used for this paper explore new channels through which institutions and majors differ in preparing students for the labor force.

The analysis in this paper highlights margins of heterogeneity in the skills a college graduate develops. I show that college degrees are not a static object. As course offerings expand and skill priorities evolve, the skills college graduates develop change over time. This time variation has important policy implications, as it suggests that different cohorts of college graduates may be differentially exposed to technology or economic shocks. A rich literature estimates returns to majors or institutions, but data limitations often preclude analysis of the interaction between institutions and majors. With the major-to-skills mapping I develop in this paper, I introduce a common unit to compare student experiences at the institution-department level.

Recent research predicts that the next round of innovation, the widespread application of artificial intelligence, may pose a greater threat to higher-skilled workers (Webb (2019); Brynjolfsson et al. (2018)). Future research and policy work will benefit from the ability to connect workers at-risk workers to the skills they developed in college.

References

David H Autor and David Dorn. The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97, 2013.

David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333, 2003.

Miroslav Beblavỳ, Brian Fabo, and Karolien Lenaerts. Demand for digital skills in the us labour market: the it skills pyramid. 2016.

Erica Blom, Brian C Cadena, and Benjamin J Keys. Investment over the business cycle: Insights from college major choice. 2015.

John Bound, Michael F Lovenheim, and Sarah Turner. Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics*, 2(3):129–57, 2010.

Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108, pages 43–47, 2018.

David Deming and Lisa B Kahn. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1):S337–S369, 2018.

Claudia Goldin and Lawrence F Katz. The race between education and technology: The evolution of us educational wage differentials, 1890 to 2005. Technical report, National Bureau of Economic Research, 2007.

Brad Hershbein and Lisa B Kahn. Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–72, 2018.

Mathew Joseph, Eileen Wall Mullen, and Deborah Spake. University branding: Understanding students' choice of an educational institution. *Journal of Brand Management*, 20(1): 1–12, 2012.

Bridget Terry Long. How have college decisions changed over time? an application of the conditional logistic choice model. *Journal of econometrics*, 121(1-2):271–296, 2004.

NCES. Integrated postsecondary education data system (ipeds), 1995-2018. Accessed: 2020-09-01.

NCES. Digest of education statistics. Report, U.S. Department of Education, National Center for Education Statistics, 2019.

Michael Webb. The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*, 2019.

