Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1042

Classifying Courses at Scale: a Text as Data Approach to Characterizing Student Course-Taking Trends with Administrative Transcripts

Annaliese Paulson
University of Michigan

Kevin Stange
University of Michigan

Allyson Flaster
University of Michigan

Students' postsecondary course-taking is of interest to researchers, yet has been difficult to study at large scale because administrative transcript data are rarely standardized across institutions or state systems. This paper uses machine learning and natural language processing to standardize college transcripts at scale. We demonstrate the approach's utility by showing how the disciplinary orientation of students' courses and majors align and diverge at 18 diverse four-year institutions in the College and Beyond II dataset. Our findings complicate narratives that student participation in the liberal arts is in great decline. Both professional and liberal arts majors enroll in a large amount of liberal arts coursework, and in three of the four core liberal arts disciplines, the share of course-taking in those fields is meaningfully higher than the share of majors in those fields. To advance the study of student postsecondary pathways, we release the classification models for public use.

VERSION: September 2024

Suggested citation: Paulson, Annaliese, Kevin Stange, and Allyson Flaster. (2024). Classifying Courses at Scale: a Text as Data Approach to Characterizing Student Course-Taking Trends with Administrative Transcripts. (EdWorkingPaper: 24-1042). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7fpa-s433

Classifying Courses at Scale: a Text as Data Approach to Characterizing Student Course-Taking Trends with Administrative Transcripts[1]

Annaliese Paulson
University of Michigan
annamp@umich.edu

Kevin Stange
University of Michigan
kstange@umich.edu

Allyson Flaster
University of Michigan
aflaster@umich.edu

Abstract

Students' postsecondary course-taking is of interest to researchers, yet has been difficult to study at large scale because administrative transcript data are rarely standardized across institutions or state systems. This paper uses machine learning and natural language processing to standardize college transcripts at scale. We demonstrate the approach's utility by showing how the disciplinary orientation of students' courses and majors align and diverge at 18 diverse four-year institutions in the College and Beyond II dataset. Our findings complicate narratives that student participation in the liberal arts is in great decline. Both professional and liberal arts majors enroll in a large amount of liberal arts coursework, and in three of the four core liberal arts disciplines, the share of course-taking in those fields is meaningfully higher than the share of majors in those fields. To advance the study of student postsecondary pathways, we release the classification models for public use.

Keywords: Liberal Arts Education, Postsecondary Curriculum, Natural Language Processing, Administrative Data

[1] Interested researchers can download the classification models developed in this paper to place their own course-level transcript data into the CCM schema and access a demonstration notebook at https://huggingface.co/collections/annamp/classifying-courses-at-scale-66e470e6126f9577f89fd417. Annaliese Paulson is grateful for support from PR/Award R305B200011 from the Institute of Education Sciences, U.S. Department of Education and a 2024 Spencer/National Academy of Education Dissertation Fellowship. The findings and conclusions contained within are those of the authors and do not necessarily reflect the positions or policies of the Institute of Education Sciences or the U.S. Department of Education. We are grateful for financial support for this project from a Propelling Original Data Science Grant from the Michigan Institute of Data Science at the University of Michigan. This manuscript was greatly improved through discussions with seminar participants from the Pathways Network at Stanford University.

I. Introduction

The courses students take while in college are fundamental to the role postsecondary education plays in shaping their lives, as conceptualized by many social science disciplines. In economics, human capital theory understands the returns to college as stemming from investment in productive capacity largely developed from classroom experiences (Becker, 1993; Arteaga, 2018); psychologists have connected the experience of specific kinds of coursework like diversity and writing courses with cognitive and affective development (Denson & Bowman, 2017; Bangert-Downs et al., 2004); education scholars have examined the particular role courses play in successful transfer from community colleges to four-year institutions (Wang, 2016); and sociologists have theorized about the role "weed-out" courses - large introductory courses in STEM - play in racial and class stratification (Weston et al., 2019). Understanding course-taking is essential to investigating the theoretical models undergirding the study of education and its consequences from many disciplinary perspectives.

At a policy level, recent administrative decisions at public institutions such as West Virginia University and the University of Vermont to cut departments in foreign languages, classics, and other liberal arts fields have also brought focus to what students in the United States are - or should be - learning while in college. Declining interest in majoring in the liberal arts and humanities has been a constant topic of public discussion, with headlines like "the suicide of the liberal arts" (Riley, 2022) and "the liberal arts may not survive the 21st century" (Harris, 2018), contrasted with the refrain that "college should be more than just vocational schools" (Devereaux, 2023). In these pieces and others, advocates argue that the liberal arts are key to developing the competencies students need in the information age, such as the ability to analyze

evidence, communicate effectively, root out misinformation, and engage in effective citizenship (Moen, 2020; Pasquerella, 2020). In contrast, some applaud the trend of fewer students majoring in the liberal arts under the belief that studying literature, philosophy, and history is a frivolous distraction from the learning that should be going on in college-- learning that is tightly connected to the needs of the labor market (Cohen, 2016). Others believe the liberal arts have real value, but are less concerned about the reduction in liberal arts programs per se, noting that students will still be able to learn these topics while majoring in other subjects, and due to the infusion of the liberal arts into professional field curricula (Heller, 2023).[2] Facts about the courses students are actually taking in college are notably absent from much of this discussion.

Clearly, from both a theoretical and policy perspective, being able to measure and examine students' course-taking -- to know the specific content they engage with and learn --is critically important for the advancement of education research and practice. Luckily, large-scale administrative data warehouses have made access to rich data containing students' course-taking histories much more available over the past two decades. However, these data are designed for administrative rather than research purposes so researchers still need to derive variables that measure the relevant theoretical constructs using often unstandardized or unstructured data (Salganik, 2018). In the case of postsecondary transcripts, for instance, a researcher may want to examine how enrollment in humanities-related courses changed after the Great Recession at regional colleges and universities across the U.S., but cannot do so without a way of easily

[2] In his essay for the New Yorker, Heller eloquently summarizes this view: "One idea about the national enrollment problem is that it's actually a counting problem: students haven't so much left the building as come in through another door. Adjacent fields aren't included in humanities tallies, and some of them are booming. Harvard's history-of-science department has seen a fifty-per-cent increase in its majors in the past five years. The humanities creature who recites Cavafy at parties might fade away, but students are still getting their vitamins. There's a lot of ethics in bioethics, after all" (2023, p. 7)

determining whether thousands of courses across these institutions – each of which have their own systems of course numbering and labeling – contain comparable and relevant content. Thus, even though the big data revolution has expanded access to large-scale administrative data, researchers lack the ability to use it to its fullest potential when examining student course-taking. This is why almost all empirical research that examines students' engagement with different subject matter and disciplinary ways of thinking is measured at the major-level, rather than course-level: standardized information about course-taking is seldom available in administrative transcripts.

In this paper, we make two major contributions to research on college student course-taking. First, we provide a rich description of trends in course-taking in liberal arts and professional fields by undergraduates at a diverse set of 18 colleges over ten cohorts of graduates. We place these trends in contrast to trends in college major, which, as noted above, have received more empirical attention and largely drive the 'decline of the liberal arts' narrative. We find that longitudinal trends in majors provide a somewhat misleading view of students' course-taking experiences. While course-taking in the liberal arts overall and humanities in particular has meaningfully declined over the past two decades, the liberal arts disciplines continue to comprise the clear majority of course-taking and the humanities, in particular, are the largest discipline as measured by course-taking. Thus, measuring students' exposure to disciplinary content based solely on the number of students who earn degrees in that field will substantively understate students' exposure to the liberal arts.

We are able to provide a rich description of course-taking trends across institutions because of our second contribution: we train machine learning models that use natural language processing techniques to classify courses from student transcripts into a standardized hierarchical

taxonomy of postsecondary courses, the College Course Map (CCM). Developed by the National Center for Education Statistics (NCES), the CCM is very similar in structure to the more widely-used Classification of Instructional Programs (CIP) taxonomy of college programs and majors. However, the CCM has - thus far - been used infrequently due to the difficulty in applying it to student transcripts. The NCES Postsecondary Education Transcript Studies (PETS) contain CCM codes, obtained through a labor- and time-intensive human annotation process, but NCES longitudinal postsecondary studies are infrequent (available every 4 or 8 years) and of relatively small scale compared to administrative data. Given ongoing calls to identify the many varied and complex curricular pathways students use to navigate the postsecondary curriculum (Kizilcec et al., 2023), researchers require tools that efficiently standardize student transcripts at scale across institutions and postsecondary institutions.

We use four waves of NCES PETS studies to train our tool and then apply it to institutional transcript data from 18 colleges and universities. Our algorithms predict CCM codes with a high degree of accuracy using course titles and subject codes. Since our training data is representative of postsecondary institutions nationally, the model is broadly applicable to many settings. We have publicly released the model weights and code to allow individual researchers, systems, and postsecondary institutions to standardize their postsecondary transcript data into the CCM schema without requiring substantial human annotation labor or financial resources.

The rest of the paper provides detailed information about our model training and demonstrates its application to an important research problem in higher education. In section II we provide background information on student majors and course-taking and review the limited prior work on the subject. In sections III, IV, and V we describe the College Course Map, our training data and feature engineering, the training of the model, and its evaluation. In VI, we

illustrate the value of our tool by using the classifications produced by our model to present new evidence on course-taking trends over time, with a focus on the distinction between professional and liberal arts fields and variances between course enrollment and majors. In VII, we discuss potential improvements that could be made to the model. In VIII, the paper concludes by summarizing potential next steps in the development of classification models like ours and new insights from our exploratory analyses. We also offer thoughts on the potential for using labeled transcript data to better understand myriad aspects of postsecondary education.

II. Trends in Studying Liberal Arts and Professional Fields

The US higher education system has historically placed greater emphasis on the liberal arts disciplines - the social and behavioral sciences, humanities, natural sciences and mathematics, and fine and performing arts - and less emphasis on occupational and professional disciplines, including engineering, education, and business, than in other countries (Andersen, 2020). The result has been a substantial number of college students choosing to major in a liberal arts field. However, the dominance of the core liberal arts disciplines has been declining since the 1970's. Between 1960 and 2000, bachelor's degrees awarded in most liberal arts and science fields declined – not just as a proportion of total degrees, but in absolute numbers – despite the number of graduates nearly doubling over that time period (Brint et al., 2005). In part this decline in liberal arts and science majors is because student decisions about field of study are responsive to labor market trends. When faced with poor labor markets, student shift from majors with unclear links to the labor market like liberal arts fields into majors with clear links and stronger pecuniary returns like engineering, nursing, and business (Blom et al., 2021). More recently, between 2010-11 and 2020-21, the total number of bachelor's degrees conferred in the

U.S. increased by 20% overall, with increases in degrees in five of the top six most popular fields: health, engineering, biology, psychology, and business (NCES, 2023).

Regardless of the popularity of specific majors, those who observe these trends inevitably assume that exposure to liberal arts coursework necessarily mirrors the trends in degrees awarded. However, the number of completed majors in a discipline may not be a wholly accurate representation of learning and engagement in that discipline due to the practice of general education and the elective model of course-taking. General education generally aims to produce graduates who have the "broad knowledge of culture, nature, and society that a well-rounded citizen needs, and to provide the skills in reading, writing, and reasoning that are necessary for academic studies" (Andersen, 2020, p.2). The most common general education model is the 'core distribution model' in which students need to take courses distributed across a variety of content areas in the core liberal arts disciplines and related subjects such as foreign language and composition (Brint et al., 2009; Warner & Koeppel, 2009).[3] These general education requirements typically comprise one-third of a student's degree requirements, comparable to the share devoted to requirements within the major (Andersen, 2020; Lattuca & Stark, 2009). Elective requirements constitute the final third of a students' degree, and they provide another avenue for students to engage in content outside of their major field of study, particularly if students pursue a professional major for pecuniary benefits but devote their electives to courses they are interested in for other reasons.

[3] For example, using a fairly restrictive definition of "the humanities" that only includes core humanities, philosophy, literature, and religion courses, research universities require students to take 2.65 humanities courses and liberal arts colleges require 3.67 courses, on average. Similarly, research universities require, on average, 4.13 natural science and mathematics courses, 2.04 social science courses, and .67 fine arts courses while liberal arts universities require 3.66 natural science and mathematics courses, 1.29 social science courses, and .48 fine arts courses (authors' calculations from data in Warner & Koeppel, 2009).

To understand if students are truly engaging with the liberal arts on a more limited scale than in the past, observers would ideally examine trends in course-taking. However, existing research on trends in course-taking largely rely on dated data. Although NCES's PETS provide rich standardized transcript data that can be linked with their surveys, these large scale surveys result in repeated cross-sectional data, with cohorts separated by five or more years. In one of the only large scale analyses using PETs data and the CCM to understand course-taking, Adelman (2004) documents several facts about three cohorts of 12th graders who earned baccalaureate degrees from 1972, 1982, and 1992. First, the thirty largest CCM codes in each cohort comprise approximately one third of all course-taking, suggesting an "empirical core curriculum." Twenty-one of those courses are the same across the three cohorts he examines, further showing that a small number of courses comprise a large portion of enrollments and that portion of enrollments has been stable over decades. The courses that make up the highest percentage of credits earned are English composition, general psychology, biology and chemistry, and calculus. Comparing the coursework of the 1982 and 1992 cohort, Adelman finds increasing enrollments in ethnic and culture studies, introductions to science, women's studies, Spanish language, crime studies, ethics, environmental resources, and computer applications, with major decreases in business fields, computer programming, and remedial English. However, it is difficult to draw policy relevant implications from Adelman's work given the latest cohort of high school graduates he examined is from more than thirty years ago.

