Technology and Big Data Are Changing Economics: Mining Text to Track Methods

Janet Currie, Princeton University and NBER

Henrik Kleven, Princeton University and NBER

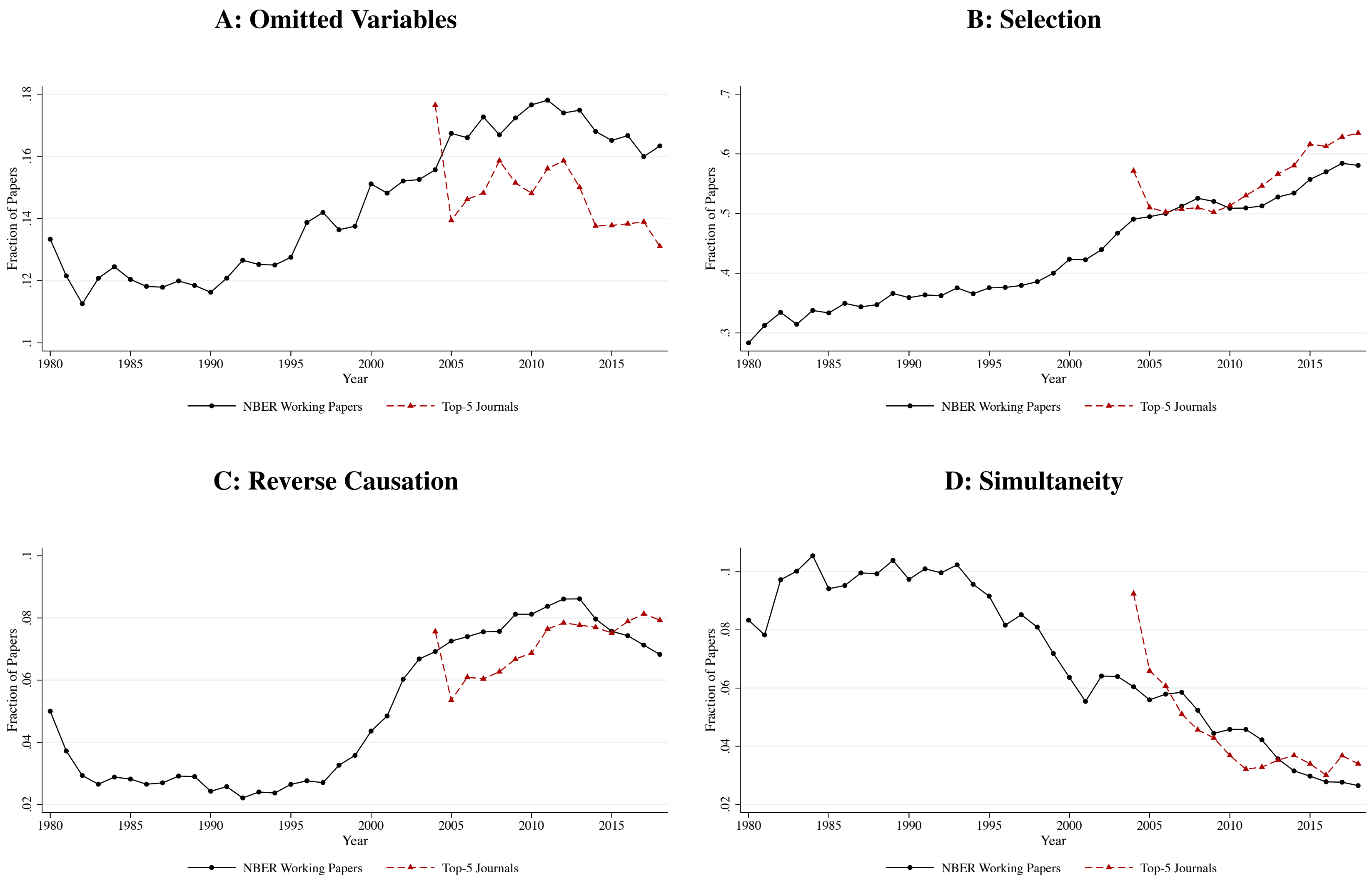
Esmée Zwiers, Princeton University

January 2020

ONLINE APPENDIX

A Supplementary Figures and Tables

Figure A.I: Identification Concerns



A: Omitted Variables

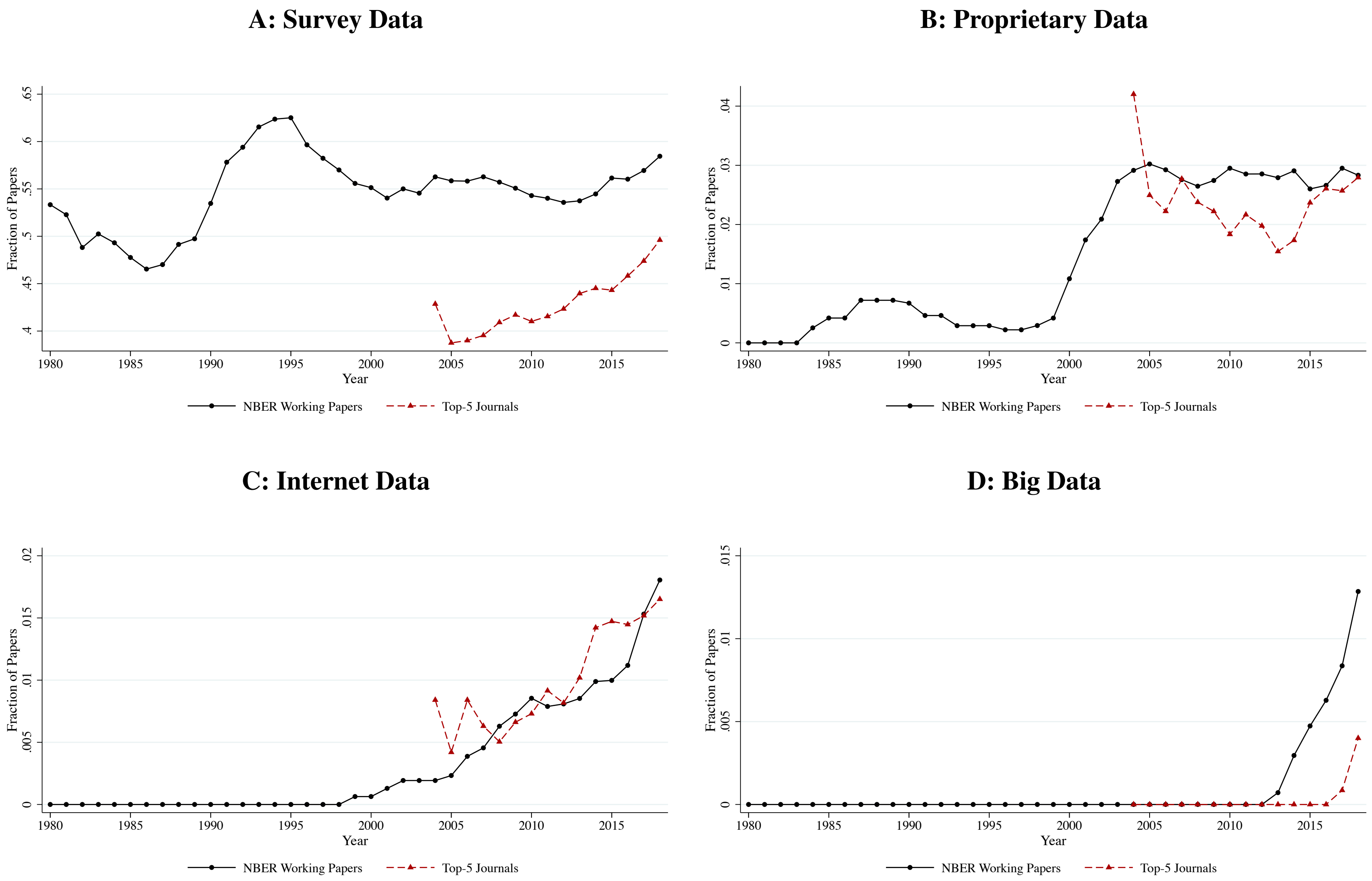
B: Selection

C: Reverse Causation

D: Simultaneity

Notes: This figure shows the fraction of papers referring to each term. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.II: Data