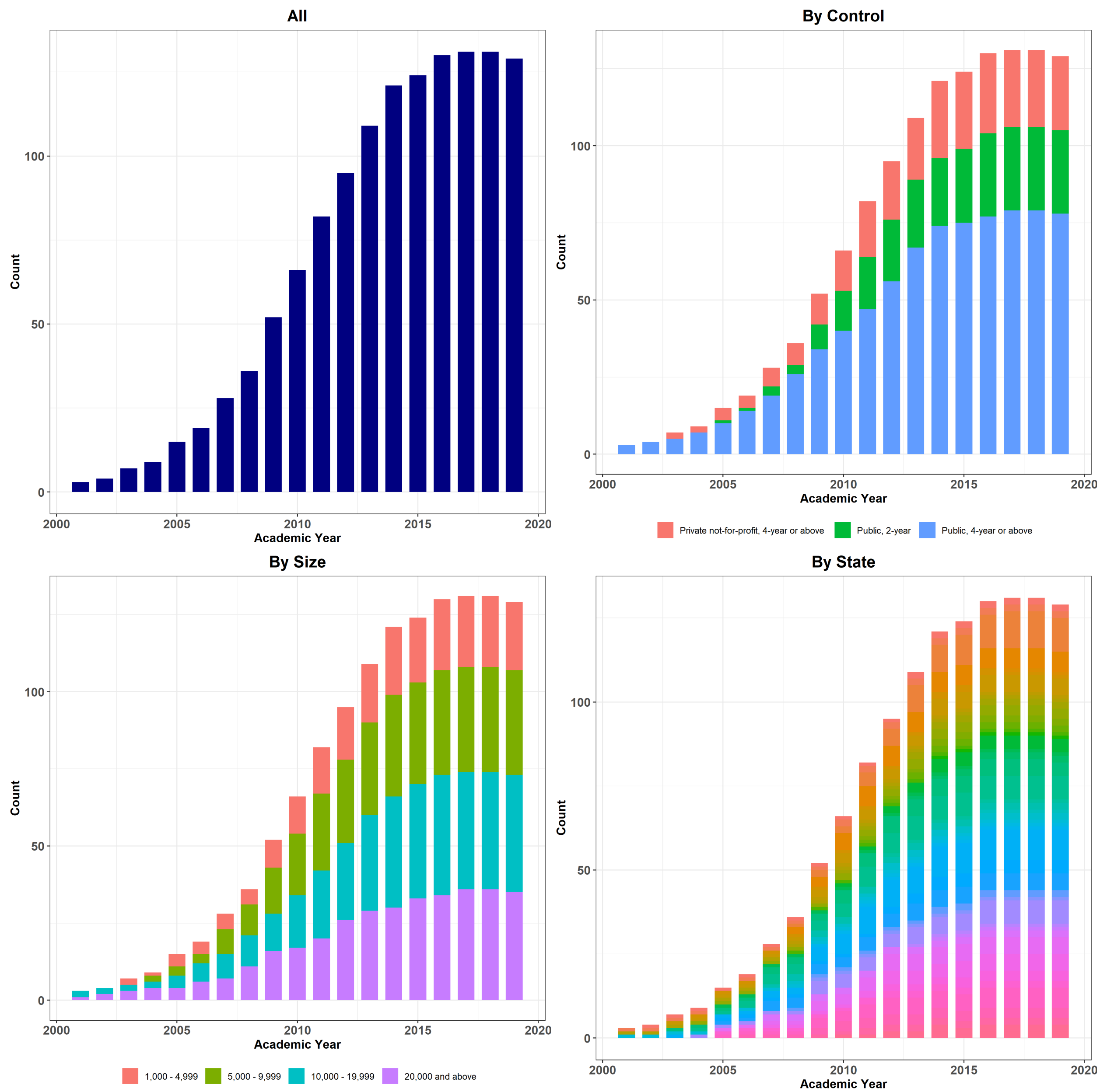
Tables and Figures

Figure 1: Example course catalog entry

```
$title
[1] "ECON 420 - Topics in Labor Economics"

$description
[1] "In this course we will study some of the most relevant aspects of modern labor markets, both from a t
heoretical and empirical point of view. We will cover topics such as the determinants of the decision to w
ork, the effects of education on labor market earnings, the effects of family background on labor market o
utcomes, discrimination in the labor maker, and the gender wage gap, among other up-to date topics. Studen
ts will learn how to use economic models to frame the relevant questions and obtain theoretical prediction
s and to analyze real life data to answer those questions and interpret the results. This course will requ
ire the use of software for statistical analysis such as Excel or Stata.Laptops are not required in class"

$credits
[1] "3"

$enforced_prerequisites
[1] "ECON 401 with a C- or better."

$primary_instructor
[1] "Reynoso,Ana"

$term
[1] "Fall 2018"

$department
[1] "LSA: Economics"

$institution
[1] "University of Michigan"
```

Source: University of Michigan, Fall 2018 Course Catalog.

Figure 2: Summary of Course Catalog Sample



Sources: Course catalog dataset, IPEDS.

Notes: Histograms count institutions in course catalog dataset by year. Histograms classify institution by (from top right, moving clockwise) institution control, state (where each shade represents a different state), and institution size category. Institutions with catalogs that span multiple years are counted in each year covered by the catalog.