Beyond NCES surveys, scholarly associations like the Modern Language Association (MLA) and the American Mathematical Society also do semi-regular surveys of course enrollments in their disciplines, resulting in snapshots of course enrollments in select disciplines. For instance, between 2005 and 2015, undergraduate enrollment in mathematics and statistics

classes rose dramatically, increasing from 1.67 million enrollments in 2005 to 2.21 million in 2015 (Blair et al., 2018); of particular note, the number of undergraduate majors in math and statistics slightly declined between 2005 and 2010, despite the substantial increase in math course-taking over that same period. Data collected by the MLA shows that undergraduate enrollment in world language classes actually increased from 1995 to 2006 - both in absolute terms and as a share of enrollment - but fell by 10 percent between 2006 and 2016 (NCES, 2019). Although such snapshots of disciplinary course-taking suggest that engagement in liberal arts subjects might not be declining at all, or at least not at the same rate as it is in liberal arts majors, it is hard to draw broad conclusions about cross-disciplinary trends from this data.

Much of the recent work on student progress and pathways through college uses institutional or state administrative data, rather than data from national cohort studies provided by NCES.[4] The expansion of administrative databases means that researchers increasingly have access to rich student transcripts that facilitate asking and answering questions about course-taking behavior. However, because student transcripts are designed for administrative purposes rather than research, they typically are messy and lack standardization across schools. While researchers may wish to measure the amount of coursework by discipline or investigate trends in specific courses like calculus, the variables provided in administrative transcripts are typically limited to subject codes, catalog numbers, and short - often abbreviated - course titles. When attempting to analyze transcripts in this way researchers are confronted with a set of less than ideal choices - they can focus on just one institution where it is feasible to draw on institutional

[4] See Kizilcec et al. (2023) for a description of recent innovative projects that measure student academic progress through college, all using institutional administrative data. McFarland et al. (2021) offer a broader overview of the emerging field of Education Data Science that has developed to leverage new novel forms of data including administrative transcripts and text data.

knowledge about the content of individual courses; they can perform labor intensive human annotation of each course; or they can associate whole department codes with CIP codes, sacrificing specificity for scale. To enable the study of course-taking across institutions at scale, researchers need easily applicable tools that map the messy data present in administrative datasets - often unstructured text - into a standardized taxonomy, such as the College Course Map.

III. College Course Map

The College Course Map (CCM) is a product of the US Department of Education's National Center for Education Statistics (NCES) that is featured prominently in their longitudinal surveys of postsecondary students. The CCM was originally developed in 1990, based on extensive conversations with experts drawn from disciplines across the postsecondary curriculum and was substantially revised in 2010 (Adelman, 1990; 2004; Bryan & Simone, 2012). Transcripts are collected by NCES for the students included in their longitudinal surveys. But, because raw administrative transcripts are unwieldy to work with, these courses are then annotated and standardized by human coders using the CCM typology (Bryan & Simone, 2012). The CCM maps each course into a six-digit code where the first two digits define a course's broad cluster (e.g. 45 - social sciences), the third and fourth digits define a course's subcategory (e.g. 45.06 - economics), and the fifth and sixth digits define a course's specific subject code (e.g. 45.0603 econometrics and quantitative economics). Across the NCES datasets we work with, there are 48 unique two-digit codes, 403 unique four-digit codes, and 2,041 unique six-digit codes. At the four-digit level, the CCM scheme is quite similar to the widely used Classification of Instructional Program (CIP) taxonomy used to characterize college majors and programs, although the six-digit level denotes specific courses rather than majors or programs. The CCM

does not explicitly embed the level of courses (e.g. introductory, lower-division, upper-division) in the numbering system, though some specific introductory courses are separately identified (for instance, introductory courses to a discipline are typically coded under a general category such as "Writing, General").

In the NCES postsecondary transcript data files, courses were annotated by human coders on the basis of course titles, subject codes, and catalog numbers, and - if necessary - course descriptions from postsecondary institutions' course catalogs (Bryan & Simone, 2012). If these pieces of information were insufficient for classification, then annotators relied on the additional context provided by the other courses on students' transcripts. However, in the data available to researchers, course descriptions drawn from course catalogs are not distributed.[5]

In Table 1, we illustrate the importance of using CCM codes to standardize transcript data across institutions. The table shows several examples of subject codes and course titles from public institutions in Texas in the format of the restricted PETS data. Note that-- for courses covering the same content-- there are a variety of course numbering systems and ways of titling courses across transcripts in student administrative data. BIO 205L and BIOL 3415 are a good example of the diverse ways the same type of course can be recorded. Without standardization, it would be difficult to determine the topical overlap of these two classes. However, if we assign CCM codes, we can see that BIO 250L and BIOL 3415 both cover molecular biology. Thus, we

[5] Standard measures of intercoder reliability such as Krippendorff's alpha were not distributed with the College Course Map data in PETS, making it difficult to evaluate the reliability of the human annotation process. However, examining the course subject codes and course titles in the PETs suggests at least minor inconsistencies across annotations. For instance, two courses offered at the same institution in different terms with identical titles and near identical course descriptions in the institution's course catalog, but a slight variation in subject number, are annotated differently: when the subject code is separated from the catalog number by a space character, the course is annotated as 11.0699 while, when the subject code is separated from catalog number with a "/" character, the course is annotated as 11.0501.

can reasonably assume that any students who have BIO 250L and BIOL 3415 on their transcripts have encountered similar course content, despite being at different institutions. It is this standardization that allows for inter-institutional data analysis, thereby fulfilling the potential of administrative data to offer population-level longitudinal data, statistical power, and the ability to detect rare events and heterogeneous effects (Figlio et al., 2017).

IV. Training Data, Preprocessing, and Feature Transformation

Our approach to standardizing courses across institutions uses supervised machine learning to classify courses in student transcript data into the codes of the 2010 CCM. In a supervised machine learning approach, we train a model to predict the most probable label conditional on a set of text features in training data that has been annotated by people (Fesler et al., 2019; Grimmer et al., 2022).

Training Data

We first develop a comprehensive corpus of labeled training data that contains both a human-annotated CCM code label for each course record and the features we will use to predict that label. Human annotated records are the gold-standard when using supervised machine learning in the social sciences (Grimmer er al., 2022). To create our full corpus, we draw on restricted data from NCES's Postsecondary Education Transcript Studies (PETS) associated with each of four recent NCES surveys: High School Longitudinal Study of 2009 (Duprey et al., 2020), Baccalaureate and Beyond Longitudinal Study of 2008-2012 (Wine et al., 2013), Beginning Postsecondary Students Longitudinal Study of 2004-2009 (Wine et al., 2011), and Beginning Postsecondary Students Longitudinal Study of 2012-2017 (Bryan et al., 2020). The High School Longitudinal Study of 2009 is a representative sample of US ninth-graders in 2009;

college transcripts were collected approximately four years after high school graduation for all those who enrolled in college. Baccalaureate and Beyond Longitudinal Study of 2008-2012 collected postsecondary transcripts for a representative sample of students graduating with a Bachelor's degree in 2008. Finally, the Beginning Postsecondary Students Longitudinal Study of 2004-2009 (2012-2017) collected transcripts for a representative sample of students who enrolled in college for the first time in 2004 (2012), with transcripts collected five years later. The latter three datasets use a two-stage sampling design (institutions, then students) and collect multiple students at each postsecondary institution. We refer users to NCES for documentation on the sampling design of these surveys.

To create our training data, we append the four PETS files together, creating a full dataset of 2,660,150 course enrollments from 71,900 students at 4,995 institutions across all four PETS surveys. In particular, the four PETS datasets we draw on contain a variable with subject and catalog number, a variable with course title for all courses that sample students enroll in throughout their postsecondary career, and the human annotated CCM label for the course.

In the following table and figures, we provide broad descriptions of the prevalence of CCM codes in the appended dataset. We highlight, in particular, that a large proportion of course enrollments occur in a small number of CCM codes with a large number of unique records for these CCM codes. Because the quality of supervised machine learning models typically increases with the number of training examples, our model will likely have strong performance on those CCM codes that comprise a large proportion of course enrollments and may be of greater interest to social scientists.

Table 2 displays the 30 CCM codes with the largest course enrollments and the proportion of total enrollments in our datasets they comprise. These 30 CCM codes comprise 32

percent of all course enrollments. This closely parallels Adelman (2004)'s analysis of postsecondary course enrollments using the previous iteration of the 2000 College Course Map in which he also found that the top 30 CCM codes by enrollments comprised approximately one third of all course enrollments. The five most frequent courses include writing, general psychology, chemistry, biology, and American history. Again, this is a similar list to Adelman's (2004) list of top 5 enrolled courses (English composition, general psychology, biology and chemistry, and calculus) — suggesting a high degree of consistency in the most popular courses across forty decades which are, notably, all courses related to liberal arts disciplines.

Figure 1 shows the cumulative proportion of all course enrollments by the relative rank of a CCM code's enrollments. We see that 80 CCM codes comprise half of all course enrollments in our data (at the red dashed line), 343 CCM codes comprise 80 percent of all enrollments (at the green dashed line), and 594 CCM codes comprise 90 percent of all enrollments (at the blue dashed line). Thus, accurately classifying approximately 600 CCM codes enables us to characterize 90 percent of courses taken by US undergraduates in our datasets. Finally, one hundred and seventy-two CCM codes enroll fewer than ten students across the four waves of PETS in our data.

We remove any courses that do not have a CCM code in the PETS datasets, reducing our dataset to 2,624,030 course enrollments. Because we are interested in predicting the CCM code of a course, given a subject number and course title, we then keep only unique subject number, course title, and CCM code triplets, resulting in a dataset of 895,490 unique subject number - course title - CCM code triplets. Finally, given that supervised machine learning models may struggle to classify records that have a limited number of training examples and that codes with a limited number of examples are of less interest since they comprise a small share of all course

enrollments, we remove a small number of CCM codes that occur fewer than 50 times. This further removes 4,170 subject number - course title - CCM code triplets comprising 6,420 course enrollments. This final step results in a final dataset of 891,320 unique subject numbers - course title - CCM code triplets that represents 2,617,550 course enrollments. In some cases, this results in a single triplet representing courses from many institutions. However, the majority of subject numbers- course title - CCM code triplets are unique across institutions: 797,560 course triplets or 89.48% of all course triplets are from one institution.

Our final dataset represents 98.4% of course enrollments in the four waves of PETS and 99.8% of course enrollments in the four waves of PETS with CCM code labels. As a result, our final dataset is largely representative of transcripts drawn from four nationally representative samples of college students: two nationally representative cohorts of beginning postsecondary students; one of graduating baccalaureate students; and one of ninth grade high school students who enroll in college. In our final dataset we retain 48 unique two-digit CCM codes, 373 unique four-digit CCM codes, and 1,622 unique six-digit CCM codes.

Preprocessing

After creating our full corpus of training data, we transform the data from its raw text form into a machine computable format by mapping the text in our corpus to a numeric representation of that text. Specifically, we transform two pieces of text available in the PETS data, course titles and subject numbers, into numeric features that we use to train our model. For example, in the PETS data we could observe a course record with subject number “BIO 205L” and course title “LAB EXPRMNT BIO-CELL & MOL BIO.” We work with course titles as they are provided in PETS. However, because of our modeling strategy, the model is unlikely to receive reliable information from the course number present in the subject number variable. For

example, consider the information present in BIO 205L. While the token “BIO" likely has a great deal of information that is useful for classification because it suggests the course is related to biological sciences, it is unlikely that the token “205L” would have useful information because institutions vary widely in how they number courses and there is no intrinsic meaning to the majority of course numbers. This course record could just as easily have been numbers 204L or 2504 at another institution. In addition, the token 205L could occur in both biology courses like “BIO 205L” and chemistry courses like “CHEM 205L” but there is a limited conceptual link between the types of courses.[6] Given this, we remove all numeric characters from the subject number strings (e.g. transforming BIOL 205L to BIOL L). For both course titles and subject numbers, we then lowercase all strings, and split sentences into individual tokens on word boundaries.

Feature Transformation

To create training and test datasets, we randomly sample 90 percent of our corpus for training and 10 percent of our corpus for testing, stratifying on six-digit CCM code to ensure representation of all codes across both the training and test datasets. This creates a training corpus with 802,190 courses and a test corpus with 89,130 courses.