A: Survey Data

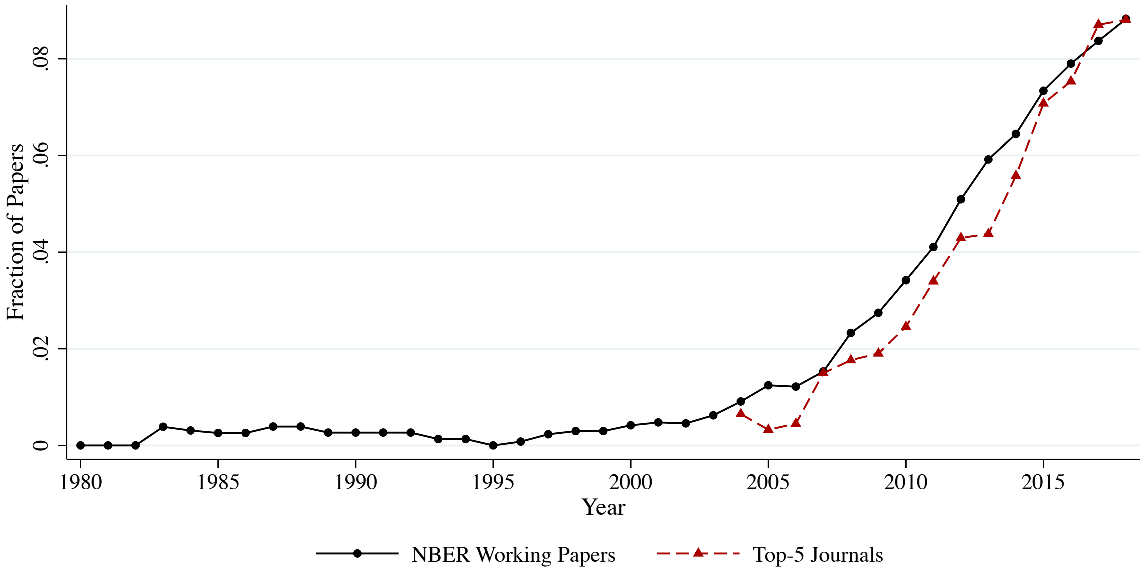
B: Proprietary Data

C: Internet Data

D: Big Data

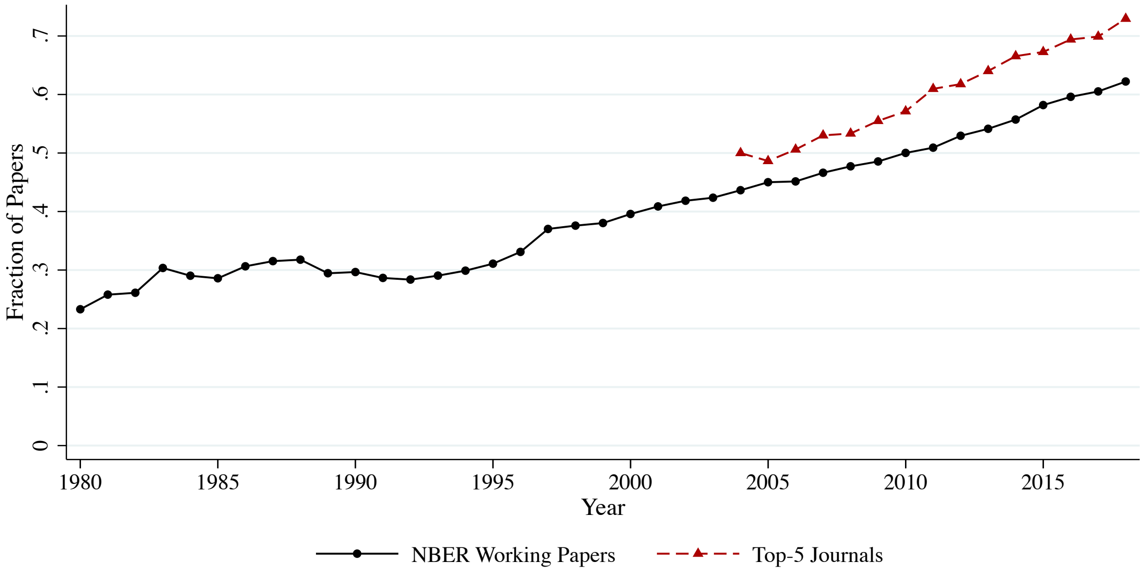
Notes: This figure shows the fraction of papers referring to each type of data. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.III: External Validity



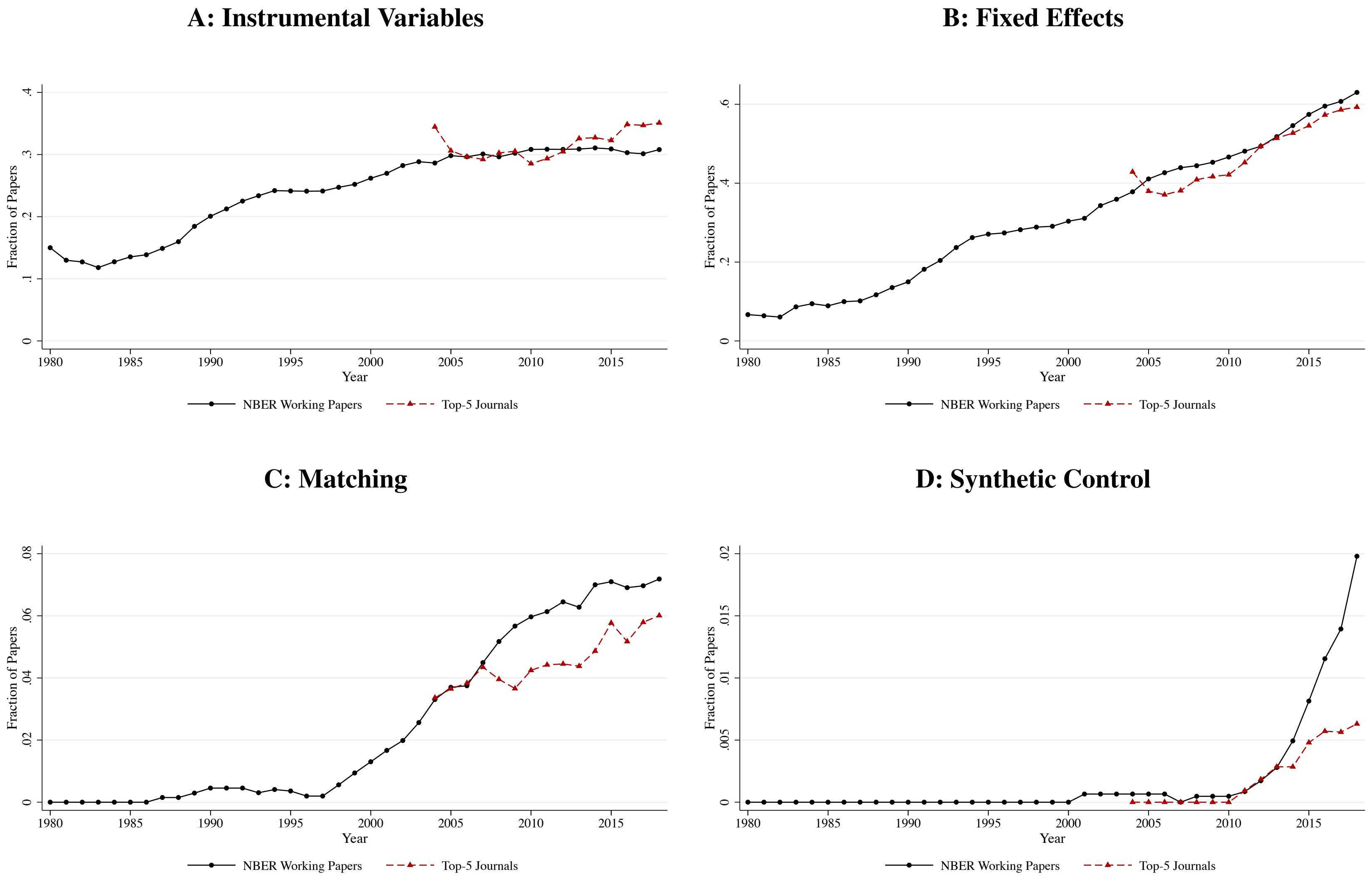
Notes: This figure shows the fraction of papers referring to external validity. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.IV: Mechanisms