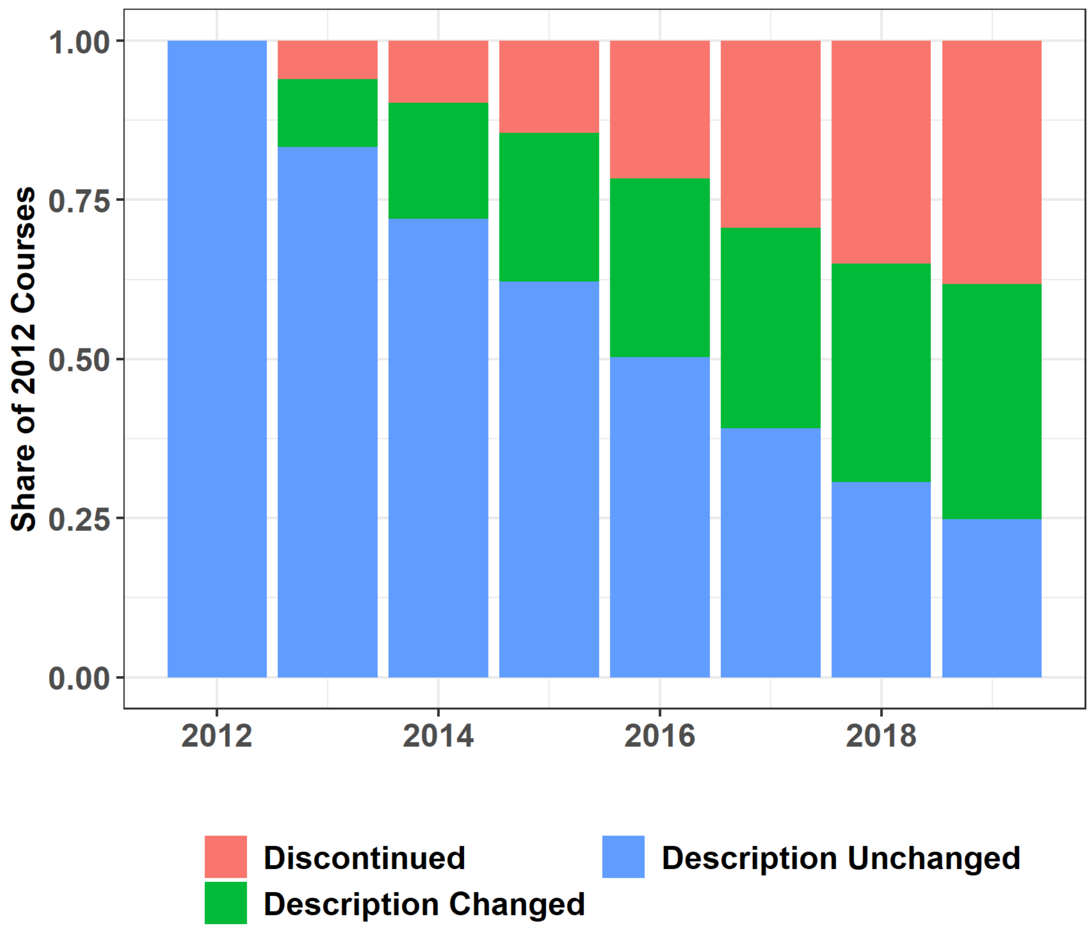
Table 1: Compare Course Catalog Sample to Population of Colleges and Universities (2018-19)

	Sample	Population
Admit Rate	65	67
	(22.1)	(19.7)
Enrollment	12,412	5,989
	(10,293.6)	(7,906.8)
Tuition	13,835	15,201
	(13,794.4)	(14,263.2)
Test Scores		
SAT Writing - 25th Percentile	532	512
	(68.8)	(69.1)
SAT Writing - 75th Percentile	630	610
	(59.9)	(63.2)
SAT Math - 25th Percentile	528	507
	(65.8)	(69.6)
SAT Math - 75th Percentile	629	606
	(66.0)	(69.9)
Count	133	2,637

Sources: Course catalog dataset, IPEDS 2018.

Notes: Standard deviations in parentheses. Population summarizes characteristics for non-profit degree-granting Title IV eligible colleges and universities in the United States as of the 2018-19 academic year.