Because it is difficult for a logistic regression model to use rare tokens effectively, it is standard practice to remove infrequent tokens (Grimmer et al., 2022). We remove any tokens that occur fewer than 50 times across the training corpus. To convert the remaining strings to a machine computable format, we then create two bag of words matrices, one for the course

[6] It is possible that course numbers could help predict whether a course is introductory or advanced if, for instance, course numbers starting with a “1” are more likely to be introductory than those starting with a “4.” This possibility is one of many reasons we discuss transformer-based approaches that better model the full semantic meaning of a course title, subject code, and catalog number in future research.

subject and one for the course title. Each bag of words matrix is a D by V matrix where D is the number of documents (course records in our corpus) and V is the number of tokens that occur 50 or more times in our training corpus. For a given document, d, and token in our vocabulary, v, cell $\{d,v\}$ of the bag of words matrix contains the number of times token v occurs in document d.

Table 3 shows an example of a bag of words matrix using the course titles from courses offered at public institutions in Texas. For instance, because the token "lab" occurs one time in the document "LAB EXPRMNT BIO-CELL & MOL BIO", the corresponding cell {"LAB EXPRMNT BIO-CELL & MOL BIO", "lab"} contains a value of one. Similarly, the token "bio" occurs twice in this document, so the corresponding cell contains a value of two. As noted above, we create a separate bag-of-word matrices for subject codes and for course titles. The subject code vocabulary contains 1,834 tokens while the course title vocabulary contains 3,571 tokens. For both the training and test datasets, we then concatenate the subject code and course title matrices together to create our final datasets.

V. Methods

Training Models

Given our derived set of features and the associated CCM labels, we train a machine learning model to predict the appropriate CCM code to apply to each course. After training, this model can then be used to predict CCM codes on unseen transcript data. For each of the two-, four-, and six-digit CCM codes, we train a regularized multinomial logistic regression model on the training dataset to predict the appropriate CCM code, given the concatenated bag of words

matrix as features.[7] Because of their relative simplicity and strong performance on text data, regularized logistic regression models are considered baseline models in text classification problems (Jurafsky & Martin, 2023).[8] This produces a logistic regression model with a set of C by V coefficients where C is the number of CCM codes and V is the number of tokens in our concatenated bag of words. Cell $\{c, v\}$ contains the learned coefficient for CCM code c associated with token v.

Throughout the process of cross-validation and evaluation we use the model's accuracy (i.e.,the percent of CCM codes correctly assigned) as our evaluation metric. To reduce the overall complexity of the model and mitigate risks of the model overfitting the training data, we use an L2 norm to penalize coefficients that are far from zero. We identify the best performing penalty term for regularization using five-fold cross-validation within the training data.

After identifying the best performing hyperparameter for our L2 norm, we refit the model on the full 90 percent training data and evaluate the accuracy of predictions on the unseen 10 percent training data. For each token in our training data, the model learns a coefficient weight for each of the possible CCM categories that minimizes the cross-entropy loss based on prediction errors in our training data. When given a new course record to classify, the model then uses those learned weight to predict the probability that the course record should be assigned to

[7] We train each of our three models using scikit-learn's LogisticRegressionCV, using fivefold cross-validation with 500 maximum iterations and defaults for all other parameters (Pedregosa et al., 2011).

[8] While it is beyond the scope of this paper to discuss appropriate machine learning strategies for distinct types of data, we provide a set of references for interested readers. Hastie et al. (2009) provide a book length introduction to machine learning methods and common models from a statistical perspective. Mullainthan and Spies (2017) provide a general overview of machine learning applications in the social sciences. Grimmer et al. (2022) provide a book length treatment of machine learning using text as data, largely drawing on examples from political science. For discipline specific overviews of text as data, we refer interested readers in education to Fesler et al. (2019), in economics to Gentzkow et al. (2019) and Dell (2024), and in sociology to Evans and Aveces (2016) and Garip & Macy (2024).

each CCM code conditional on the course record's subject code and course title. Assuming the joint distribution of text and true CCM codes the new course record is drawn from resembles that of our training data, our model is able to predict the statistically most probable CCM code for that record and our model's performance on the unseen test dataset will approximate performance on new data. This provides us with final models that can be used to predict the most probable CCM code for each course at the two-, four-, and six-digit level.

Model Evaluation

To approximate the model's performance on unseen administrative data, we evaluate the performance of the predictive models on each of the two-, four-, and six-digit levels. Because the features used to predict CCM courses vary at the course-level but many analytical uses will focus on student-level course enrollments, we perform our final evaluation at both the course and enrollment levels. In the course-level evaluation, our sample is the set of unique subject code - course title - CCM code triplets. In the enrollment-level evaluation, we calculate the number of course enrollments associated with each unique subject code - course title - CCM code and weight the evaluation accordingly. At both the course and enrollment level, we evaluate the accuracy of our predictive models on both the training and test data.

Further, because we can easily derive higher level predictions from lower-level predictions, we evaluate the accuracy of the four-digit model in predicting two-digit codes and the six-digit model in predicting two- and four-digit codes. For instance, if our six-digit model predicted the six-digit CCM code 45.0603 - econometrics and quantitative economics, we derived the two-digit code 45 - social science - and the four-digit code 45.06 - economics. We can then compare the accuracy of these derived codes against the ground-truth classification made by human annotators. The results of our model evaluation are summarized in Table 4.

As Table 4 shows, our models have strong performance on both the two- and four-digit levels, respectively classifying 87.4 and 79.5 percent of course enrollments correctly. Further, despite the six-digit classification posing a challenging task of classifying short strings with abbreviations and uncommon spellings into one of 1,622 categories, our model correctly classifies 71.9 percent of course enrollments correctly at the six-digit level. Comparing our model's classification accuracy on courses and course enrollments, we further see that the model's accuracy is higher when we weight records according to their total enrollments.

Examining the accuracy of the two- and four-digit codes derived from the six-digit classifications, we can see that they perform comparably to the models trained to directly classify two- and four-digit levels. Thus, even when the model classifies a course incorrectly at the six-digit level, it frequently classifies those courses into similar CCM codes at higher levels in the typology. In part this may be because, as noted above, there are inconsistencies in the annotated training data at finer levels of granularity.

For researchers interested in describing course-taking trends across sectors, a plausible concern is that the model may perform better on some types of institutions than others, worrying that our model performs when classifying the course-taking at large R1 institutions while performing worse at small community colleges because larger institutions may have better administrative data available that makes classification easier. In Appendix A, we present evidence that our model performs well across institutional sector and level by linking courses in the PETS data with institutional characteristics from the Integrated Postsecondary Education Data System (IPEDS). Across the sectors and levels for which PETS has a meaningful number of courses, we see relatively similar performance although the model does perform best on the four

sectors and levels with the most representation in our training data: public two- and four-year institutions and private for- and not-for profit four-year institutions.

Although general guidelines for adequate performance in text classification tasks in the social sciences are difficult to articulate given the many possible use cases (Grimmer et al., 2022), our model performance is in line with or out-performs text classification in other social science settings. For instance, Atalay et al. (2020) classify the text of newspaper job ads into four- and six-digit Standard Occupational Classification codes with 53 percent accuracy on the four-digit task and 36 percent accuracy on the six-digit task. Cuffe et al. (2022) use the text of establishments collected from the Google Places API to classify employers into two-digit North American Industry Classification System codes with 59 percent accuracy. Using these results as a baseline of acceptable model performance, our classification model appears well-suited for the task of classifying courses into appropriate CCM codes at scale.

Error Analysis

Despite the model's relatively high accuracy, it is important to understand the types of classification errors our model is making and the limitations of our modeling strategy. While adequate for social science research, our model's performance suggests there may be room for improvement with more sophisticated modeling strategies. We perform an error analysis, examining a random sample of 100 records from the test data that the six-digit classification model classifies incorrectly and identify the type of error the model. We identify systematic errors in 96 of the 100 sampled records. This process of error analysis helps researchers understand the kinds of errors our supervised machine learning model is making, identifying systematic types of errors and suggesting modifications to our modeling strategy would be fruitful (Géron, 2022).

In eight of the 100 cases, it is not clear that the model's prediction is incorrect and that the human annotation is correct. For instance, our model classifies a 'higher education leadership' course into the code for higher education administration, 13.0406, while the human annotation classifies this course into the code for educational leadership, 13.0401. Similarly, the model classifies an engineering course in thermodynamics into the engineering physics code, 14.1201, while the human annotation places the course in the general physics code 40.0801. In an additional three of the 100 courses, the model confuses the other category and the general category for a particular discipline. This likely occurs because the division between these two codes is more ambiguous than that of specific courses within a discipline.

In 26 of the 100 cases, the model appears to be incorrect because our logistic regression model does not allow for interactions between tokens (i.e., the contextual meaning of words), and individual tokens are strongly associated with specific codes. For instance, the model classifies an 'atmospheric sciences' course with "orientation" in the title into the code for college orientation courses, 90.9997, rather than the general code for atmospheric sciences and meteorology, 40.0401. This occurs because the single token "orientation" is strongly related to the college orientation code, although a human annotator considering the context of the title and subject code would conclude the course is an orientation to the subject of atmospheric sciences, not to college.

In 53 of the 100 cases, it is not clear that a course title and subject number as recorded in administrative transcripts will be sufficient to classify the courses at the six-digit CCM code. For instance, the model incorrectly classifies some physical therapy courses because titles like "Physical Therapy" cannot identify whether a course should be classified as physical therapy, 51.2308, or physical therapy assistant, 51.0806. It may be difficult to improve on our six-digit

classification models in these cases without the context provided by the course catalogs used by human annotators. In an additional two of the 100 cases, we encounter similar challenges with polysemy (or words that have the same spelling but different meanings depending on the context of language). Issues of polysemy would also be addressed by additional information from course catalogs.

In four of the 100 cases, the model encounters an uncommon token that has clear meaning to human annotators but occurs fewer than 50 times in the training dataset and is not included in our bag of words representation of the text. For instance, because the text in our dataset is drawn from administrative records that may have character limits on the length of strings, we see uncommon abbreviations of common tokens like “apprectn” for “appreciation.” While a human annotator can intuit that “apprectn” is an abbreviation for “appreciation”, our bag of words model treats each token independently and does not understand the shared meaning of these two tokens. Because “apprectn” occurs fewer than 50 times, it is not included in our bag of words representation of the text. In the context of a course like “Theater apprectn,” this results in a classification of theater arts, 50.0501, rather than the code for drama and theater appreciation, 36.0117. Issues of uncommon tokens would likely be addressed by incorporating the richer text of course descriptions or by fine-tuning large language models like RoBERTa for sequence classification (Liu et al., 2019) that have access to a larger vocabulary through their pre-training.

VI. Application

Having demonstrated the *ability* of our model to classify courses at scale, we now turn to demonstrating the *utility* of doing so. In this section, we use the classifications produced by our model to present new evidence on course-taking trends in colleges and universities in the United States. As discussed in Section II, this is an improvement over simply looking at the number of

degrees earned for understanding the true level of engagement with different disciplinary content among undergraduates.[9]

To fully illustrate the utility of our classification model, we take a descriptive exploratory approach, looking at who enrolls in courses that teach liberal arts and professional field content from several different angles. Our explorations begin by comparing how enrollment in majors compares to enrollment in coursework, both within the humanities and liberal arts and in other broad disciplinary groupings. We also examine the degree to which students in different majors take courses in a variety of disciplines shedding light on the breadth of students' learning during college. Finally, we use the size of our administrative data to examine mathematical course-taking across the intersection of race and sex. The results contextualize the 'liberal arts are in decline' narrative while pointing to surprising differences in the breadth of subjects students in different liberal arts and professional majors engage with.

Data and Methods

We document trends in course-taking with transcript data from the College and Beyond II (CBII) study (Courant et al, 2022). CBII is distributed by the Inter-university Consortium for Political and Social Research (ICPSR) and contains student record and transcript data on over 1 million bachelor's-seeking undergraduates enrolled from 2000-2021 at 19 public colleges across seven postsecondary systems in the United States (Bell et al., 2022). While not nationally representative, these institutions are quite diverse, including large research universities and

[9] Others have lamented the lack of course-level data as well. See Schmidt (2018), discussing the state of the humanities since the Great Recession in *The Atlantic*: "College degrees are a somewhat problematic metric: I'd rather see information about the type and level of courses that undergraduates take. This kind of data is hard to come by except at an anecdotal level."

liberal arts colleges in urban and suburban campuses across seven states.[10] In the following analyses, we focus on baccalaureate graduates from six systems comprised of 18 institutions that provided course titles in their transcript data and for whom we can make CCM classifications. We construct an analytical sample of all baccalaureate graduates at these systems between 2009 and 2019, comprising 455,389 students. We link student-level information with transcript data, comprising 17,992,577 course enrollments. We identify the major each student graduated with, and in cases in which the student graduated with two or more majors, use the student's first major as indicated by the systems.

The student transcript data in CBII contains a subject number (e.g., CHEM 241) and course title (e.g., Intro to Chem Analy) for each course a student enrolled in. To standardize this information across institutions we apply our four-digit CCM classification model to the analytical sample. Our model assigns a four-digit CCM code to each course enrollment in the CBII data that is not missing a subject code and course title. We then associate each four-digit CCM code in the data with one of the following disciplines: humanities, social and behavioral sciences, natural sciences, fine and performing arts, business and management, health and medical, education, engineering, and other/unknown. These disciplinary groupings were developed by American Academy of Arts and Sciences (AAAS) for their Humanities Indicators Project.[11] This creates a course enrollment-level dataset that can be disaggregated by discipline.