Notes: This figure shows the fraction of papers referring to mechanisms. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.V: Other Empirical Methods



A: Instrumental Variables

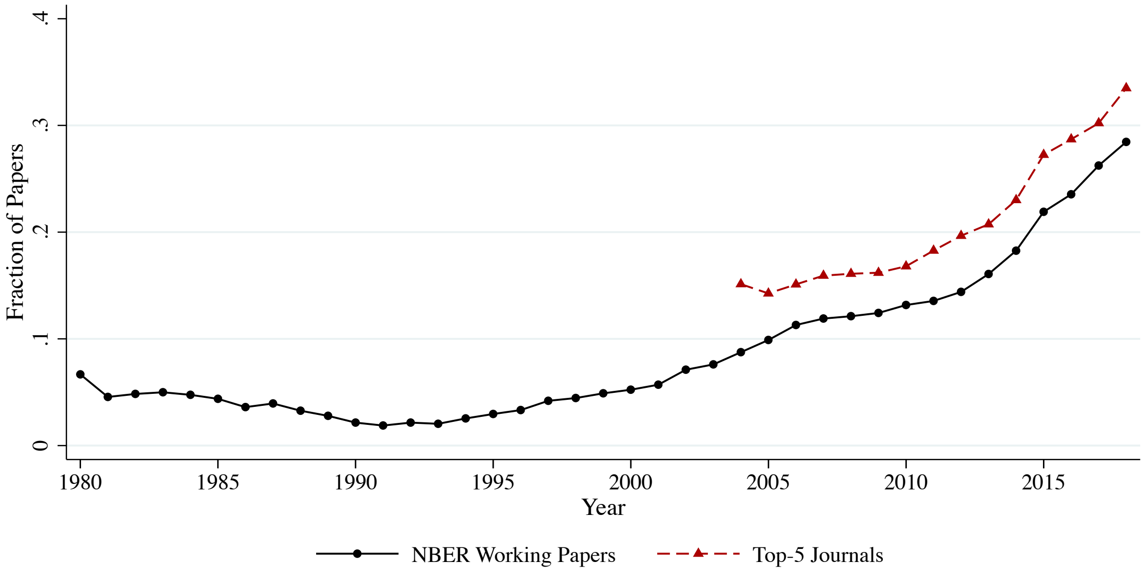
B: Fixed Effects

C: Matching

D: Synthetic Control

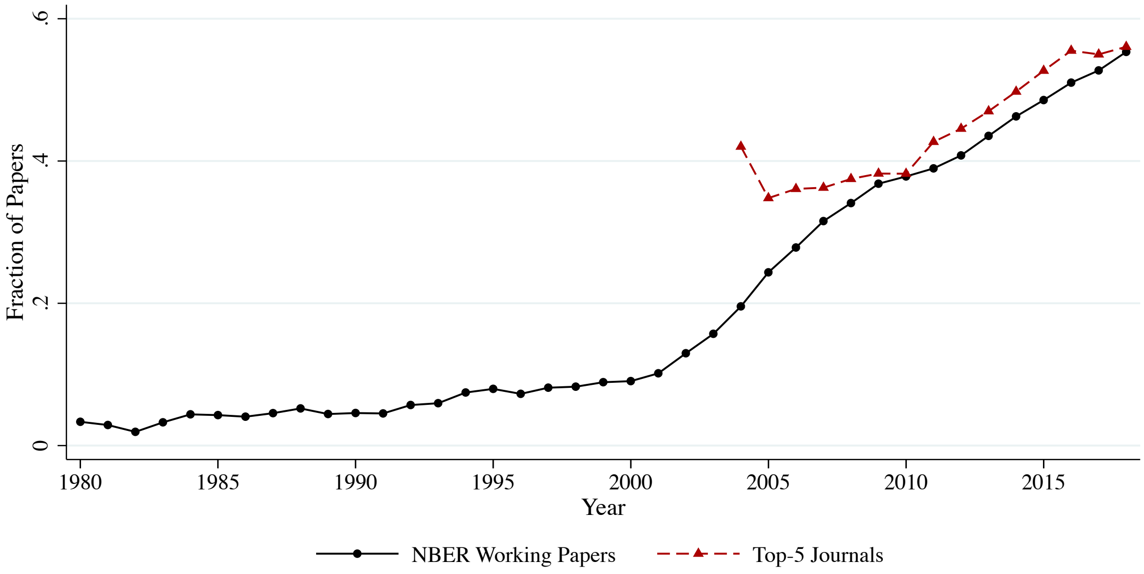
Notes: This figure shows the fraction of papers referring to each method. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.VI: Confidence Interval



