# 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

December 2019

*Contact information for the authors: Currie (corresponding author): Center for Health and Wellbeing, Princeton University, Princeton, NJ 08544, jcurrie@princeton.edu; Kleven: Department of Economics, Princeton University, Princeton, NJ 08544, kleven@princeton.edu; Zwiers: Center for Health and Wellbeing, Princeton University, Princeton, NJ 08544, ezwiers@princeton.edu. AEA/ASSA Session: Empirical Practice in Economics: Challenges and Opportunities; Chair: Guido Imbens; Discussant: Lawrence Katz. We thank Dana Scott for outstanding research assistance.

The last 40 years have seen huge innovations in computing technology and in the availability of all types of data. It has become commonplace to process data sets derived from millions of adminstrative records or by using (as in this paper) new methods of data generation such as text mining. While we like to think that we can choose what to sip from the fire hose of data that is now available, the torrent is also sweeping the field of economics in specific directions that we attempt to document in this paper. We argue that new data often requires new methods, which in turn can inspire new data collection.

# I Data and Methods

Our data come from two sources: The first is all papers in the National Bureau of Economic Research (NBER) working paper series between January 1, 1980 and June 30, 2018, and the second is all papers published in the "Top Five" economics journals (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies) between January 1, 2004 and August 2019. Because our focus is on the ways that new data and methods are changing the profession, we focus on the field of applied microeconomics. For top-five papers, we define aapplied micro using the Journal of Economic Literature (JEL) codes suggested by Card and DellaVigna (2013) with the addition of category I3 for "Welfare, Wellbeing, and Poverty" and category Q for environmental economics. For the NBER working papers, we include papers in the following programs: Aging; Children; Development; Education; Health Care; Health Economics; Industrial Organization; Labor Studies; Political Economy; Public Economics; International Trade; and Environment and Energy. We end up with a sample of 2,825 top-five papers and 10,324 NBER working papers.

Figure I shows that the number of top-five papers has increased over time, especially since 2008, and that the number of applied microeconomics papers has increased even faster. As a result, the fraction of applied microeconomics papers published in major general interest journals has increased from 55-60% to about 75%.

For each paper, we use a plain text version excluding references for the analysis. We use a series of regular expression (regex) searches, implemented using Python, to find keywords and phrases. The full details of our data and methods are provided in appendix B, while Table A.I lists the search categories we consider and the specific trigger phrases within each category. For most categories, we simply search for any instances of each trigger phrase. Take the category 'Event Study' as an example: here we search for the trigger phrases 'event stud' and 'event-stud', including a wild card at the end. The wild card ensures that we capture all permutations of 'event stud' and 'event-stud' such as 'event studies' or 'event-study specification.' This particular search is not case-sensitive, so that 'Event Study,' 'Event study,' and 'event study' will all be counted. For other categories and trigger phrases (for example, the phrase 'DiD' in the difference-in-differences category), it is important that the search is case sensitive.

For some categories, we search for specific phrases while conditioning on using the word 'data.' That is, our search focuses only on papers that mention the word 'data' or any permutation of the word 'data' at least once. An example would be 'clustering,' where we are interested in capturing empirical papers talking about standard errors rather than, say, theoretical papers talking about clustering around a discontinuous incentive. Conditioning on 'data' reduces (by may not entirely eliminate) the likelihood of picking up false positives. For some categories, we use more complex search instructions. To illustrate, for the category 'Survey Data', we look for any instance in which the words 'survey' and 'data' are both mentioned within two full stops. The search for 'identification' is based on more involved algorithm, because we want to capture language like "the behavioral response is identified using exogenous tax variation" without counting papers that use permutations of the word 'identification' in ways unrelated to causal research designs. In general, our search algorithms have been designed by trial and error, with the intention of minimizing the prevalence of Type I and Type II errors. Even if some errors remain, this is not a major concern due to the fact that our main focus is on trends rather than on levels.