[10] The participating institutions include all four-year institutions City University of New York, Georgia College and State University, Indiana University-Bloomington, Truman State University, University of California-Irvine, University of Houston, and the University of Michigan. Researchers can apply to access the data at https://www.icpsr.umich.edu/web/about/cms/4369

[11] Additional information about the Humanities Indicators Project and the classification of disciplines into these categories is available at https://www.amacad.org/humanities-indicators/higher-education/bachelors-degrees-humanities. The AAAS disciplines are created for six-digit CIP codes but there is not perfect correspondence between CIP and CCM codes. For instance, the CCM code for advanced statistics is 27.0598 but there is not a corresponding CIP code. To map CCM codes to disciplines, we collapse the AAAS categories to the four-digit level

Table 5 displays an example course enrollment record from the CBII dataset for each of the disciplinary groupings.

To create a major level dataset, we perform an analogous process; we collapse student majors to the disciplinary level by mapping four-digit CIP codes to the same disciplinary groupings from the AAAS Humanities Indicators project.[12] This creates a major-level dataset that can be disaggregated by discipline. In Table 6, we show similar examples by major, with the major title, the associated Classification of Instructional Programs Code, and the discipline that major is assigned to.

Aggregate Measures. In the following sections, we examine longitudinal trends in disciplinary majors and disciplinary coursework by graduating cohort. For discipline *i* in academic year *j*, the value of discipline *i*'s coursework in year *j* is given by:

$$Prop.of\ Coursework_{ij} = \frac{Number\ of\ Course\ Enrollments\ in\ Discipline_i\ among\ \ Year_j\ Graduates}{Number\ of\ Course\ Enrollments\ among\ Year_j\ Graduates}$$

Similarly, for discipline *i* in academic year *j*, the value of discipline *i*'s majors in year *j* is given by:

$$Prop.of\ Majors_{ij} = \frac{Number\ of\ Majors\ in\ Discipline_i\ among\ Year_j\ Graduate}{Number\ of\ Year_j\ Graduates}$$

and assign each four-digit code to the discipline that the majority of six-digit codes belonged to. In a small number of cases, we manually annotate a AAAS code because there is no four-digit CIP code that corresponds to the four-digit CCM code (for example, the four-digit CCM code 90.99 contains courses like study abroad and college orientation. There is not a corresponding four-digit CIP code and we manually annotated this CCM code as "Other/Unknown."

[12] Four of the six CBII systems provided six-digit CIP codes. For the two systems that did not provide this data, a CBII team member annotated the text descriptions of each major with the appropriate six-digit CIP code, using institutional websites if necessary. We use the four-digit CIP code derived from these six-digit codes to map majors into the AAAS disciplines.

With these measures created, we then describe course-taking over the past decade at institutions that participated in the CBII study.

Results

Majors vs. Course-taking. In Figure 2, we contrast the proportion of total majors in the four liberal arts disciplines with the proportion of total course enrollments in those disciplines over time. For instance, in Panel A we examine trends in humanities majors and coursework. Consistent with broader narratives concerning the decline of the humanities, Panel A shows a dramatic decline in humanities majors (from 13.7% in 2009 to 9.7% in 2019). Panel B shows a slight increase in natural science majors while Panel C and D show slight decreases in fine arts majors and social science majors, respectively.

Although trends in course-taking in each discipline qualitatively mirror the direction of trends in majors in that discipline, course-taking in the humanities, natural sciences, and fine arts is much more common than majoring in those disciplines. When measured as a proportion of enrollments, the humanities and natural science are twice as large as when measured by majors in our sample. In contrast to the three other core liberal arts disciplines, course-taking in the social sciences is less common than majors. Still, however, the social sciences comprise approximately 17 percent of course-taking over the decade under study. In sum, the liberal arts disciplines comprise a clear majority of course-taking throughout the time period we study (66 percent of courses among graduates from 2009 and 63 among graduates from 2019).

In Figure 3, we perform the same analysis within the professional disciplines, contrasting the proportion of majors in the four disciplines with the proportion of enrollments. We observe an increase in the share of engineering majors and relatively stable shares of education, business,

and health majors. Again, trends in coursework qualitatively mirror trends in majors over time directionally. However, across all four professional disciplines, the proportion of coursework in a discipline is meaningfully lower than that of majors. In contrast to majoring in the liberal arts, majoring in professional disciplines is more common than course-taking in those disciplines.

Comparing trends across disciplines, we see that in 2009, the humanities comprised the third largest discipline by major, but, by 2019, the three largest disciplines in our sample by major are business, the social sciences, and engineering. If the number of humanities majors in 2009 had stayed constant throughout the decade our data covers, they would still comprise the third largest discipline by major in 2019. However, although there has been a decline in humanities course-taking from 2009 to 2019, the humanities are still the largest discipline by course-taking with the natural sciences representing the second largest discipline and the social sciences representing the third largest discipline. Although professional majors are approximately 50 percent of all majors in 2009, increasing to 55 percent in 2019, three of the four liberal arts disciplines continue to form the academic core of postsecondary course-taking.

In Figure 4, we take a broader view of longitudinal trends in degrees and course-taking in liberal arts and professional disciplines by aggregating our measures further to liberal arts - the humanities, social sciences, natural sciences, and fine arts - and professional disciplines - engineering, education, business, and health and medical disciplines. In Figure 4, Panel A, we present trends in majors over the ten-year time period of our sample, showing a meaningful decline in liberal arts majors. In 2009, graduate's majors were approximately split between the liberal arts and professional discipline but, by 2019, approximately ten percentage points more graduates were majoring in the professional disciplines than the liberal arts disciplines.

We contrast this widely discussed trend in liberal arts majors with that of trends in enrollments in the liberal arts in Figure 4, Panel B. Although course-taking enrollments in the liberal arts decline over our panel's time span, they continue to comprise the clear majority of course-taking. In 2009, approximately sixty-six percent of course-taking was drawn from liberal arts disciplines with a slight decline to sixty-three percent in 2019. In contrast to the ten-percentage point advantage to majoring in professional disciplines relative to the liberal arts in 2019, we observe a twenty-seven percentage point advantage to liberal art enrollments relative to professional coursework.

Course Enrollments by Major. Figure 5 shows the proportion of coursework taken in each discipline, disaggregated by major. In Panel A, we restrict our population to just those students that graduate with a major in the humanities and consider the proportion of those students' course enrollments taken in each discipline, by graduating year. Students that graduate with a humanities major in 2009 took about approximately half of their coursework in their "home" discipline (i.e., the discipline they are majoring in). However, this proportion is declining over time from 51.5% in 2009 to 48.0% of course enrollments in 2019. In other words, humanities majors are taking slightly fewer humanities courses now than they used to take. For humanities majors, two of the three remaining core liberal arts disciplines – the social sciences and the natural sciences - comprise the next two largest disciplines of coursework followed by other/unknown and the final core liberal arts discipline, fine and performing arts. As a result, the four core liberal arts disciplines comprise 83.4% percent of humanities majors' course enrollments in 2009 and 81.2% of humanities majors' course enrollments in 2019.

Panels B, C, and D examine the proportion of course enrollments in each discipline amongst majors in the natural sciences, fine and performing arts, the social and behavioral sciences, and, respectively. Students majoring in the natural sciences and fine and performing arts display relatively similar trends to students in the humanities, with between fifty and sixty percent of course enrollments in their home discipline and coursework in the humanities and social and behavioral sciences comprising the next two most common disciplines from which they draw course enrollments. The number of courses they take in their home discipline appears fairly constant over time.

In contrast, students majoring in the social and behavioral sciences enroll in between ten and twenty percentage points less coursework in their home discipline. Social science majors take meaningfully more coursework in the humanities and natural sciences than their peers majoring in other liberal arts disciplines. Among students who graduated with a social science degree in 2009, the social sciences comprise 37.1% of course enrollments. The natural sciences comprise 15.9% course enrollments, the humanities comprise 25.2% of course enrollments and the fine and performing arts comprise 4.2% of course enrollments. As a result, 82.3% of 2009 social science majors' course enrollments are drawn from the core liberal arts disciplines but just over a third of their course enrollments are drawn from their home discipline. In contrast, 83.4% of 2009 humanities majors' coursework is drawn form the core liberal arts disciplines but 51.5% of that coursework is in the humanities. Social and behavioral science majors spend relatively less time specializing in their home discipline and more time in other liberal arts fields. This likely explains why the social sciences was the only discipline amongst the liberal arts disciplines with a greater proportion of majors than course enrollments

In Figure 6 we examine the proportion of coursework in the professional disciplines: business and management, education, engineering, and health. Among students majoring in the four professional disciplines between 2009 and 2019, three of the four core liberal arts disciplines – the humanities, natural sciences, and social sciences - comprise the three largest non-home discipline course enrollments. Among students that graduate with engineering or education majors, for instance, coursework in the natural sciences comprise more than one fifth of their course enrollments.

Across all disciplines, we see little change in the distribution of disciplines they draw their coursework from between 2009 and 2019. The sole exception to this is business and management. In Figure 6, Panel D, we see an apparent increase in disciplinary coursework, and thus specialization, among students graduating with business and management degrees between 2009 and 2019. In 2009, among students that graduate with a business and management degree, 47.6% of course enrollments are drawn from business and management disciplines and by 2019 this increases to 52.2% of course enrollments in business and management.

The relative stability of coursework by discipline stands in contrast to aggregate trends in disciplinary majors and coursework which show large shifts across disciplines, notably large decreases in humanities majors and increases in engineering majors. We do not see evidence that students that had an interest in majoring in the humanities but ultimately choose a professional degree would enroll in more humanities coursework. If it were true that prospective liberal arts students were more likely to major in professional degrees over time while enrolling in more liberal arts coursework, we anticipate that the proportion of liberal arts coursework professional majors enroll in would be increasing as more students major in those disciplines. We do not

observe this trend. Indeed, when we observe changes in disciplinary makeup of courses among majors in business, we see a slight increase in specialization within the discipline, not a decrease.

Mathematics Coursework and Majors. Since the beginning of the liberal arts tradition, mathematics has been a core component of a traditional liberal arts education (Kimball, 2010). In this section, we document trends in mathematics coursework and majors. In Figure 7, we present longitudinal trends in the proportion of coursework and the proportion of majors in mathematics. Panel A shows a large and consistent gap between coursework in mathematics and majors. Over our sample period, we estimate that mathematics coursework increases from six to seven percent of total coursework while majors increase from one to two percent.

Relative to surveys, one of the virtues of administrative data is that we have sufficient sample size to examine heterogeneity of relationships at the intersection of identities (Figlio et al., 2017; Viano & Baker, 2018). In Panel B, C, and D, we further analyze course enrollments in mathematics by sex, by race/ethnicity, and by the intersection of race/ethnicity and sex.[13] We stress in these analyses that they are purely descriptive and reflect the association between the proportion of math coursework students enroll in and the socially constructed categories of race and sex recorded in administrative data. In Panel B, we show longitudinal trends in mathematics majors and course enrollments by sex, with male students enrolling in approximately two

[13] While the size of our administrative data allows us to conduct analyses within finer-grained subgroups than that of survey data like the intersection of sex and gender, our measures of sex and race are constrained by the nature of administrative data and, even in the case of our large dataset, contain small populations that do not meet disclosure risk review requirements. In the case of our measures of race/ethnicity, the measures reflect the federal reporting categories for IPEDS, although this provides only a coarse-grained measure of the spectrum of socially constructed race and ethnicity groups (Viano & Baker, 2020). Regarding our measures of sex, partner systems in the CBII system typically provided binary variables with male and female values while two systems providing a small number of values outside this binary that we cannot provide descriptive statistics on because they do not meet minimum sample sizes for disclosure risk review. Further, at all partner systems, it is unclear whether the system provided sex assigned at birth, legal sex or gender at time of matriculation, or legal sex or gender identity at some other point in time, consistent with the broader lack of conceptual clarity around measures of sex and gender in education research (Garvey et al., 2019).

percentage points more mathematics courses than female students. In Panel C, we show longitudinal trends in mathematics course enrollments across four racial/ethnic groups - white, Black, Asian, and Hispanic students. Interestingly, the share of enrollments in math are quite similar in 2019 for white, Hispanic, and Black students. Finally in Panel D, we show longitudinal trends in mathematics course enrollments across the intersection of sex and those four racial/ethnic groups. We find that gaps across the intersection of race and sex are roughly additive. Across the four racial/ethnic groups we examine, male students enroll in between one and a half and two percentage points more mathematics courses relative to female students.

VII. Potential Improvements to Classification Models

In this work, we have shown that our regularized logistic regression model - a baseline text classification model - provides strong enough performance on the classification task to allow us to meaningfully examine course-taking patterns. Given the lessons of our error analysis, however, there are likely ways to improve our classification models. This section describes possible improvements of our modeling strategy including going beyond the bag of words representations of our text data, incorporating additional textual information like course descriptions, and using the structure of student transcripts in classification.