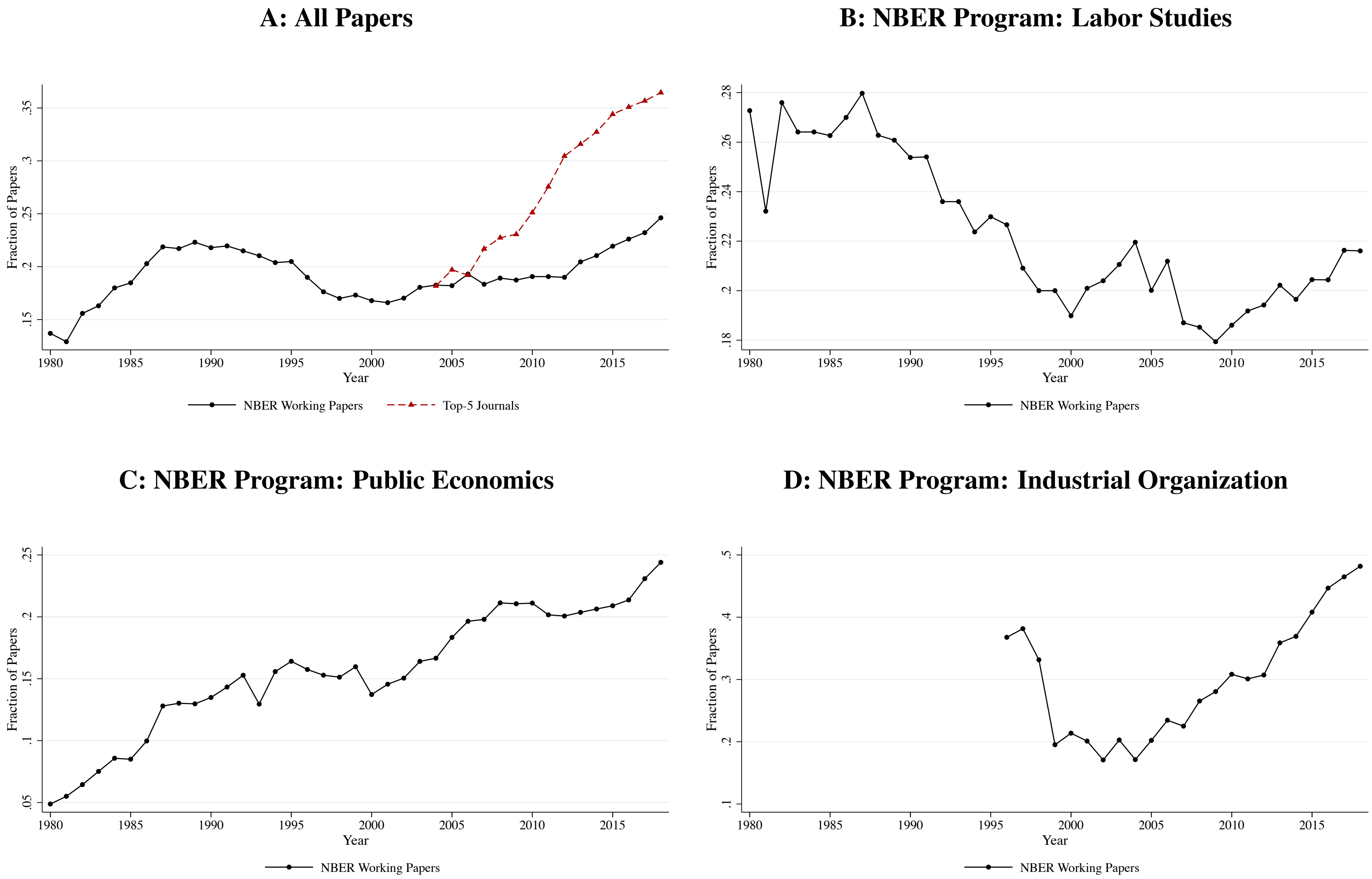
Notes: This figure shows the fraction of papers referring to confidence intervals. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.VII: Clustering



Notes: This figure shows the fraction of papers referring to clustering. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.VIII: Structural Methods: Structural Model


A: All Papers

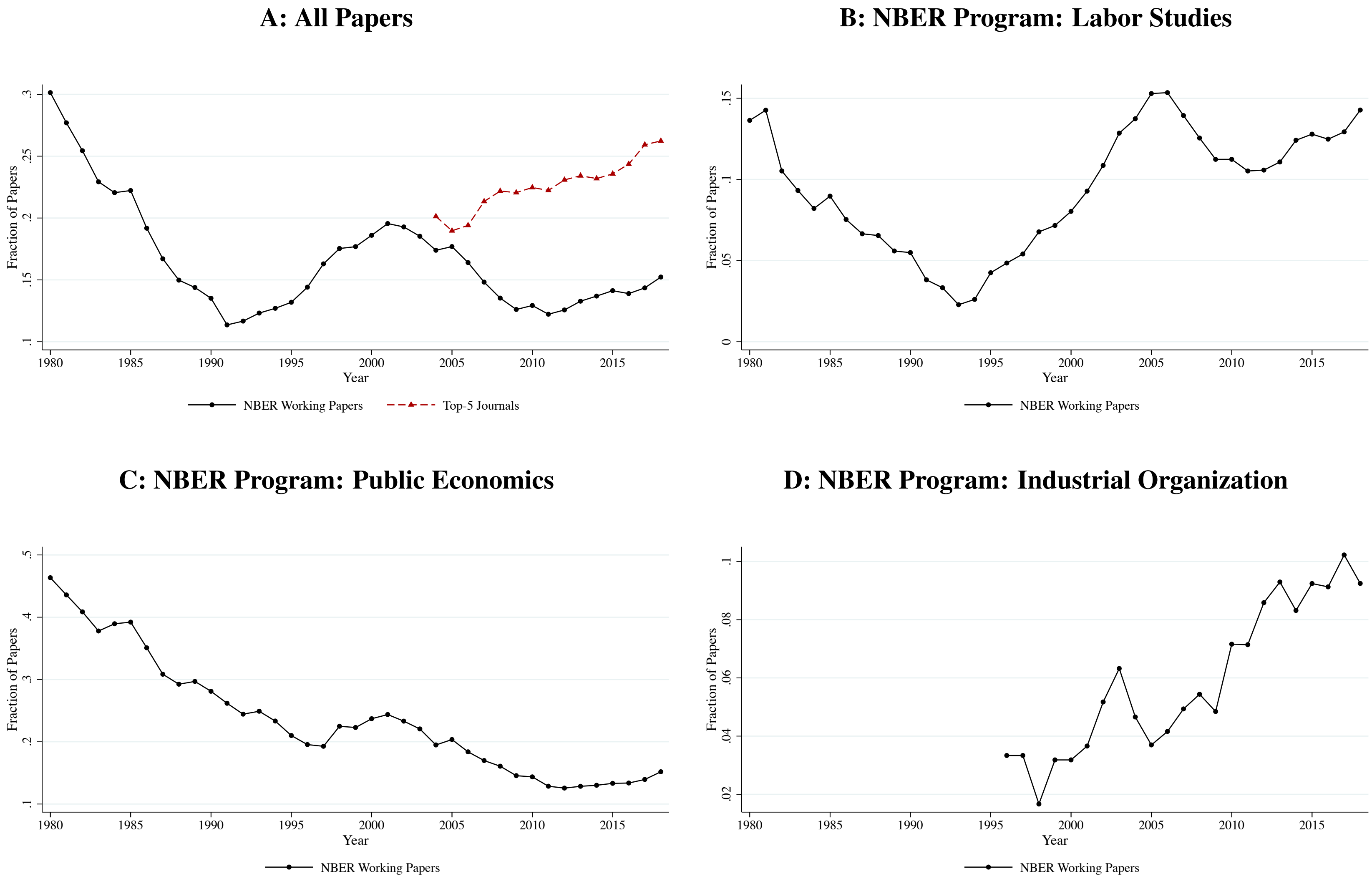
B: NBER Program: Labor Studies

C: NBER Program: Public Economics

D: NBER Program: Industrial Organization

Notes: This figure shows the fraction of papers referring to structural methods. Panel A shows all papers in applied microeconomics, while Panels B-D focus on NBER working papers within specific programs (LS, PE, and IO). The IO series omit the first 5 data points, because of the low number of papers in the early years of the program. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.IX: Structural Methods: General Equilibrium