# II Results

In this section, we document methodological changes in applied microeconomics by plotting the time series of methods-related words and phrases since 1980 (for NBER working papers) and 2004 (for top-five papers).[1] The four panels of Figure II highlight different dimensions of what has been dubbed the "credibility revolution" in economics (Angrist and Pischke 2010). Panel A illustrates a virtually linear rise in the fraction of papers, in both the NBER and top-five series, which make explicit reference to identification. This fraction has risen from around 4 percent to 50 percent of papers. Figure A.I in the online appendix shows the time trends of specific identification concerns: omitted variables, selection biases, reverse causation, and simultaneity. Curiously, concerns about simultaneity seems to have fallen out of fashion, perhaps being replaced by reverse causation.

With this focus on cleaner identification has come a somewhat slower rise in the use of experimental and quasi-experimental methods, illustrated in Panel B. Currently, over 40 percent of NBER papers and about 35 percent of top-five papers make reference to randomized controlled trials (RCTs), lab experiments, difference-in-differences, regression discontinuity, event studies, or bunching. We unpack these individual methods further below.

Panel C shows a very similar pattern in references to administrative data. The NBER series starts increasing in the mid-1990s, rising to about 30 percent today. The top-five series shows a similar increase, but with a lag of about three years. Appendix Figure A.II considers alternative data sources. References to survey data have declined since the mid-1990s, and they are less frequent in top-five journals than in NBER working papers. By contrast, references to proprietary data, internet data, and big data have increased over time, although the exact timing varies across categories. The term Big Data suddenly sky-rockets after 2012, with a more recent uptick in the top five.

Panel D depicts what we have called the "Graphical Revolution" in applied economics. It tracks the ratio of words about figures (charts, graphs, figures, and plots) to words about tables.

[1] Related, Brice and Montesinos-Yufa (2019) documents methodological changes in economics based on analyzing the titles of top-journal papers. Our analysis is based on the full texts of all papers.

The importance of figures relative to tables has increased substantially over time and in two phases. The first phase happened in the 1990s and likely reflects the diffusion of new software such as STATA that made it easier to create impactful figures. The second phase has happened in the last 10-12 years and is still ongoing. This reflects the increasing use of administrative datasets, which lend themselves to compelling graphical representation using raw data and non-parametric approaches. It is interesting to note that the top-five journals tend to be more graphically oriented than the NBER working paper series. This may reflect the type of paper that is selected into the top five or editorial decisions that favor graphical over tabular evidence.

The next two figures unpack the rise in experimental and quasi-experimental methods. Figure III documents the increased focus on field and laboratory experiments. Panel A shows a sharp rise in the fraction of NBER working papers discussing randomized controlled trials since 2005, and especially since 2010. Given RCTs are particularly common in the field of Development Economics, the later rise is likely an artifact of the creation of the NBER program in Development in 2012. The top-five series shows a much gentler rise in RCTs, to about 5 percent of papers by 2019. Laboratory experiments have grown steadily in popularity since the late 1990s, which is connected to the rise of Behavioral Economics during this time period (see Kleven 2018). Interestingly, lab experiments have been more impactful in top-five journals than in NBER working papers, with mentions in more than 10 percent of top-journal publications in their peak year.[2] More recently, the popularity of lab experiments has turned downwards in both series.

The rise of experiments has generated its own critiques (e.g., Deaton 2010; Imbens 2010), one of which is a concern about external validity. Appendix Figure A.III shows that discussions of external validity begins in the late 1990s and rises sharply in both the NBER and top-five series after 2005. This time pattern mirrors the rise of lab and field experiments quite closely. One possible reaction to external validity concerns is to supplement treatment effect estimates with evidence on mechanisms, thereby allowing readers to better gauge if and when the estimates can be extrapolated to other settings. Consistent with this, Appendix Figure A.IV shows an impressive

[2] The difference between top journals and NBER working papers may reflect a European tilt in the use of lab experiments.

rise in the fraction of applied micro papers discussing mechanisms, from about 20 percent to about 60 percent in the NBER series. The fraction of top-five papers discussing mechanisms is even higher, more than 70 percent today, suggesting that editors either select papers that provide evidence on mechanisms or push authors to add such evidence as part of the editorial process.