Bag of Words Representations

In this work, we used one of the baseline models for text classification in natural language processing, regularized logistic regression, because it is simple to implement and computationally undemanding for a state data center or university institutional research office to use for inference. Although our logistic regression models perform well enough for our analysis and many social science applications, applications of state-of-the-art natural language processing

typically use large language models, large neural networks using some variant of the Transformer architecture (Dell, 2024; Tunstall et al., 2022). These models are trained on language modeling tasks such as masked language modeling with large quantities of text, allowing the model to learn abstract mathematical representations of text that better capture semantic meaning and the interdependencies of language. A large language model can then be fine-tuned on a specific classification task like ours. As a result of the language modeling task, the model has mathematical representations of text that allow for similar tokens like "apprectn" and "appreciation" to share meaning and allows for the interdependence of meaning, like the relationship between "atmospheric sciences orientation" referring to an orientation to the atmospheric sciences, not a college orientation. We have explored preliminary work fine-tuning the base RoBERTa model on our classification tasks (Liu et al., 2019), finding meaningful gains in classification accuracy. A fine-tuned RoBERTa model correctly classifies 84% of courses at the two-digit level and 90% of enrollments, 75% of courses at the four-digit level and 82% of enrollments, and 65% of courses at the six-digit level and 75% of enrollments. This represents an improvement of approximately three percent on the accuracy of classifying enrollments at each of the two-, four-, and six-digit levels.

Incorporating Course Descriptions

The human annotation process for the College Course Map relies on course descriptions but our models use just the subject code and course title as recorded in administrative transcripts (Bryan & Simone, 2012). As a result, our models do not have access to all the data required for human annotators to make classifications. As discussed in the error analysis, 55 of the 100 cases reviewed during our error analysis seem to require additional information beyond the subject codes and course titles provided with the PETS data. For instance, we noted the model will

struggle to classify some physical therapy courses because titles like “Physical Therapy” are insufficient to identify whether a course should be classified as physical therapy, 51.2308, or physical therapy assistant, 51.0806. However, adding the additional context about course content present in a record’s course description would likely allow us to differentiate between these two types of courses. Because institution course catalogs are typically publicly available online, and often organized in searchable web databases, it is possible to collect course descriptions at scale and link them with the transcripts from the PETS study, allowing the model to disambiguate between courses with similar titles but different content.

Hierarchical Classification

Our models do not rely on the hierarchical nature of the College Course Map codes, but there is likely useful information in each of the two-, four-, and six-digit labels. For instance, knowing that a statistics course belongs to the two-digit social sciences code, 45, instead of the two-digit mathematics and statistics code, 27, may allow a model to make a correct classification at the six-digit level of “Research Methodology and Quantitative Methods,” 45.0102, rather than the six-digit “Statistics, General” code, 27.0501. Hierarchical classification techniques offer a variety of alternatives to the flat classification approach we used that allow models to use the additional information present in the hierarchy of labels (for a review, see Silla & Freitas, 2011). Rather than train three separate models for the two-digit, four-digit, and six-digit classification task, future research might train a single model that incorporates the hierarchical structure of the CCM annotations. Indeed, in a similar setting to ours - classifying the text of patents into a standardized hierarchical taxonomy - Pujari et al. (2021) find that Transformer models that incorporate hierarchical information reach state of the art levels of performance by sharing information across classification tasks in each level of the hierarchy.

Classifying Courses in the Context of Transcripts

Our classification has removed course textual features - subject code and title - from students' actual transcripts and the institutions they attend. However, courses' location on transcripts and student majors may contain useful additional information. For instance, Adelman (2004) discusses how human annotators use the surrounding context of a student's transcript to disambiguate approximately 4,000 courses, discussing the case of a course with title "composition" and no indication of department that could plausibly be either music composition or Russian composition. However, the course occurred in the student's third year, in the context of a transcript with many Russian courses, so the course was assigned to Russian language, for that student. Jiang & Pardos (2020) show that the structure of student transcripts contain information that can be leveraged for classification tasks and offer a useful starting point for incorporating the structure of transcripts into future classification models.

VIII. Conclusion

In documenting trends in course-taking at scale over two decades at 18 four-year colleges, we make five main conclusions. First, liberal arts course-taking is more common than liberal arts majors, comprising approximately two-thirds of all course-taking and three of the four liberal arts disciplines - the humanities, social sciences, and natural sciences - are the three most common course-taking disciplines. Conversely, professional majors are more prevalent than professional course-taking. Thus, evaluating disciplines based on majors overstates the place of pre-professional disciplines and understates the role that the liberal arts continues to play as the academic core of postsecondary education in the US. Measured as a proportion of coursework, the humanities and natural sciences are twice as prevalent as when measured as a proportion of majors. In part, this is likely because all students must complete a general

education curriculum, drawing primarily from the liberal arts disciplines, and, in part, this is likely because the disciplines build on the core liberal arts disciplines - for instance, engineering requires a substantial amount of mathematics. While we have shown descriptively that the core liberal arts are a larger proportion of total coursework, we have not examined why this is. Future research might explore the role of structure and student choice within that structure play in these trends by accounting for general education and major requirements in the analysis. Further, researchers might investigate how the gendered and racialized dynamics of iterative course-taking decisions and differences in major selection can contribute to these course-taking patterns (building on, for example, Baker & Orona, 2020 and Dalberg et al., 2024).

Second, we contextualize narratives about the decline of the humanities. While humanities course-taking has experienced a decline over the decade of our study, humanities course-taking still represents the plurality of courses. Given this, budget decisions made on the basis of majors rather than course-taking will understate the true role of the humanities in postsecondary education.

Third, many liberal arts experiences are infused in professional majors course-taking. Reflecting one of the core aims of general education programs to ensure all college graduates receive a well-rounded liberal education, three of the four liberal arts disciplines are the most common disciplines students majoring in professional majors take courses in outside of their home discipline.

Fourth, we demonstrate our applications' ability to examine heterogeneous relationships due to the large sample size of administrative data: we examine trends in mathematics course-taking across the intersection of race/ethnicity and sex. In our context, we find that differences across sex and race/ethnicity are roughly additive.

Our examination of trends in course-taking at multiple institutions and at large scale was made possible our fifth major contribution, an algorithm we developed to classify courses by CCM code. We found that using subject codes and course titles is sufficient to accurately predict CCM codes for course enrollments at a fine-grained level 72 percent of the time. However, we see this algorithm as offering a baseline for future classification models that draws on more sophisticated modeling strategies and incorporates additional data from transcripts and course catalogs. Drawing on our error analysis, we outlined several steps for improving our classification algorithms that will likely further improve the models, allowing researchers to efficiently classify courses into the standardized CCM typology at scale. We released our current best performing classification models as open-source software.[14]

The rise of large-scale administrative data in higher education has provided researchers with unprecedented access to student transcripts, allowing researchers to trace the course-taking pathways students use to navigate postsecondary curricula. However, linking courses from administrative transcripts to theoretical constructs is difficult because course transcripts are largely unstandardized. Our classification algorithms provide tools to make these course-taking pathways legible across institutions, allowing researchers to measure course-taking in the context of the traditional benefits of administrative data detailed by Figlio et al. (2017): the ability to more credibly estimate causal effects and examine heterogeneous effects within subgroups. We believe this opens up new avenues of research by, for example, letting researchers exploit

[14] Due to the fact that our logistic regression models rely on information about text derived from restricted use data when defining the bag-of-words representations of text, the model weights can only be distributed to users through the National Center for Education Statistics' restricted-use data licensing program. However, researchers can download the RoBERTa classification models developed in this paper to place their own course-level transcript data into the CCM schema and access a demonstration notebook at https://huggingface.co/collections/annamp/classifying-courses-at-scale-66e470e6126f9577f89fd417. We anticipate further refinements to our model strategies inspired by our error analysis and other improvements will be available at that link.

random shocks to understand how changes in financial circumstances might shift students out of liberal arts coursework and into more professional course-taking trajectories or explore gendered and racial differences in the ways students navigate the curriculum at scale.

References

Adelman, C. (1990). *A college course map: Taxonomy and transcript data based on the postsecondary records, 1972-1984, of the high school class of 1972*. US Department of Education.

Adelman, C. (2004). *The empirical curriculum: Changes in postsecondary course-taking, 1972-2000*. Institute of Education Sciences, US Department of Education.

Andersen, H. (2020). General education. In M. E. David & M. J. Amey (Eds.), *The SAGE Encyclopedia of Higher Education*. SAGE Publications, Inc.

Arteaga, Carolina, 2018. "The Effect of Human Capital on Earnings: Evidence from a Reform in Colombia's top University." *Journal of Public Economics*. Vol 157, January. pp 212-225

Atalay, E., Phongthiengtham, P., Sotelo, S., & Tannenbaum, D. (2020). The Evolution of Work in the United States. *American Economic Journal: Applied Economics*, *12*(2), 1–34. https://doi.org/10.1257/app.20190070

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis. *Review of Educational Research*, *74*(1), 29–58.

Baker, R., & Orona, G. A. (2020). Gender and Racial Differences in Awareness and Consideration of Curricular Programs: Exploring a Multistage Model of Major Choice. *AERA Open*, *6*(3). https://doi.org/10.1177/2332858420937023

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

Bell, D., Byrd, W.C., Flaster, A., Koester, B., Leonard, S., Manley, K., Nishimura, R., Paulson, A., & Stange, K.M. (2022). *College and Beyond II User Guide*. https://www.icpsr.umich.edu/web/about/cms/4369

Blair, R., Kirkman, E., & Maxwell, J. (2018). *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States: Fall 2015 CBMS Survey*. American Mathematical Society. https://doi.org/10.1090/cbmssurvey/2015

Blom, E., Cadena, B. C., & Keys, B. J. (2021). Investment over the Business Cycle: Insights from College Major Choice. *Journal of Labor Economics*, *39*(4), 1043–1082. https://doi.org/10.1086/712611

Bradburn, M., & Townsend, R. (2023). *Humanities Indicators Project*. American Academy of Arts and Sciences. https://www.amacad.org/humanities-indicators/higher-education/bachelors-degrees-humanities

Brint, S., Riddle, M., Turk-Bicakci, L., & Levy, C. S. (2005). From the Liberal to the Practical Arts in American Colleges and Universities: Organizational Analysis and Curricular Change. *The Journal of Higher Education*, *76*(2), 151–180. https://doi.org/10.1080/00221546.2005.11778909

Brint, S., Proctor, K., Murphy, S. P., Turk-Bicakci, L., & Hanneman, R. A. (2009). General Education Models: Continuity and Change in the U.S. Undergraduate Curriculum, 1975–2000. *The Journal of Higher Education*, *80*(6), 605–642. https://doi.org/10.1080/00221546.2009.11779037

Bryan, M., Caperton, S.A., Cooney, D., and Dudley, K. (2020). *2012 Beginning Postsecondary Students Longitudinal Study (BPS:12) Postsecondary Education Transcript Study (PETS)*

Data File Documentation (NCES 2021-176). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2021/2021176.pdf

Bryan, M. & Simone, S. (2012). *2010 College Course Map* (NCES 2012-162rev). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2012/2012162rev.pdf

Cohen, P. (2016, February 22). A rising call to promote STEM education and cut liberal arts funding. *The New York Times*. https://www.nytimes.com/2016/02/22/business/a-rising-call-to-promote- stem-education-and-cut-liberal-arts-funding.html

Courant, P. N., Flaster, A, Jekielek, S., Levenstein, M., McKay, T. A., & Stange, K.M. (2022) *College and Beyond II (CBII) Administrative Data, [United States], 2000-2021* [Data set], https://doi.org/10.3886/ICPSR38488.v2

Cuffe, J., Bhattacharjee, S., Etudo, U., Smith, J. C., Basdeo, N., Burbank, N., & Roberts, S. R. (2022). Using Public Data to Generate Industrial Classification Codes. In K. G. Abraham, R. S. Jarmin, B. C. Moyer, & M. D. Shapiro (Eds.), *Big Data for Twenty-First-Century Economic Statistics* (pp. 229–246). University of Chicago Press. https://doi.org/doi:10.7208/chicago/9780226801391-010

Dalberg, T., Cortes, K. E., & Stevens, M. L. (2024). Major Selection as Iteration: Observing Gendered Patterns of Major Selection Under Elective Curriculums. *AERA Open*, *10*. https://doi.org/10.1177/23328584241249600

Dell, M. (2024). *Deep Learning for Economists* (Working Paper 32768). National Bureau of Economic Research. https://doi.org/10.3386/w32768

Denson, N., & Bowman, N. A. (2017). Do Diversity Courses Make a Difference? A Critical Examination of College Diversity Coursework and Student Outcomes. In M. B. Paulsen (Ed.), *Higher Education: Handbook of Theory and Research* (pp. 35–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-48983-4_2

Devereaux, B. C. (2023, April 2). Opinion | Colleges Should Be More Than Just Vocational Schools. *The New York Times*. https://www.nytimes.com/2023/04/02/opinion/humanities-liberal-arts-policy-higher-education.html

Duprey, M.D., Pratt, D.J., Wilson, D.H., Jewell, D.M., Brown, D.S., Caves, L.R., Kinney, S.K., Mattox, T.L., Smith Ritchie, N., Rogers, J.E., Spagnardi, C.M., and Wescott, J.D. (2020). High School Longitudinal Study of 2009 (HSLS:09) *Postsecondary Education Transcript Study and Student Financial Aid Records Collection Data File Documentation* (NCES 2020-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2020/2020004.pdf

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, *42*(1), 21–50. https://doi.org/10.1146/annurev-soc-081715-074206

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. *Journal of Research on Educational Effectiveness*, *12*(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Figlio, D., Karbownik, K., & Salvanes, K. (2017). The Promise of Administrative Data in Education Research. *Education Finance and Policy*, *12*(2), 129–136. https://doi.org/10.1162/EDFP_a_00229

Garip, F., & Macy, M. W. (2024). Machine Learning in Sociology: Current and Future Applications. In *The Oxford Handbook of the Sociology of Machine Learning*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197653609.013.11

Garvey, J. C., Hart, J., Metcalfe, A. S., & Fellabaum-Toston, J. (2019). Methodological Troubles with Gender and Sex in Higher Education Survey Research. *The Review of Higher Education*, *43*(1), 1–24. https://doi.org/10.1353/rhe.2019.0088

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Harris, A. (2018, December 13). The Liberal Arts May Not Survive the 21st Century. *The Atlantic*. https://www.theatlantic.com/education/archive/2018/12/the-liberal-arts-may-not-survive-the-21st-century/577876/

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. https://hastie.su.domains/ElemStatLearn/

Heller, N. (2023, February 27). The End of the English Major. *The New Yorker*. https://www.newyorker.com/magazine/2023/03/06/the-end-of-the-english-major

Jiang, W., & Pardos, Z. A. (2020). Evaluating Sources of Course Information and Models of Representation on a Variety of Institutional Prediction Tasks. *International Educational Data Mining Society*.