A: All Papers

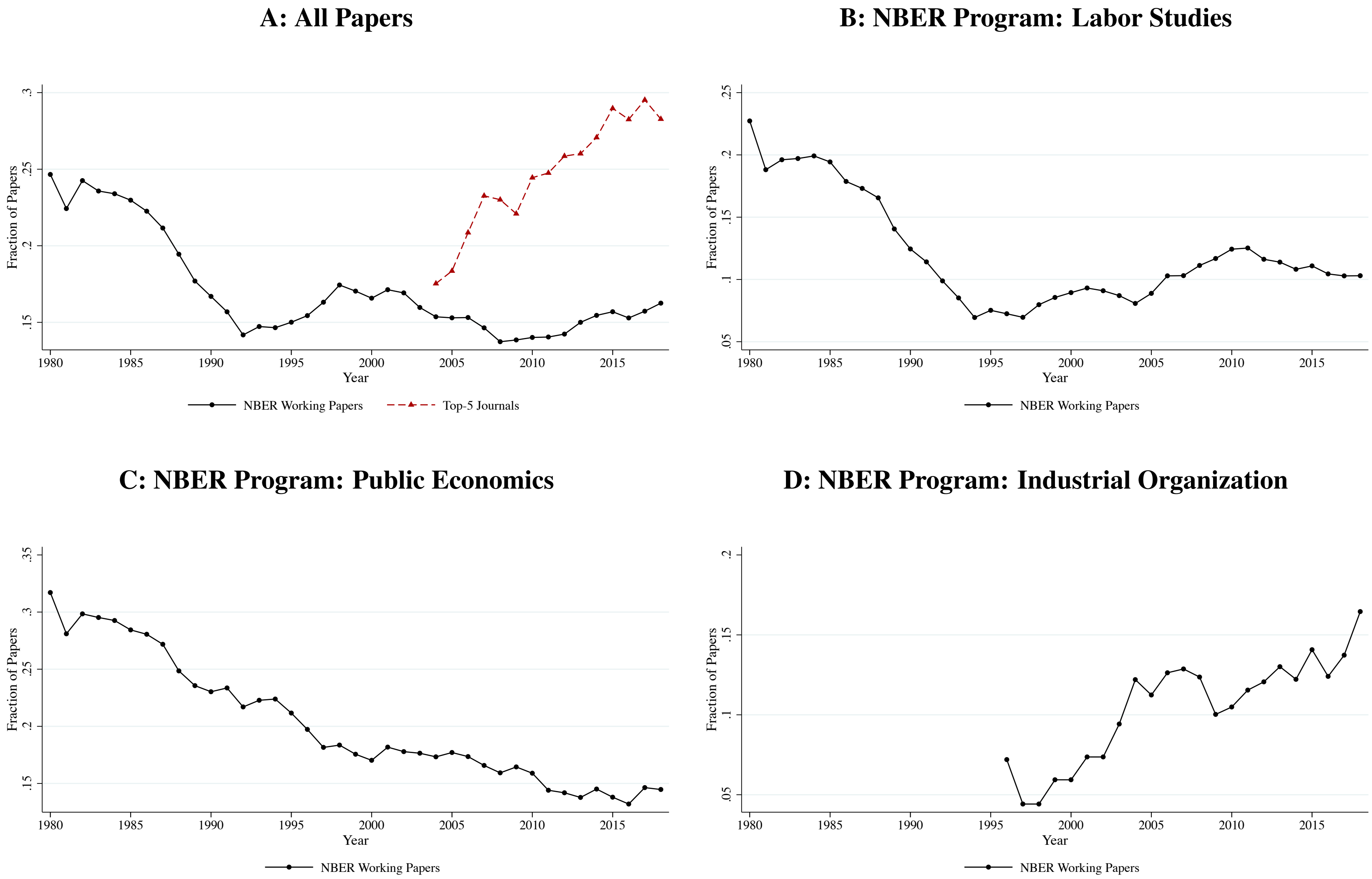
B: NBER Program: Labor Studies

C: NBER Program: Public Economics

D: NBER Program: Industrial Organization

Notes: This figure shows the fraction of papers referring to general equilibrium. Panel A shows all papers in applied microeconomics, while Panels B-D focus on NBER working papers within specific programs (LS, PE, and IO). The IO series omit the first 5 data points, because of the low number of papers in the early years of the program. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.X: Structural Methods: Functional Forms



Notes: This figure shows the fraction of papers referring to functional forms. Panel A shows all papers in applied microeconomics, while Panels B-D focus on NBER working papers within specific programs (LS, PE, and IO). The IO series omit the first 5 data points, because of the low number of papers in the early years of the program. See Table A.I for a list of terms. The series show 5-year moving averages.

Table A.I: Search Categories and Trigger Phrases

Category	**Trigger Phrases**	**Outcome**	**Case Sensitive**	**Wildcard at end**	**Cond. on ’data’**
Administrative Data	‘administrative data’, ‘admin data’, ‘administrative-data’, ‘admin-data’, ‘administrative record’, ‘admin record’, administrative regist’, ‘admin regist’, ‘register data’, ’registry data’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Big Data	‘big data’, ‘big-data’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Binscatter	‘binscatter’, ‘bin scatter’, ‘binned scatter’	Fraction of papers with at least 1 phrase	No	Yes	No
Bunching	‘bunching’	Fraction of papers with at least 1 phrase	No	Yes	No
Clustering	‘cluster’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Confidence Interval	’confidence interval’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Data	’data’	Fraction of papers with at least 1 phrase	No	Yes	No
Difference-in-Differences	‘Difference in Diff’, ‘Difference in diff’, ‘difference in diff’, ‘Difference-in-Diff’, ‘Difference-in-diff’, ‘difference-in-diff’, ‘Differences in Diff’, ‘Differences in diff’, ‘differences in diff’, ‘Differences-in-Diff’, ‘Differences-in-diff’, ‘differences-in-diff’, ‘diff-in-diff’, ‘d-in-d’, ‘DiD’	Fraction of papers with at least 1 phrase	Yes	Yes	No
Event Study	‘event stud’ ‘ event-stud’	Fraction of papers with at least 1 phrase	No	Yes	No
External Validity	‘external validity’, ‘external-validity’, ‘externally valid’, ‘externally-valid’	Fraction of papers with at least 1 phrase	No	Yes	No
Figure	‘graph’, ‘figure’, ‘plot’, ‘chart’	Average word count per paper	No	Yes	No

Fixed Effects	‘FE’, ‘Fixed Effect’, ‘Fixed effect’, ‘fixed effect’, Fixed Effects’, ‘Fixed effects’, ‘fixed effects’, ‘Fixed-Effect’, ‘Fixed-effect’, ‘fixed-effect’, ‘Fixed-Effects’, Fixed-effects’, ‘fixed-effects’	Fraction of papers with at least 1 phrase	Yes	No	Yes
Functional Forms	‘CES’, ‘constant elasticity of substitution’, ‘Constant Elasticity of Substitution’, ‘Constant elasticity of substitution’, ‘Cobb-Douglas’, ‘Cobb Douglas’, ‘Stone Geary’, ‘Stone-Geary’, ‘CRRA’, ‘coefficient of relative risk-aversion’, ‘coefficient of relative risk aversion’, ‘Coefficient of relative risk-aversion’, ‘Coefficient of relative risk aversion’, ‘Coefficient of Relative Risk-Aversion’, ‘Coefficient of Relative Risk Aversion’, ‘CARA’, ‘constant absolute risk aversion’, ‘constant absolute risk-aversion’, ‘Constant absolute risk aversion’, ‘Constant absolute risk-aversion’, ‘Constant Absolute Risk Aversion’, ‘Constant Absolute Risk-Aversion’, ‘translog’, ‘Translog’	Fraction of papers with at least 1 phrase	Yes	No	No
General Equilibrium	‘general equilibr’, ‘general-equilibr’	Fraction of papers with at least 1 phrase	No	Yes	No
Identification	**Sentence structure:** search for sentences that have the term ’identif’ in combination with any of the terms: ’effect’, ’response’, ’impact’, ’elasticit’, ’parameter’, or ’coefficient’ with maximum two words in between. Note that even though the search includes wildcards at the end, we exclude any match with the word ’effective’. **Also search for these terms:** ’causal identification’, ’causally identified’, ’identification strategy’, ’identification approach’, ’identification assumption’, ’identifying assumption’, ’identifying variation’, ’empirical identification’, ’over identified’, ’over-identified’, ’under identified’, ’under-identified’, ’identification properties’, ’identification test’, ’identification problem’,	Fraction of papers with at least 1 phrase	No	Yes	No