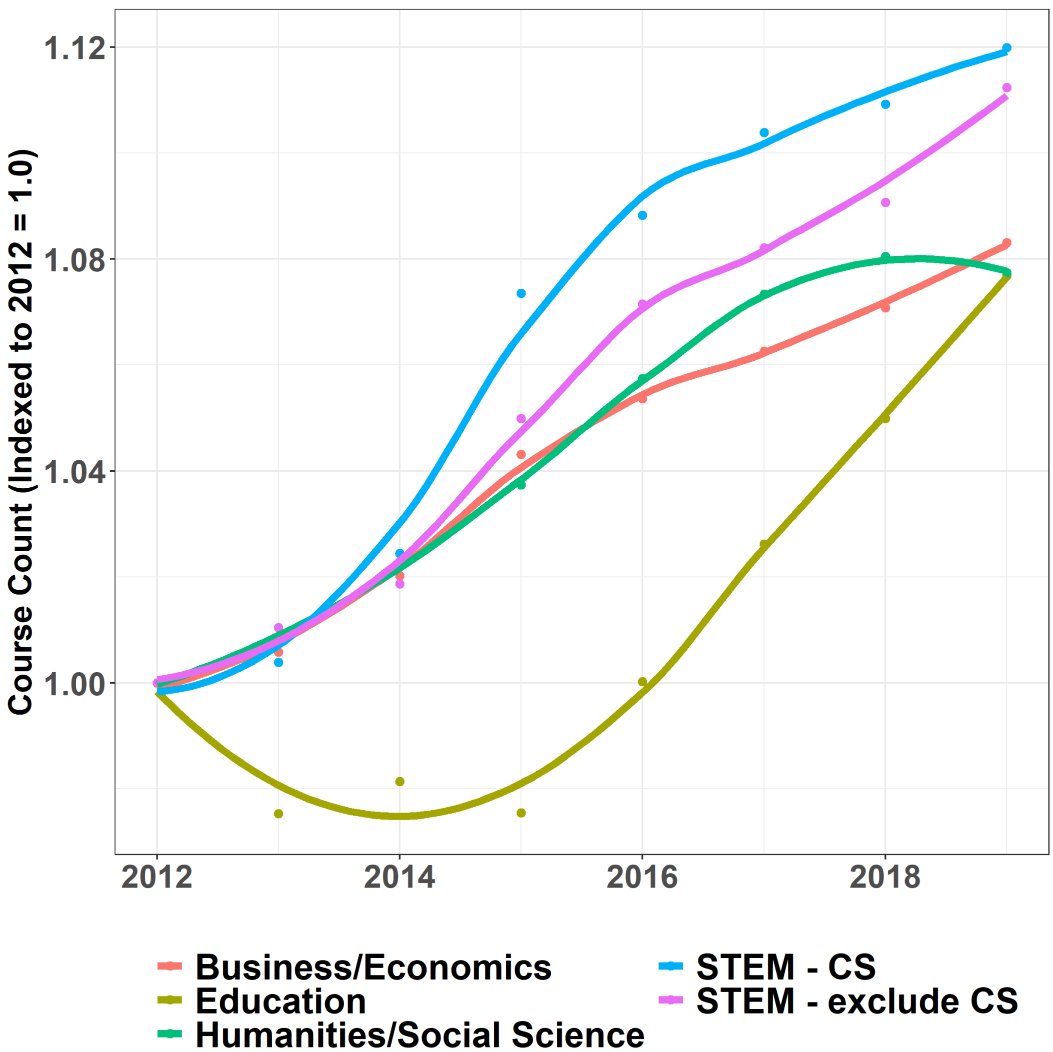
Figure 3: Course description survival, 2012 - 2020



Source: Course catalog dataset, IPEDS 2018.

Notes: Figure restricts to panel of institutions in course catalog dataset each year 2012-13 to 2019-20. Figure restricts to courses offered in the 2012-13 academic year and tracks the survival rate of the course description in subsequent years. A course is defined as a unique department-course number-institution triple. A course changes description when the same course is listed with a description that does not match the 2012-13 description. A course is dropped when it is not offered in any future years covered by the catalog dataset. Courses can transition from changed description to dropped. Annual survival rates are calculated for each institution, then averaged with weights proportional to undergraduate enrollment in 2018-19 from IPEDS.

Figure 4: Growth in course offerings, 2012 - 2020



Source: Course catalog dataset, IPEDS 2018.