Jurafsky, D., & Martin, J. (2023). *Speech and Language Processing* (3rd ed.). Self Published. https://web.stanford.edu/~jurafsky/slp3/

Kimball, B. A. (2010). *The Liberal Arts Tradition: A Documentary History*. University Press of America.

Kizilcec, R. F., Baker, R. B., Bruch, E., Cortes, K. E., Hamilton, L. T., Lang, D. N., Pardos, Z. A., Thompson, M. E., & Stevens, M. L. (2023). From pipelines to pathways in the study of academic progress. *Science*, 380(6643), 344–347. https://doi.org/10.1126/science.adg5406

Lattuca, L. R., & Stark, J. S. (2009). *Shaping the college curriculum: Academic plans in context*. John Wiley & Sons.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education Data Science: Past, Present, Future. *AERA Open*, *7*. https://doi.org/10.1177/23328584211052055

Moen, M. (2020). Opportunity knocks for liberal education: How liberal education has an exceptional opportunity to help fix what most ails our nation today (opinion). *Inside*

Higher Ed.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

National Center for Education Statistics. (2019). Table 311.70: Course enrollments in languages other than English compared with total enrollment at degree-granting postsecondary institutions, by enrollment level, institution level, and language: Selected years, 1965 through 2016. *Digest of Education Statistics*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/digest/d19/tables/dt19_311.70.asp

National Center for Education Statistics. (2023). Undergraduate Degree Fields. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/cta .

Pasquerella, L. (2020). Foreword. In *Teaching for Liberal Learning in Higher Education* by Mary Taylor Huber (p. v). American Association of Colleges & Universities.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Researc*h, 12, 2825–2830.

Pujari, S. C., Friedrich, A., & Strötgen, J. (2021). A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in Information Retrieval* (pp. 513–528). Springer International Publishing. https://doi.org/10.1007/978-3-030-72113-8_34

Riley, N. (2022, December 2). The "Suicide" of the Liberal Arts. *Wall Street Journal*. https://www.wsj.com/articles/the-suicide-of-the-liberal-arts-higher-education-students-teachers-educators-degree-america-learning-11669994410

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schmidt, B. (2018, August 23). *The Humanities Are in Crisis*. The Atlantic. https://www.theatlantic.com/ideas/archive/2018/08/the-humanities-face-a-crisisof-confidence/567565/

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1–2), 31–72. https://doi.org/10.1007/s10618-010-0175-9

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. O'Reilly Media, Inc.

Viano, S., & Baker, D. J. (2020). How Administrative Data Collection and Analysis Can Better Reflect Racial and Ethnic Identities. *Review of Research in Education*, *44*(1), 301–331. https://doi.org/10.3102/0091732X20903321

Wang, X. (2016). Course-Taking Patterns of Community College Students Beginning in STEM: Using Data Mining Techniques to Reveal Viable STEM Transfer Pathways. *Research in Higher Education*, *57*(5), 544–569. https://doi.org/10.1007/s11162-015-9397-4

Warner, D. B., & Koeppel, K. (2009). General education requirements: A comparative analysis. *The Journal of General Education*, *58*(4), 241–258.

Wells, C. A. (2016). Realizing General Education: Reconsidering Conceptions and Renewing Practice. *ASHE Higher Education Report*, *42*(2), 1–85. https://doi.org/10.1002/aehe.20068

Weston, T. J., Seymour, E., Koch, A. K., & Drake, B. M. (2019). Weed-Out Classes and Their Consequences. In E. Seymour & A.-B. Hunter (Eds.), *Talking about Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education* (pp. 197–243). Springer International Publishing. https://doi.org/10.1007/978-3-030-25304-2_7

Wine, J., Janson, N., and Wheeless, S. (2011*). 2004/09 Beginning Postsecondary Students Longitudinal Study (BPS:04/09) Full-scale Methodology Report* (NCES 2012-246). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2012/2012246.pdf

Wine, J., Janson, N., Siegel, P., Bennett, C. (2013). *2008/09 Baccalaureate and Beyond Longitudinal Study (B&B:08/09) Full-scale Methodology Report* (NCES 2014-041). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2014/2014041.pdf

Appendix A

Model Performance by Institutional Control and Level

In Table A1, we evaluate our baseline logistic regression model's classification enrollment weighted accuracy on the test set on two-, four-, and six-digit classification by IPEDS sector, showing that our model's performance is comparable across a diversity of institutions in the US postsecondary system. Note that because this is an enrollment weighted accuracy, we report the accuracy across the 262,280 course enrollments that occur within course sections in the test dataset, not on the accuracy among the 89,130 course sections. Because not all CCM transcripts have term and academic codes, it is not possible to exactly link courses to institutional sectors as they are when courses are offered: we approximate the sector of the institution offering the course with the institution's current sector as of 2023. We successfully match 99.7% of course enrollments in the test set with a sector from IPEDS. While our models perform adequately across all sectors, we do find that, perhaps unsurprisingly, classification accuracy is generally poorer on private for-profit 2 and less than 2-year institutions and public less than 2-year institutions; these are the institutions for which we have less training data, with the three sectors collectively comprising just 1 percent of total course enrollments in the training set.

Table 1
Examples of Courses from Public Texas Institutions in PETS Format

Subject Code	Course Title	College Course Map Code
BIO 205L	LAB EXPRMNT BIO-CELL & MOL BIO	26.0204 - Molecular Biology
BIOL 3415	INTRO TO MOLECULAR BIOL	26.0204 - Molecular Biology
ANTH 2414	BIOLOGICAL ANTH	45.0202 - Physical and Biological Anthropology
BIOL 1333	INTRO BIOL	26.0101 - Biology/Biological Sciences General
ANT 5073	ADV. BIOLOGICAL ANTHROPOLOGY	45.0202 - Physical and Biological Anthropology

Table 2
Thirty College Course Map Codes with Largest Enrollments

CCM Code	CCM Title	Enrollments	Prop. Of Total Enrollments
23.1301	Writing, General	87,670	0.033
42.0101	Psychology, General	52,290	0.020
40.0501	Chemistry, General	49,240	0.019
26.0101	Biology/Biological Sciences, General	46,800	0.018
54.0102	American History United States	42,060	0.016
16.0905	Spanish Language and Literature	39,420	0.015
27.9995	Calculus I, Calculus II, Calculus III, Calculus IV, Calculus for Life Science, Calculus for Economics, Calculus for Business, Calculus for Technology, Applied Calculus, Calculus for Decision-Making, Survey of Calculus and/or Short-Course Calculus	37,060	0.014
45.1101	Sociology	37,050	0.014
40.0801	Physics, General	34,720	0.013
52.0301	Accounting	33,180	0.013
27.0197	Intermediate Algebra, Pre-Collegiate Algebra, Elementary Algebra, Basic Algebra, Preparatory Algebra and/or Pre-Algebra Math	32,140	0.012
27.0102	Mathematics, General	29,800	0.011
45.1002	American Government and Politics (United States)	27,530	0.010
26.9996	Anatomy and Physiology, Applied Anatomy and/or Applied Physiology (Service Courses)	24,490	0.009
31.0501	Health and Physical Education/Fitness, General	23,940	0.009
90.9997	College orientation, freshman orientation and/or orientation and study skills for college	23,610	0.009
27.0501	Statistics, General.	21,020	0.008
40.0504	Organic Chemistry	20,270	0.008
23.1401	General Literature	19,320	0.007
09.0101	Speech Communication and Rhetoric	17,350	0.007
09.0100	Communication, General	16,700	0.006
09.0196	Public Speaking, Debate and/or Forensics	15,990	0.006
54.0196	World civilization/world history and/or modern world	15,140	0.006
52.1401	Marketing/Marketing Management, General	14,930	0.006
50.0903	Music Performance, General	14,780	0.006
45.0696	Macroeconomics, aggregate economic analysis, income and employment, growth theory, macroeconomic theory, macroeconomic analysis, income analysis, income policy, income and business cycles, business fluctuations, national income and/or national economy	14,460	0.006

Table 2 Cont.

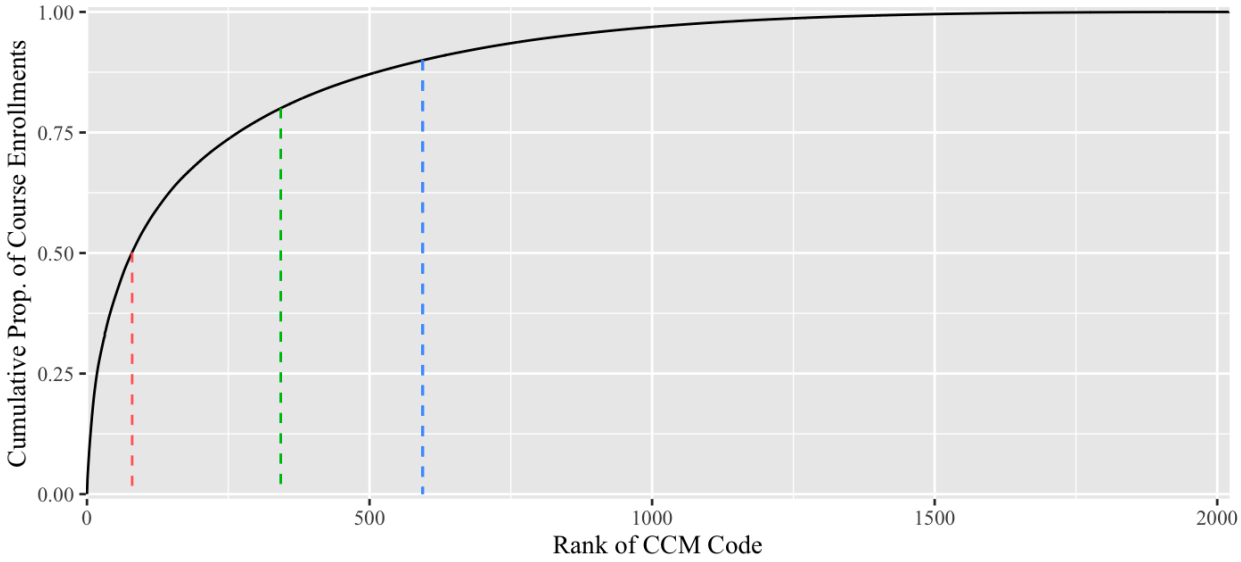
CCM Code	CCM Title	Enrollments	Prop. Of Total Enrollments
45.0697	Microeconomics, microeconomic theory, microeconomic analysis, price theory, theory of demand, theory of cost, theory of the firm, economic analysis of the firm, production economics and/or production prices	14,440	0.006
38.0101	Philosophy	14,350	0.005
45.0601	Economics, General	13,950	0.005
50.0703	Art History, Criticism and Conservation.	13,300	0.005

Notes: Table reports the 30 CCM codes that appear the most frequently in the four combined PETS samples used in the study. Numbers are rounded to the nearest ten to preserve confidentiality of NCES restricted data.

Source: Authors' calculations from U.S. Department of Education, National Center for Education Statistics, High School and Beyond Longitudinal Study of 2009 Seniors, "Postsecondary Education Transcript Study", 04/09 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", 12/17 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", and 2008/09 Baccalaureate and Beyond Longitudinal Study, "Postsecondary Education Transcript Study."