	’identification issue’, ’problem with identification’, ’problems with identification’, ’issue with identification’, ’issues with identification’, ’problem identifying’, ’problems identifying’, ’issue identifying’,’issues identifying’, ’threat to identification’, ’threats to identification’, ’threat for identification’, ’threats for identification’, ’over identifying’, ’over-identifying’, ’under identifying’, ’under-identifying’, ’partial identification’, ’partially identified’, ’non-parametric identification’, ’nonparametric identification’, ’non parametric identification’, ’non-parametrically identified’, ’nonparametrically identified’, ’non parametrically identified’, ’identification condition’, ’identifying condition’, ’condition for identification’, ’conditions for identification’, ’condition for identifying’, ’conditions for identifying’, ’point identification’, ’point-identification’, ’point identified’, ’point-identified’, ’point identifying’, ’point-identifying’, ’set identification’, ’set-identification’, ’set identified’, ’set-identified’, ’set identifying’, ’set-identifying’, ’identification analysis’, ’weak identification’, ’identification result’, ’identification argument’, ’identification framework’, ’identification scheme’				
Internet Data	‘internet data’, ‘internet-data’, ‘web data’, ‘web-data’, ‘scraped data’, ‘scraped-data’, ‘scrape data’, ‘scraping data’, ‘search data’, ‘search-data’, ‘google data’, google-data’, ‘social media data’, ‘google trend’, ‘google-trend’, ‘google search’, ‘google-search’, ‘google ngram’, ‘google n-gram’, ‘google books ngram’, ‘google books n-gram’	Fraction of papers with at least 1 phrase	No	Yes	Yes

Instrumental Variables	‘Instrumental Variable’, ‘Instrumental variable’, ‘instrumental variable’, ‘Instrumental-Variable’, ‘Instrumental-variable’, ‘instrumental-variable’, ‘Two Stage Least Squares’, ‘Two stage least squares’, ‘two stage least squares’, ‘2SLS’, ‘TSLS’, ‘valid instrument’, ‘exogenous instrument’, ‘IV Estimat’, ‘IV estimat’, ‘IV-estimat’, ‘IV Specification’, ‘IV specification’, ‘IV-specification’, ‘IV Regression’, ‘IV regression’, ‘IV-regression’, ‘IV Strateg’, ‘IV strateg’, ‘IV-strateg’, ‘we instrument’, ‘I instrument’, ‘paper instruments’, ‘exclusion restriction’, ‘weak first stage’, ‘simulated instrument’	Fraction of papers with at least 1 phrase	Yes	Yes	Yes
Lab Experiments	‘Laboratory Experiment’, ‘Laboratory experiment’, ‘laboratory experiment’, ‘Lab Experiment’, ‘Lab experiment’, ‘lab experiment’, ’Dictator Game’, ‘Dictator game’, ‘dictator game’, ’Ultimatum Game’, ‘Ultimatum game’, ‘ultimatum game’, ’Trust Game’, ‘Trust game’, ‘trust game’ , ’Public Good Game’, ’Public good game’, ‘public good game’, ’Public Goods Game’, ‘Public goods game’, ‘public goods game’, ‘Z-tree’, ‘zTree’, ‘ORSEE’, ‘show-up fee’, ‘laboratory participant’, ‘lab participant’	Fraction of papers with at least 1 phrase	Yes	Yes	No
Machine Learning	‘machine learning’, ‘lasso’, ‘random forest’	Fraction of papers with at least 1 phrase	No	Yes	No
Matching	‘propensity score’, ‘propensity score matching’, ’propensity-score matching’, ‘matching estimat’, ‘nearest neighbor matching’, ’nearest-neighbor matching’, ‘nearest neighbour matching’, ’nearest-neighbour matching’, ‘caliper matching’, ‘stratification matching’, ‘exact matching’, ‘one to one matching’, ‘one-to-one matching’, ‘kernel matching’, ‘inverse probability matching’, ’inverse-probability matching’	Fraction of papers with at least 1 phrase	No	Yes	Yes

Mechanisms	’mechanism’	Fraction of papers with at least 1 phrase	No	Yes	No
Omitted Variables	‘omitted variable’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Preanalysis Plan	‘pre-analysis plan’, ‘pre analysis plan’, ‘preanalysis plan’	Fraction of papers with at least 1 phrase	No	Yes	No
Precisely Estimated	‘precisely estimated’, ‘precisely-estimated’	Fraction of papers with at least 1 phrase	No	Yes	No
Precisely Estimated Zero	‘precisely estimated zero’, ‘precisely-estimated zero’	Fraction of papers with at least 1 phrase	No	Yes	No
Proprietary Data	‘proprietary data’, ‘confidential data’, ‘nonpublic data’, ‘non-public data’, ‘proprietary-data’, ‘confidential-data’, ‘nonpublic-data’, ‘non-public-data’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Quasi- and Natural Experiments	‘quasi experiment’, ‘quasi-experiment’, ‘quasiexperiment’, ‘natural experiment’, ‘natural-experiment’	Fraction of papers with at least 1 phrase	No	Yes	No
RCTs	‘Randomized Controlled Trial’ , ‘Randomized controlled trial’, ‘randomized controlled trial’, ‘Randomized Control Trial’, ‘Randomized control trial’, ‘randomized control trial’, ‘Randomized Field Experiment’, ‘Randomized field experiment’, ‘randomized field experiment’, ‘Randomized Controlled Experiment’, ‘Randomized controlled experiment’, ‘randomized controlled experiment’, ‘Randomised Controlled Trial’ , ‘Randomised controlled trial’, ‘randomised controlled trial’, ‘Randomised Control Trial’, ‘Randomised control trial’, ‘randomised control trial’, ‘Randomised Field Experiment’, ‘Randomised field experiment’, ‘randomised field experiment’, ‘Randomised Controlled Experiment’, ‘Randomised controlled experiment’, ‘randomised controlled experiment’, ’Social Experiment’, ’Social experiment’, ’social experiment’, ‘RCT’	Fraction of papers with at least 1 phrase	Yes	Yes	No

Regression Discontinuity	‘Regression Discontinuit’, ‘Regression discontinuit’, ‘regression discontinuit’, ’Regression-discontinuity’, ’regression-discontinuity’, ‘Regression Kink’, ‘Regression kink’, ‘regression kink’, ‘RD Design’, ‘RD design’, RD-design’, ‘RD Estimat’, ‘RD estimat’, ‘RD-estimat’, ‘RD Model’, ‘RD model’, ‘RD-model’ , ‘RD Regression’, ‘RD regression’, ‘RD-regression’, ‘RD Coefficient’, ‘RD coefficient’, ‘RD-coefficient’, ’RK Design’, ’RK design’, ’RK-Design’, ’RK-design’, ’RKD’, ’RDD’	Fraction of papers with at least 1 phrase	Yes	Yes	No
Reverse Causation	‘reverse causa’, ’reverse-causa’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Selection	‘selection’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Simultaneity	‘simultaneity’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Structural Model	**Sentence structure**: we search for instances where, within two full stops, the term ’structural’ is mentioned in combination with either ’model’, ’specification’, ’estimate’, or ’parameter’. **Also search for these terms:** ‘Structural Model’, ‘Structural model’, ‘structural model’, ‘Method of Moments’, ‘Method of moments’, ‘method of moments’, ’Method-of-Moments’, ’Method-of-moments’, ’method-of-moments’, ‘Berry, Levinsohn, Pakes’, ‘Berry, Levinsohn and Pakes’, ‘Berry, Levinsohn, and Pakes’, ‘BLP’, ‘Structural General Equilibrium Model’, ’Structural general equilibrium model’, ’structural general equilibrium model’, ‘GMM’, ’Maximum Likelihood Estimat’, ’Maximum likelihood estimat’, ’maximum likelihood estimat’, ’Maximum-Likelihood Estimat’, ’Maximum-likelihood estimat’, ’maximum-likelihood estimat’, ’MLE’	Fraction of papers with at least 1 phrase	Yes	Yes	No