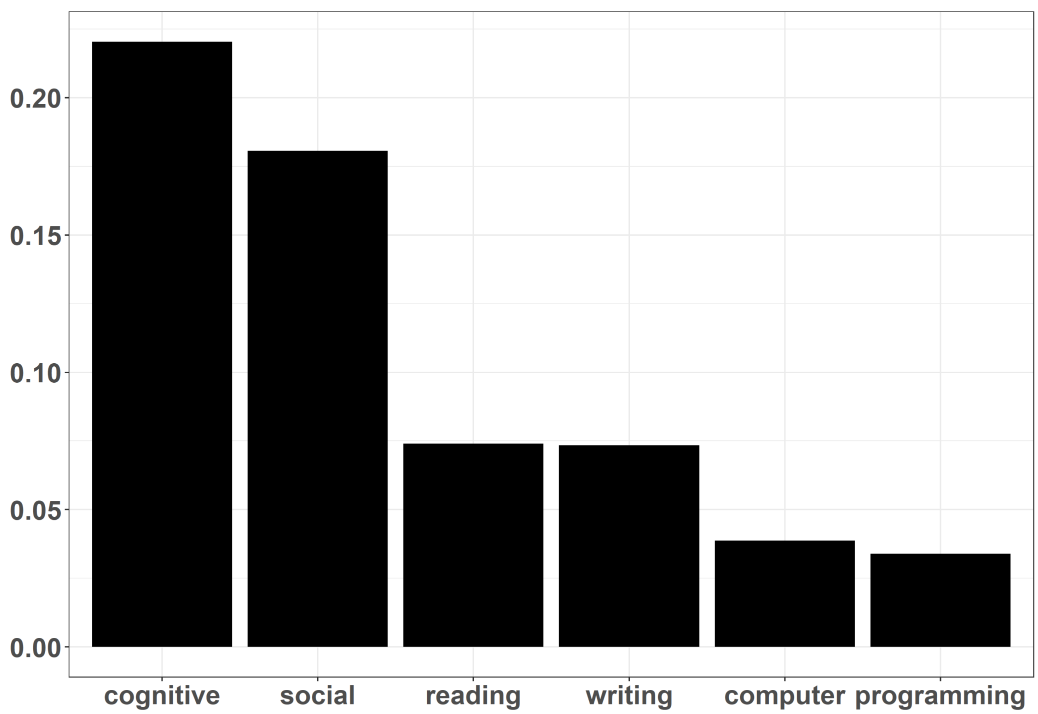
Notes: Figure restricts to panel of institutions in course catalog dataset each year 2012-13 to 2019-20. Figure counts unique courses courses offered during each academic year. A course is defined as a unique department-course number-institution triple, regardless of how frequently the course is offered within the year. Course in each department group are counted at the institution level, then averaged across institutions with weights proportional to undergraduate enrollment in 2018-19 from IPEDS. Departments are mapped to four-digit CIP codes, which are then classified according to the five department categories in the figure. The mapping will be available as an Appendix table.

Table 2: Keywords for high earnings skills

Skill	Keyword
Computer	Computer, Microsoft, Excel, Powerpoint
Programming	Machine Learning, Artificial Intelligence, Coding, Programming, Spreadsheet, C++, Java, Python, C, R, Stata, Fortran, VBA, html, Ruby, Julia
Writing	Write, Writing
Reading	Read, Reading
Cognitive	Problem Solving, Research, Analysis, Analyze, Critical Thinking, Calculus, Algebra, Statistics
Social	Communication, Team, Group, Collaborate, Negotiate, Present

Notes: Adapted from Deming and Kahn (2018).

Figure 5: Average skill exposure across all institutions (% of courses)



Source: Course catalog dataset, IPEDS 2018.

Notes: Figure restricts to course data in 2019-20 academic year. Each course is scored as "intensive" in a skill if its course description contains at least one keyword corresponding to the skill (defined in Table 2). Courses offered multiple times during the year are counted once in the average, with skill intensity score equal to the highest skill intensity score across descriptions. Averages exclude courses with missing or empty course description. Skill intensity scores are averaged across all courses offered within each institution, then averaged across institutions with weights proportional to undergraduate enrollment in 2018-19 from IPEDS.