Figure 1

Cumulative Proportion of Course Enrollments by Rank of CCM Codes



Notes: Authors' calculation of cumulative proportion of course enrollments by rank of CCM Code.

Source: Authors calculations from U.S. Department of Education, National Center for Education Statistics, High School and Beyond Longitudinal Study of 2009 Seniors, "Postsecondary Education Transcript Study", 04/09 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", 12/17 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", and 2008/09 Baccalaureate and Beyond Longitudinal Study, "Postsecondary Education Transcript Study."

Table 3
Example Bag of Words Matrix

Course Title	lab	exprmnt	bio	cell	&	mol	intro	to	molecular	biol	biological	anth
LAB EXPRMNT BIO-CELL & MOL BIO	1	1	2	1	1	1	0	0	0	0	0	0
INTRO TO MOLECULAR BIOL	0	0	0	0	0	0	1	1	1	1	0	0
BIOLOGICAL ANTH	0	0	0	0	0	0	0	1	0	0	1	1
INTRO BIOL	0	0	0	0	0	0	1	0	1	1	0	0

Table 4

Accuracy of Models Predicting Two-, Four-, and Six-Digit CCM Codes

CCM Level	Train Accuracy on Courses	Test Accuracy on Courses	Train Accuracy on Enrollments	Test Accuracy on Enrollments
Two Digit CCM Code	0.817	0.808	0.874	0.874
Four Digit CCM Code	0.723	0.703	0.803	0.795
Six Digit CCM Code	0.696	0.603	0.768	0.719
Two Digit CCM Code Derived from Four Digit CCM Code	0.817	0.804	0.876	0.874
Two Digit CCM Code Derived from Six Digit CCM Code	0.857	0.808	0.900	0.876
Four Digit CCM Code Derived from Six Digit CCM Code	0.777	0.707	0.837	0.799

Notes: Trained multinomial logistic regression models classification accuracy on unweighted and enrollment weighted train and test datasets.

Source: Authors calculations of accuracy and enrollment weighted accuracy of predicted CCM codes from U.S. Department of Education, National Center for Education Statistics, High School and Beyond Longitudinal Study of 2009 Seniors, "Postsecondary Education Transcript Study", 04/09 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", 12/17 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", and 2008/09 Baccalaureate and Beyond Longitudinal Study, “Postsecondary Education Transcript Study”

Table 5
Example Course Enrollment Records from CBII with Predicted CCM Codes and Disciplinary Groupings

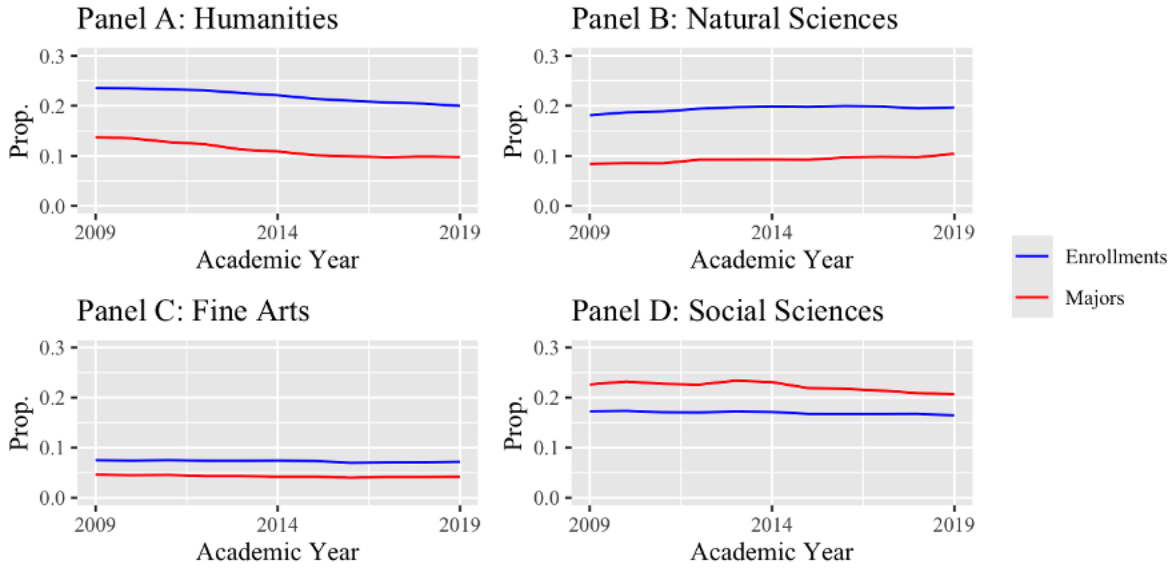
Subject Number	Course Title	Predicted Four-Digit CCM Code	Disciplinary Grouping
CHEM 241	Intro to Chem Analy	40.05 – Chemistry	Natural Sciences
ECON 101	Principles Econ I	45.06 – Economics	Social and Behavioral Sciences
CLCIV 101	Ancient Greek World	54.01 – History	Humanities
ENS 349	Univ Choir	50.09 – Music	Fine and Performing Arts
ACC 271	Prin Acctg I	52.03 – Accounting	Business and Management
EDUC 461	Teach Curr Material	13.03 – Curriculum & Instruction	Education
NURS 220	Women's Hlth	51.38 - Nursing	Health and Medical
EECS 477	Intro to Algorithms	11.01 – Computer & Information Sciences	Engineering
ITEC 4357	Digital Forensics	43.01 – Criminal Justice	Other/Unknown

Table 6
Example Major Records with CIP Codes and Disciplinary Groupings

Major Name	Four-Digit CIP Code	Disciplinary Grouping
Biology	26.01 – Biology, General	Natural Sciences
Psychology	42.01 – Psychology, General	Social and Behavioral Sciences
English	23.01 – English Language and Literature, General	Humanities
Art, Studio	50.07 – Fine and Studio Arts	Fine and Performing Arts
Business Management	52.02 – Business Management, Administration, and Operations	Business and Management
Elementary Education	13.12 – Teacher Education and Professional Development, Specific Levels and Methods	Education
Nursing	51.38 – Registered Nursing, Nursing Administration, Nursing Research and Clinical Nursing	Health and Medical
Computer Science	11.01 – Computer and Information Sciences, General	Engineering
Criminal Justice	43.01 – Criminal Justice and Corrections	Other/Unknown

Figure 2

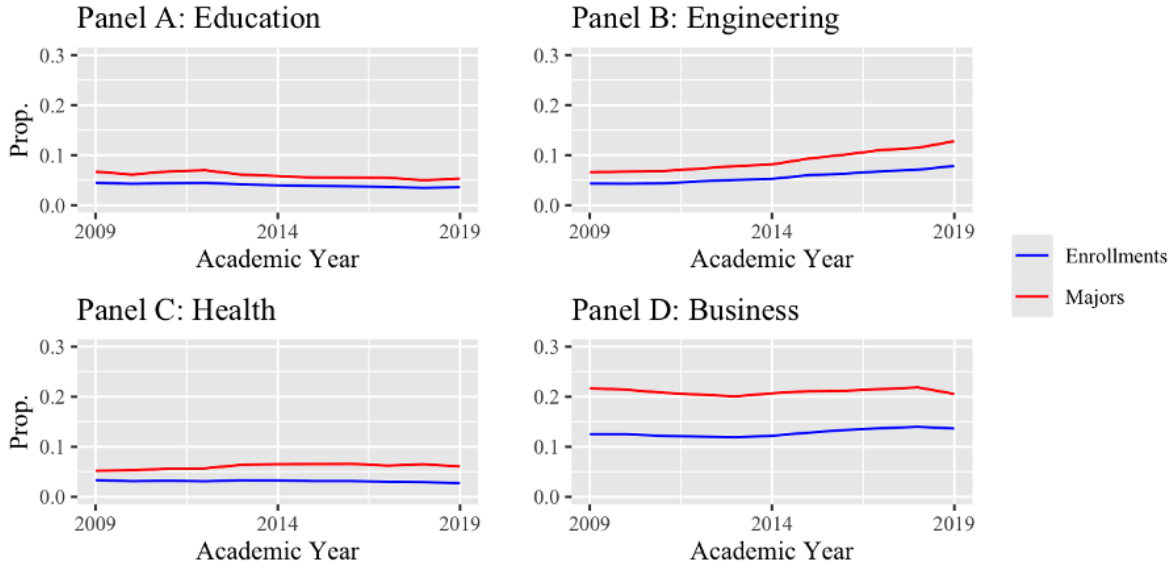
Longitudinal Trends in Liberal Arts Majors and Course-taking



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of course-taking and majors in the Humanities over time, Panel B displays the share in the Natural Sciences, Panel C displays the share in the Fine and Performing Arts, and Panel D displays the share in the Social and Behavioral Sciences. Disciplinary definitions are taken from the American Academy of Arts and Sciences Humanities Indicators Project.

Figure 3

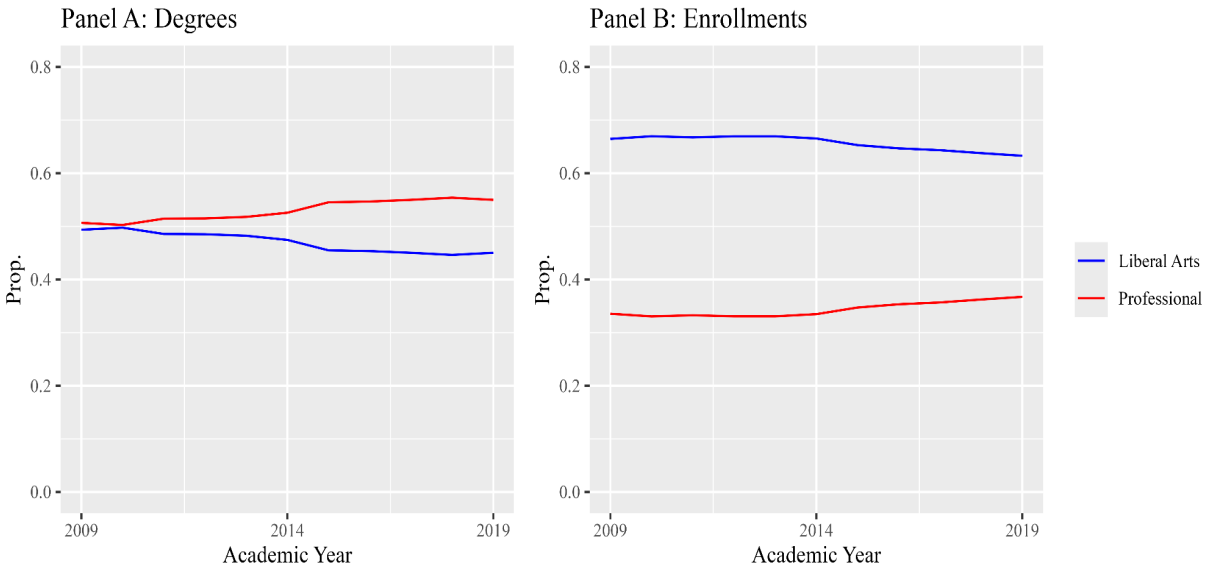
Longitudinal Trends in Professional Majors and Course-taking



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of course-taking and majors in the education over time, Panel B displays the share in engineering, Panel C displays the share in health and medical fields, and Panel D displays the share in business and management. Disciplinary definitions are taken from the American Academy of Arts and Sciences Humanities Indicators Project.