Survey Data	**Sentence structure**: we search for instances where the term ‘survey’ and ’data’ are mentioned within two full stops.	Fraction of papers with at least 1 phrase	No	Yes	Yes
Synthetic Control	‘synthetic control’	Fraction of papers with at least 1 phrase	No	Yes	Yes
Table	’table’	Average word count per paper	No	Yes	No
Text Analysis	‘natural language processing’, ‘text analys’, ‘computational linguistics’, ‘speech processing’, ‘n-gram’, ‘ngram’, ‘n gram’, ‘textual analys’, ‘language processing’, ’language analys’, ‘text data’, ‘text mining’, ‘mining text’, ‘text regression’, ‘tokeniz’	Fraction of papers with at least 1 phrase	No	Yes	No

B Data and Methods Appendix

B.1 Data: Top-5 Papers

Our sample of "top-5" economics journals includes papers published in the American Economic Review (AER), Econometrica (ECMA), the Journal of Political Economy (JPE), the Quarterly Journal of Economics (QJE), and the Review of Economic Studies (ReStud). We retrieve PDFs of the papers from the respective journals' websites and focus on all published papers between 2004 and 2019. Specifically, we collect papers from the AER until its 7th issue of 2019, Econometrica until its 3rd issue of 2019, JPE until its 3rd issue of 2019, RES until its 4th issue of 2019, and QJE until its 3rd issue of 2019. The paper's JEL codes are collected directly from the paper or from the Econlit database. We disregard papers published in AER Papers & Proceedings, papers that are labeled comments, notes or replies, and special issues. Eventually we end up with a sample of 4,344 papers that cover a little over 15 years.

We focus on applied microeconomic papers, and we use the paper's JEL codes for our sample selection. Card and DellaVigna [2013] create a classification to map different JEL codes into different fields. We use their fields of Labor (J, I2), Industrial Organization (L), International (F), Public Economics (H), Health and Urban Economics (I0, I1, R, K), Development (O), and Lab experiments (C9) to define the group of applied microeconomic papers. We add Welfare, Wellbeing, and Poverty (I3), and Agriculture and Natural Resource Economics/Environmental and Ecological Economics (Q) to this classification to come our final definition of applied microeconomic papers. We select all papers that report a JEL code in either one of these JEL categories - note that papers can have JEL codes in multiple categories - and our final sample consists of 2,830 applied microeconomic papers.[1]

[1] Figure B.I shows the number of papers in our top-5 journal sample over time. Table B.I shows the distribution of papers across JEL codes before and after our selection of applied micro papers.

B.2 Data: NBER Working Papers

Our sample of NBER Working Papers consists of all working papers published through the National Bureau of Economic Research (NBER) Working Paper Series from the inception of the program in 1975 to the end of June 2018, when the data for this project were scraped from the NBER website. There are 24,449 papers in the generalized Working Paper series, 313 papers in the Technical Working Paper (TWP) Series (papers focused on econometrics and other methodological contributions), 135 papers in the Working Paper Series on Historical Factors in Long-Run Growth, and 166 Reprint Series papers, for a total of 25,063 papers across the programs.

For the analyses presented in this paper, we drop from the NBER data set all papers written before 1980. Additionally, we restrict our analysis to papers in NBER Working Paper Program categories associated with applied microeconomics. These categories are: Aging; Children; Development Economics; Education; Health Care; Health Economics; Industrial Organization; Labor Studies; Political Economy; Public Economics; International Trade; and Environment and Energy. Since papers may be tagged to more than one program, to select applied microeconomics papers, we drop all papers that are tagged to programs *other* than the ones specified here. For example, a paper that is only tagged to Public Economics would be included in our sample, but a paper that is tagged to both Public Economics and Economic Fluctuations and Growth would be excluded. Counts of papers in each category that we classify as applied microeconomics papers are detailed in Table B.II. Within the set of applied microeconomics in our final sample, we have 62 in the reprint program, 61 in the Technical Working Paper Series, and 10,201 standard working papers. This gives us a total of 10,324 papers in the NBER data set.[2]

The more recent of these papers are available in plain text format at the nber.org URL associated with the paper.[3] For papers for which a plain text version is available, we scrape the text of the paper directly from the link. For papers for which there is no plain text version available (typically older papers), we scrape the PDF version of the paper from the NBER website. Bibliographic data

[2] Figure B.I shows the number of papers in the NBER Working Papers sample over time.

[3] For example, Working Paper 25524 is available in plain text format at https://www.nber.org/papers/w25524.txt.

– including title, author(s), year and date of publication, and abstract – are available for each paper in the Series at the .bib URL associated with the paper.[4]

B.3 Data Processing

We use a series of Python scripts (files ending in .py) to process our data and count the number of times each phrase of interest is mentioned in a paper. In what follows, we will first outline how we processed our data to get from the paper PDFs to a plain text file that could be used for text mining. Afterwards we will discuss how we searched for the categories mentioned in Table A.I.

B.3.1 Converting PDFs to text

Relevant code files: `pdf2txt.py, convertPDF.py`

We obtain PDFs of each paper in the Top 5 journal data set, as well as for each paper in the NBER data set for which we are not able to obtain a text file directly from the NBER website. We use the PDFMiner program, wrapped in the `pdf2txt` package, to convert the text of each PDF paper to plain text format.[5]

B.3.2 Text cleaning

Relevant code file**:** `gibberishDetector.py`

In the NBER data set we encountered several papers that were transcribed from PDF to text as an unreadable jumble of letters and numbers. We identify and drop these text files from our data set by reviewing the five most common words in each transcribed text file and dropping those that meet several common indicative criteria, such as having "!" in the five most common words or having each of the five most common words be a single character, e.g. "t". We manually verified the accuracy of the algorithm by spot-checking discarded files. To our knowledge, this transcription error only occurred in the NBER data set and was not present in any papers in the data set of

[4] E.g., bibliographic data for Working Paper 25524 are available at https://www.nber.org/papers/w25524.bib.

[5] Documentation for PDFMiner is available here.

top-5 journal articles.

Relevant code files: `screenForCids.py` and `screenForLigs.py`

We observed two common classes of easily replaceable transcription errors: (1) the text pattern "(cid:###)", where each '#' represents a digit, and (2) transcription of ligatures, such as "fi", as a single character. In case (1), the text patterns do not replace characters in words, so we simply delete them. In case (2), we identify types of ligature transcription errors using regular expressions ("regex") of common words that contain them, such as "financial", and replace the single-character ligatures with their multi-character counterparts so that our script can read and match texts to our set of trigger words.

Relevant code file: `textCleanUp.py`

Once we obtain a text file for each paper, we clean the text by identifying and replacing additional common transcription errors. We first manually inspected a sample of transcribed papers and identified common UTF-8 transcription errors that could be easily remedied, such as the transcription of an "œ" as "\textbackslash xc5\textbackslash x92", and replaced them with readable characters. We also replaced ligatures identified in `screenForLigs.py` using this script.