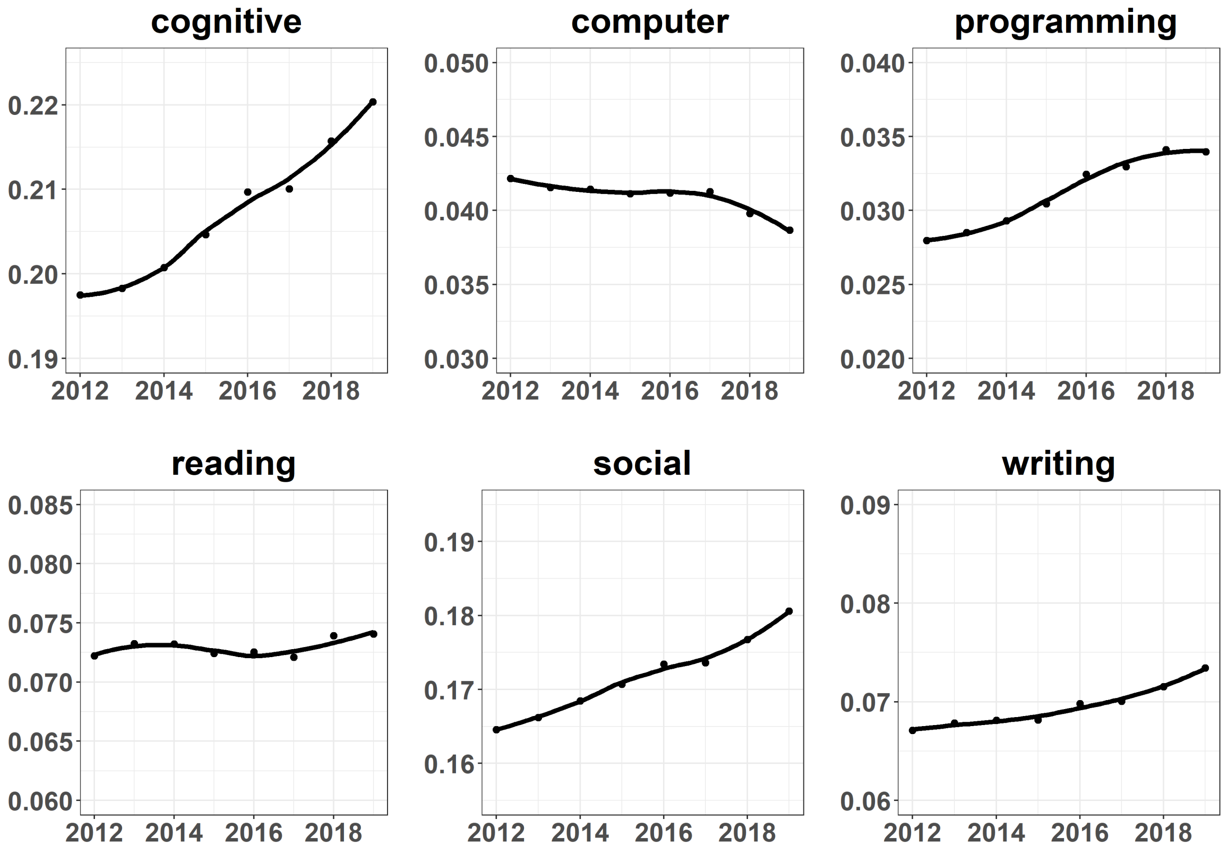
Table 3: Extremes in Skill Intensity (% of courses)

	Cognitive		Computer	
Top	Mathematics and Statistics	0.56	Computer-Related Fields	0.30
Bottom	Leisure Studies	0.10	Liberal Arts and History Fields	0.01
	Programming		**Reading**	
Top	Computer-Related Fields	0.30	Literature and Languages Fields	0.30
Bottom	Pharmacy	0.00	Public Affairs, Health, Policy	0.01
	Social		**Writing**	
Top	Communications Fields	0.49	Journalism	0.42
Bottom	Finance	0.06	Pharmacy	0.00

Source: Course catalog dataset, IPEDS 2018.

Notes: Table restricts to course data in 2019-20 academic year. Each course is scored as "intensive" in a skill if its course description contains at least one keyword corresponding to the skill (defined in Table 2). Courses offered multiple times during the year are counted once in the average, with skill intensity score equal to the highest skill intensity score across descriptions. Averages exclude courses with missing or empty course description. Departments classified according to Blom et al. (2015) classification of four-digit CIPs. Skill intensity scores are averaged by department group across all courses offered within each institution, then averaged across institutions with weights proportional to undergraduate enrollment in 2018-19 from IPEDS.