Figure 4

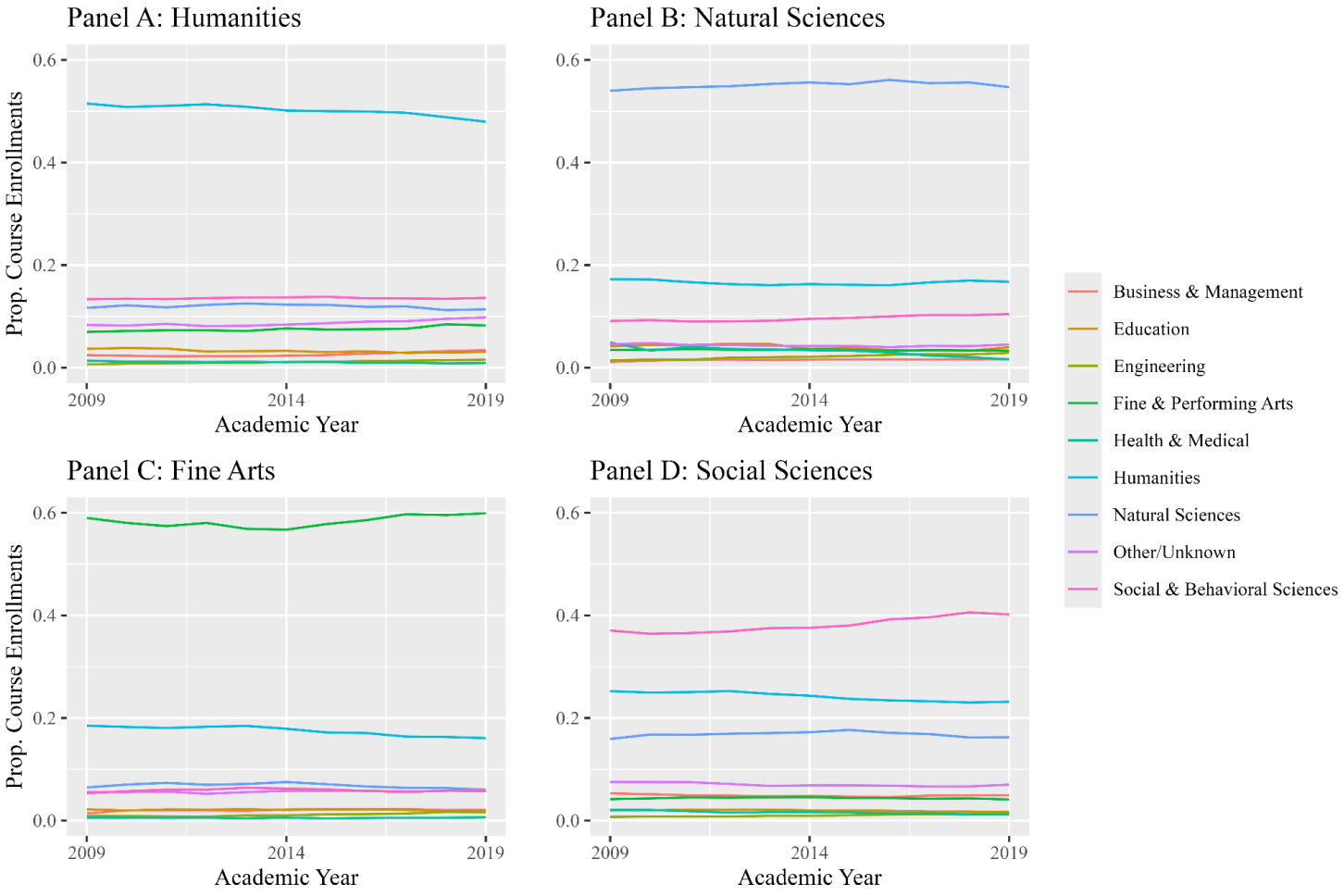
Longitudinal Trends Liberal Arts and Professional Disciplines Majors and Coursework



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of course-taking and majors in the liberal arts disciplines over time and Panel B displays the share in professional disciplines. Liberal arts disciplines include the humanities, social and behavioral sciences, the natural sciences, and the fine and performing arts. Professional disciplines include education, engineering, health and medical fields, and business and management. Disciplinary definitions are taken from the American Academy of Arts and Sciences Humanities Indicators Project.

Figure 5

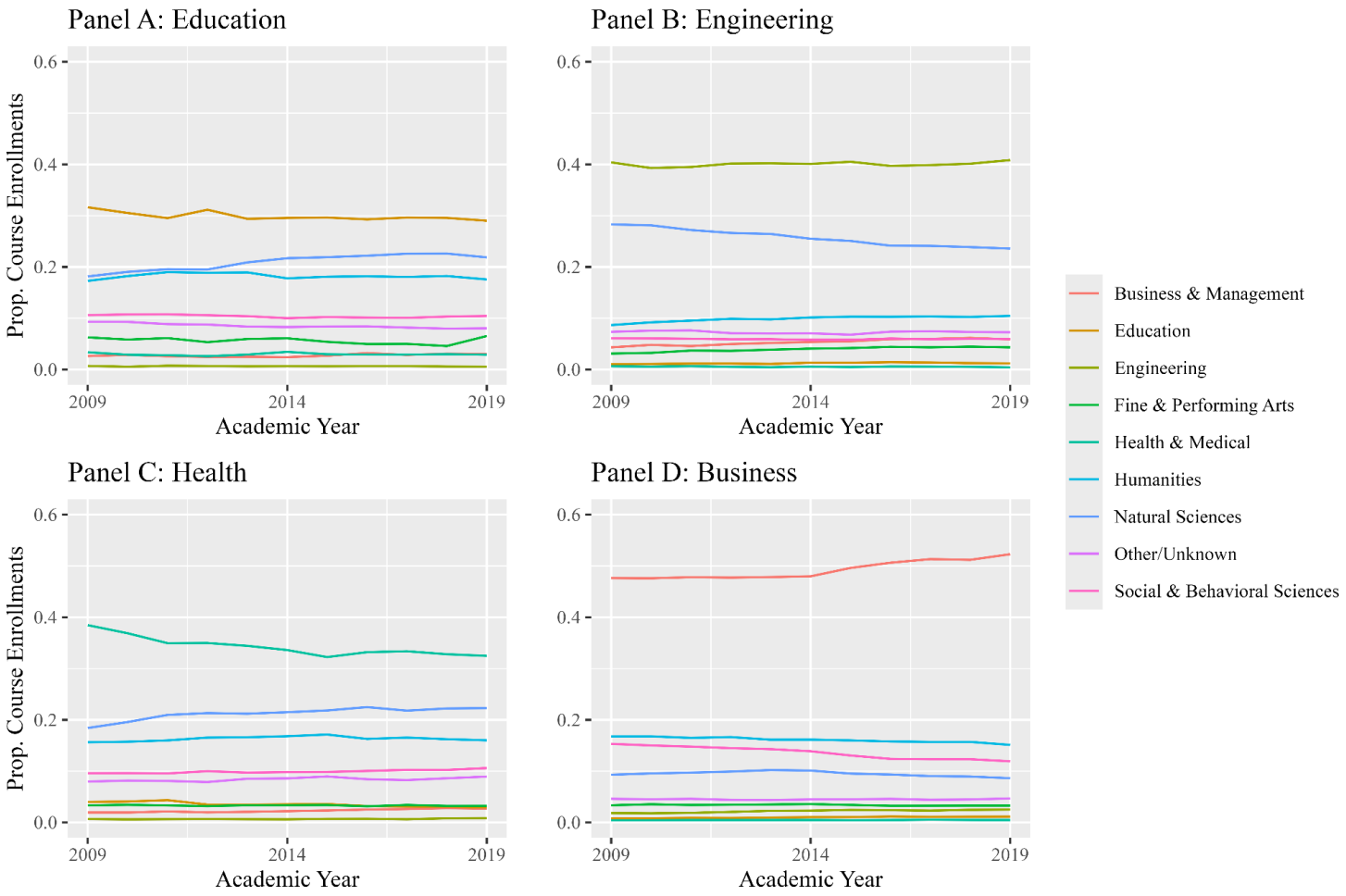
Longitudinal Trends in Disciplinary Course-taking by Major, Liberal Arts



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of course-taking in nine disciplines among students that major in the humanities over time, Panel B displays the share in the natural sciences, Panel C displays the share in the fine and performing arts, and Panel D displays the share in social and behavioral sciences. Disciplinary definitions are taken from the American Academy of Arts and Sciences Humanities Indicators Project.

Figure 6

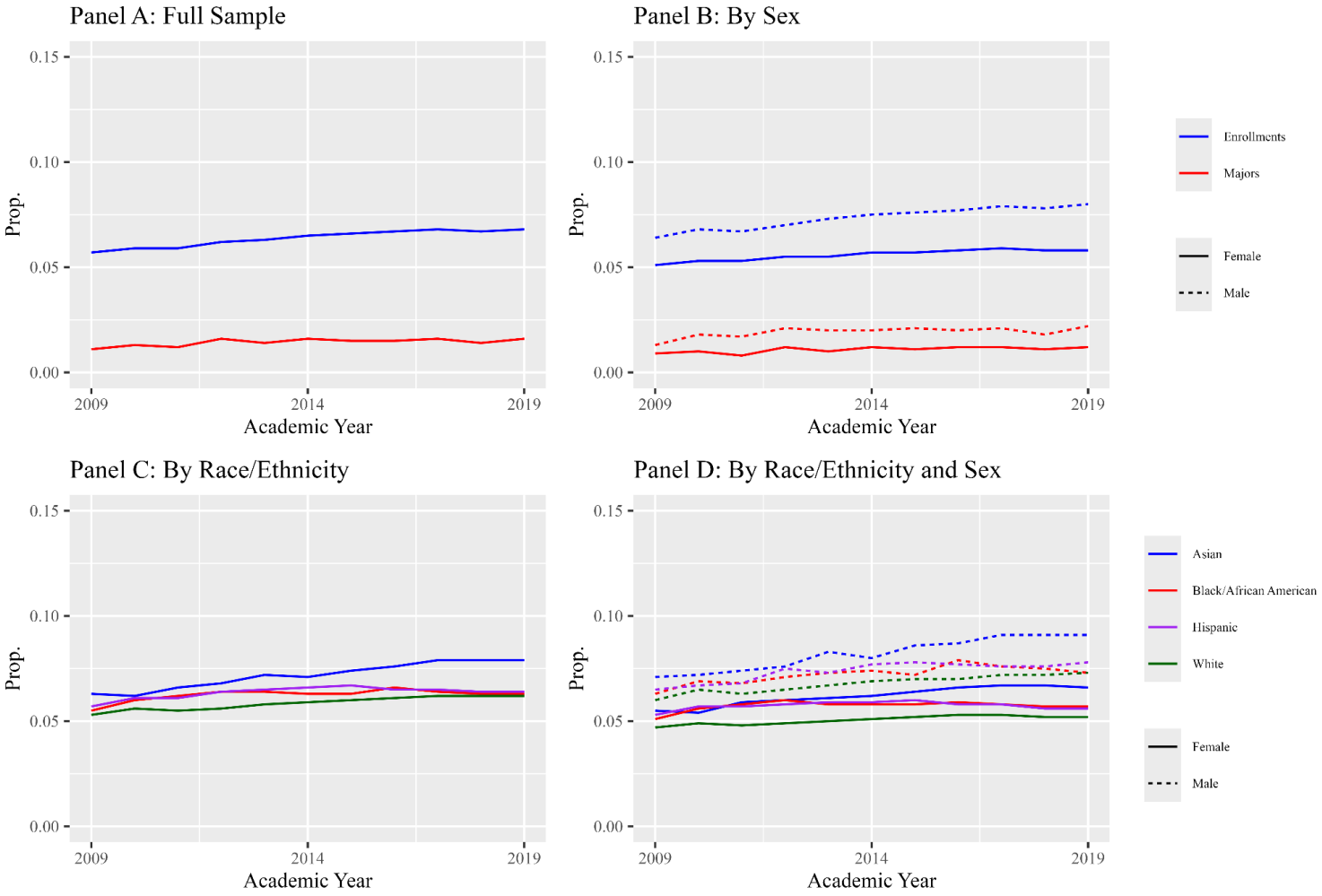
Longitudinal Trends in Disciplinary Course-taking by Major, Professional Disciplines



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of course-taking in nine disciplines among students that major in education over time, Panel B displays the share of course-taking in engineering, Panel C displays the share of course-taking in health and medical fields, and Panel D displays the share of course-taking in business and management. Disciplinary definitions are taken from the American Academy of Arts and Sciences Humanities Indicators Project.

Figure 7

Longitudinal Trends in Mathematics Course-Taking by Race/Ethnicity and Sex



Notes: Sample includes graduates of 18 institutions between 2009 and 2019. Panel A displays the proportion of enrollments and degrees in mathematics in the full sample and Panel B displays the proportion of enrollments and degrees in mathematics by sex. Panel C and Panel D display the proportion of enrollments in mathematics coursework by race/ethnicity and by the intersection of race/ethnicity and sex respectively.

Table A1

Model Classification of Enrollment-Weighted Accuracy by IPEDS Sector

	Number of Unique Courses	Two Digit Accuracy	Four Digit Accuracy	Six Digit Accuracy
Private for-profit, 2-year	4,400	0.826	0.671	0.627
Private for-profit, 4-year or above	12,010	0.871	0.791	0.689
Private for-profit, less-than 2-year	1,890	0.828	0.729	0.639
Private not-for-profit, 2-year	1,510	0.842	0.781	0.704
Private not-for-profit, 4-year or above	72,630	0.858	0.772	0.697
Private not-for-profit, less-than 2-year	210	0.783	0.642	0.491
Public, 2-year	48,800	0.910	0.842	0.774
Public, 4-year or above	118,640	0.878	0.798	0.721
Public, less-than 2-year	770	0.744	0.681	0.616
Sector Unknown in IPEDS	50	0.771	0.625	0.521
Did Not Match with IPEDS	1370	0.793	0.730	0.710

Notes: We do not report accuracy on a small number of courses whose institutional sector is "Administrative Unit." Numbers are rounded to the nearest ten to preserve confidentiality of NCES restricted data.

Source: Authors calculations of accuracy and enrollment weighted accuracy of predicted CCM codes by IPEDS Sector from U.S. Department of Education, National Center for Education Statistics, High School and Beyond Longitudinal Study of 2009 Seniors, "Postsecondary Education Transcript Study", 04/09 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", 12/17 Beginning Postsecondary Students Longitudinal Study, "Postsecondary Education Transcript Study", and 2008/09 Baccalaureate and Beyond Longitudinal Study, "Postsecondary Education Transcript Study"