B.3.3 Word Counts

We use a series of regular expressions (regex) searches, implemented using Python, to search for key words and phrases in the paper texts.[6] Before searching for any set of terms, we drop the references section from each paper to avoid incorrectly identifying papers based on the presence of key words in the titles of cited literature. We do so by finding instances of either "references", "works cited", or "bibliography" and identifying the instance with the highest count of the word "Journal" in the text immediately following it.[7] Once we identify the start of the references section,

[6] Documentation for regular expressions (regex) is available here.

[7] Specifically we focus on the first 5,000 characters after the mention of any of the three words referring to the cited literature.

we drop all text from the beginning of the references to the start of the appendix, if applicable.

Table A.I shows the search term categories that we use, and the trigger phrases we use for each category. For most search terms, we use a dictionary approach in which we simply search for instances of each trigger phrase, with switches indicating whether the terms are case-sensitive and/or end in a wildcard[8] as specified in Table A.I. To illustrate, for the search category 'Event Study' we look for the trigger phrases 'event stud' and 'event-stud'. Our search is not case-sensitive, implying that we will be both capitalized and noncapitalized versions of the trigger phrases: e.g. Event Study, Event study, and event study. For 'Event Study' we also include a wildcard at the end, which implies that we will capture all permutations of 'event stud' and 'event-stud': e.g. event study and event studies. For some search term categories, for example 'Administrative Data' and 'Big Data' we condition on the search term category 'Data'. This implies that we only include papers in our search that mention the word 'data' or any permutation of the word 'data' more than once. Our standard search script is specified in `wordcountsAppliedMicro_NBER.py` and `wordcountsAppliedMicro_top5.py` for the NBER and Top 5 datasets respectively. As is defined in Table A.I, we run the standard script four times for different search term categories: (1) with wildcards, case sensitive; (2) with wildcards, case insensitive; (3) without wildcards, case sensitive; and (4) without wildcards, case insensitive. For categories with more complex search instructions – such as when we search for one word from each of two categories within a given sentence – we run a separate script. To illustrate, for the category 'Survey Data' we look for any instance in which the word 'survey' and 'data' are mentioned within two full stops. We run this separate script for four search categories: survey data, identification, and structural models.

We then count the number of instances trigger phrases from a search term appear in the full text of a paper. Our figures show five-year moving averages, and are based on each paper having at least one mention of a trigger phrase in a search term category.

[8] A 'wildcard' refers to a character that can be substituted for zero or more characters in a search string. For example, the phrase 'estimat*', where '*' indicates a wildcard, will match to any of the phrases 'estimation', 'estimate', or 'estimator'.

B.3.4 Word Counts – Identification

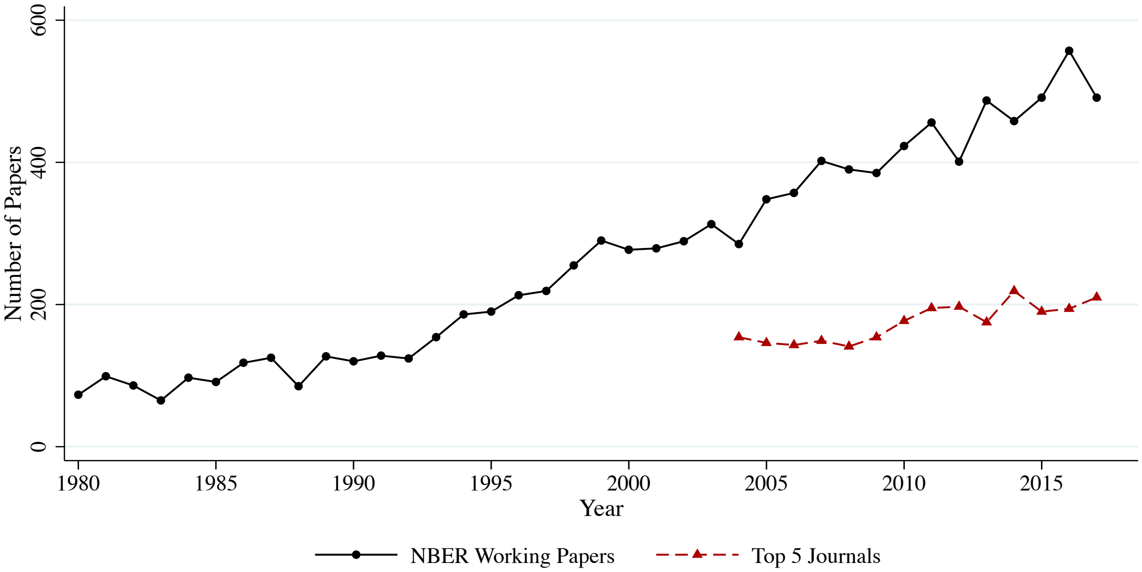
To search for the term 'identification', we require a more precise search procedure. We use two types of searches: (1) searching for a particular sentence structure with more flexible terms; and (2) searching for a set of specific terms that are associated with the phenomenon we aim to measure. We describe these two searches below.

Sentence structure. We search for sentences that match the following pattern. Between 2 full stops ('.') a matching sentence must contain the following features in the given order:

1. 'identif' with a wildcard at the end, followed by
2. any 0-2 words, followed by
3. any of the following terms: 'effect', 'response', 'impact', 'elasticit', 'parameter', or 'coefficient', with a wildcard at the end of the term. We exclude instances where the found term is the word 'effective'.

Specific trigger phrases. In addition to the flexible search pattern described above, we also identify the terms associated with identification that we specify in Table A.I. Note that we permit wildcards at the end of each of these trigger phrases.

Figure B.I: Number of papers in sample over time



Notes: The graphs shows the number of papers in our sample of NBER Working Papers and our sample of Top-5 journals over time.

Table B.I: Top-5 papers by JEL code

JEL Code	Field name	Count all Papers	Count Applied Micro Papers
A	General Economics and Teaching	17	9
B	History of Economic Thought, Methodology, and Heterodox Approaches	13	2
C	Mathematical and Quantitative Methods	857	386
D	Microeconomics	2,265	1,253
E	Macroeconomics and Monetary Economics	770	433
F	International Economics	416	416
G	Financial Economics	711	335
H	Public Economics	498	498
I	Health, Education, and Welfare	540	540
J	Labor and Demographic Economics	939	939
K	Law and Economics	169	169
L	Industrial Organization	762	762
M	Business Administration and Business Economics/Marketing/Accounting/Personnel Economics	191	161
N	Economic History	211	167
O	Economic Development, Innovation, Technological Change, and Growth	647	647
P	Economic Systems	90	74
Q	Agriculture and Natural Resource Economics/Environmental and Ecological Economics	194	194
R	Urban, Rural, Regional, Real Estate, and Transportation Economics	254	254
Y	Miscellaneous Categories	0	0
Z	Other Special Topics	165	117
N of Papers		4,344	2,830

Notes: Papers can have JEL codes in more categories. Our Applied Micro classification includes JEL codes: C9, F, H, I, J, K, L, O, Q, R.

Table B.II: NBER papers by category

Category Code	Category Name	Papers
AG	Aging	1,075
CH	Children	1,159
DEV	Development Economics	468
ED	Education	1,075
HC	Health Care	1,076
HE	Health Economics	1,541
IO	Industrial Organization	712
LS	Labor Studies	3,519
POL	Political Economy	529
PE	Public Economics	3,464
ITI	International Trade and Investment	1,285
EEE	Environment and Energy	745
All Papers	All Papers	10,324

References

David Card and Stefano DellaVigna. Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–61, 2013.