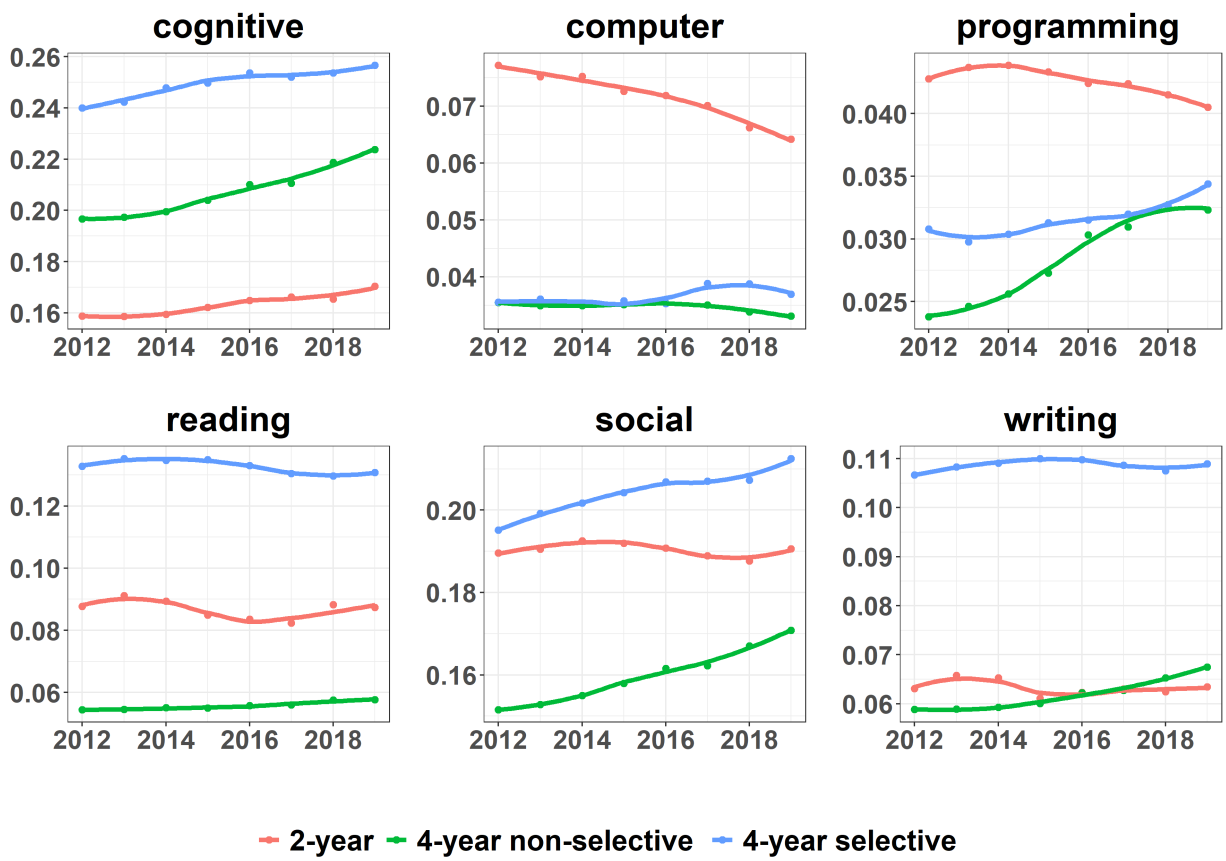
Figure 6: Growth in skill offerings over time (% of courses)



Source: Course catalog dataset, IPEDS 2018.

Notes: Figure plots the share of courses intensive in each of the high earnings skill categories. Figure restricts to panel of institutions in course catalog dataset each year 2012-13 to 2019-20. Each course is scored as "intensive" in a skill if its course description contains at least one keyword corresponding to the skill (defined in Table 2). Courses offered multiple times during the year are counted once in the average, with skill intensity score equal to the highest skill intensity score across descriptions. Averages exclude courses with missing or empty course description.

Figure 7: Growth in skill offerings over time by institution selectivity (% of courses)



Source: Course catalog dataset, IPEDS 2018.

Notes: Figure plots the share of courses intensive in each of the high earnings skill categories by institution selectivity over time. Figure restricts to panel of institutions in course catalog dataset each year 2012-13 to 2019-20. Institutions are classified by control. Four-year institutions are "selective" if the 25th percentile ACT composite score is greater than or equal to 24, and "non-selective" otherwise. Each course is scored as "intensive" in a skill if its course description contains at least one keyword corresponding to the skill (defined in Table 2). Courses offered multiple times during the year are counted once in the average, with skill intensity score equal to the highest skill intensity score across descriptions. Averages exclude courses with missing or empty course description.