Figure IV drills down on specific quasi-experimental methods: difference-in-differences, regression discontinuity, event studies, and bunching. These methods have all become more popular over time, in roughly the order named. The use of difference-in-differences was virtually non-existent until 1990 and then starts growing.[3] It is quite striking that, today, almost 25 percent of all NBER working papers in applied micro make references to difference-in-differences. Regression discontinuity approaches start gaining popularity around 2000, following the early contributions by Angrist and Lavy (1999) and Hahn, Todd, and Van der Klaauw (2001), published as NBER working papers a couple of years prior.[4] Event studies and bunching approaches are more recent, having exploded during the last decade. Both of these approaches are closely linked to the increased use of administrative data sources, which are critical to the effective implementation of these data-demanding approaches. Over time, event studies have become almost synonymous with difference-in-differences: It is now rare to use difference-in-differences without showing an event study graph, and conversely it is rare to show event studies without a control group. As a result, the sharp rise in the use of event studies over the last ten years goes hand in hand with the increased slope of the difference-in-differences series during this time period. The modern bunching approach starts with Saez (2010), although the NBER working paper version of that paper was published more than ten years prior.[5]

One might wonder whether these new methods have grown at the expense of older empirical methods such as instrumental variables and fixed effects. However, Appendix Figure A.V suggests

[3]The very first papers that mention difference-in-differences estimators in our data are Ashenfelter and Card (1985) and Card and Sullivan (1988), which appear as NBER working papers in 1984 and 1987, respectively. As far as we are aware, the very first paper to use a difference-in-differences approach is Ashenfelter (1978), although that paper did not use the difference-in-differences language.

[4]There were other RD papers around this time (reviewed by, e.g., Hahn, Todd, and Van der Klaauw 2001).

[5]Bunching was discussed quite a bit in the early 1980s in the context of the non-linear budget set approach to labor supply estimation (see Kleven 2016).

that this is not the case. Mentions of instrumental variables and fixed effects have both grown continuously since the 1980s, while mentions of matching methods have been growing since the mid-1990s.[6] The fact that old and new methods appear to be complements rather than substitutes suggests that another outgrowth of the credibility revolution is the rise of the "collage" approach to empirical work. Authors no longer hang their hats on a single method or dataset, but attempt to make a case based on a more multi-pronged approach.

Figure V documents another implication of the rise in "big data" and more credible research designs, namely that authors have become increasingly concerned with whether their estimates are precisely estimated, and not merely with whether they are significantly different from zero in a statistical sense. The focus on precision has grown continuously since the beginning of our series, and is almost identical in the NBER and top-five series.[7] Panel B of Figure V highlights a specific dimension of this change: the concept of a "precisely estimated zero." This concept, virtually non-existent until around 2010, has seen a sharp rise within the last decade. We view this as a positive development, because it holds the promise to reduce the publication bias that arises in a world where empiricists hunt for "large effects" in order to impress editors and drop projects that show "no effects." We still have a ways to go, however. The idea that a precisely estimated zero could be as useful as a precisely estimated non-zero is only mooted in a little over 1 percent of papers.

Figure VI considers a number of recent developments and innovations. Binscatter plots have become a more popular way of visualizing big data since they were used in Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011). Discussions of preanalysis plans have trended up sharply since 2012, when the American Economics Association voted to create a registry for them. Machine learning is the most popular among these brand new methods, with mentions in 2.5 percent of NBER working papers. Text analysis like we do here has also become more common,

[6] Synthetic control methods are a relative newcomer, showing rapid growth in the NBER working paper series since 2010.

[7] Closely related, authors have become increasingly concerned with the correct measurement of standard errors. As an example, Figure A.VI in the online appendix shows that, after year 2000, there has been a massive increase in attention paid to clustering of standard errors.

with mentions in about 1 percent of NBER working papers in 2019.

Finally, Figures A.VII-A.IX in the online appendix consider time trends in references to structural methods. Specifically, we focus on words and phrases related to structural models, general equilibrium, and specific functional forms (see Table A.I for details). These figures provide an example of a case where it is important to consider heterogeneity across subfields in applied micro. While there is no clear time trend in applied micro as a whole (see Panel A of each figure), this masks relatively strong secular trends within specific subfields (see Panels B-D of each figure). For example, structural models have been on the decline in Labor Economics, whereas they have been on the rise in Public Economics and Industrial Organization.

# III Conclusion

In the late 1960s and 1970s, new computing methods and data sources reshaped economics and made it a more applied field. Fields like Labor and Public Economics shifted their focus from theory and discussions of institutions and case studies towards estimating quantities such as returns to schooling and behavioral elasticities. Micro data from surveys in developing countries made it possible to focus on the determinants of household and individual well-being. Large-scale social experiments such as the Negative Income Tax Experiments and the RAND Health Insurance Experiment were conducted and evaluated for the first time. In turn, these new data prompted a flowering of interests in new methods such as selection and panel data approaches.

We may be at a similar turning point in the field today, with a proliferation of new data and methods. If history is any guide, some methods will stick and others will prove to be a flash in the pan. However, the larger trend towards demanding greater credibility and transparency from researchers in applied economics will likely continue.

# References

ANGRIST, JOSHUA D., AND VICTOR LAVY (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114(2), 533–575.

ANGRIST, JOSHUA D., AND JÖRN-STEFFEN PISCHKE (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24(2), 3–30.

ASHENFELTER, ORLEY (1978): "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, pp. 47–57.

ASHENFELTER, ORLEY, AND DAVID CARD (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*, 67(4), 648–660.

BRICE, BRANDON, AND HUGO M. MONTESINOS-YUFA (2019): "The Era of Evidence," Working Paper.

CARD, DAVID, AND STEFANO DELLAVIGNA (2013): "Nine Facts about Top Journals in Economics," *Journal of Economic Literature*, 51(1), 144–61.

CARD, DAVID, AND DANIEL SULLIVAN (1988): "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56(3), 497–530.

CHETTY, RAJ, JOHN N. FRIEDMAN, NATHANIEL HILGER, EMMANUEL SAEZ, DIANE WHITMORE SCHANZENBACH, AND DANNY YAGAN (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 126(4), 1593–1660.

DEATON, ANGUS (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48(2), 424–55.

HAHN, JINYONG, PETRA TODD, AND WILBERT VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

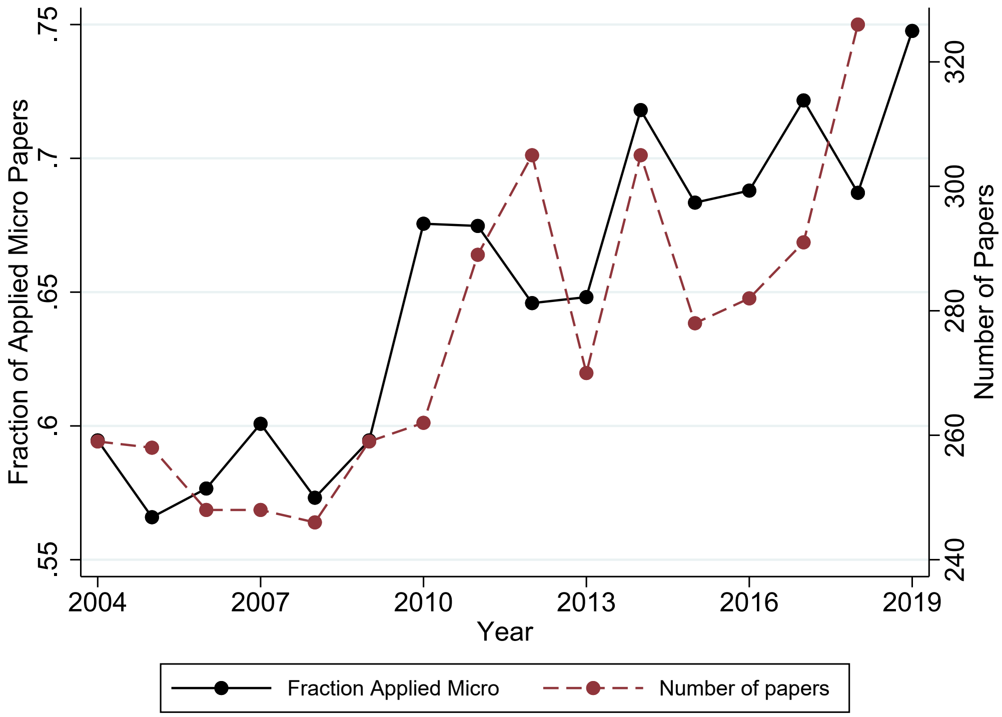
IMBENS, GUIDO W. (2010): "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48(2), 399–423.

KLEVEN, HENRIK J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

——— (2018): "Language Trends in Public Economics," Slides.

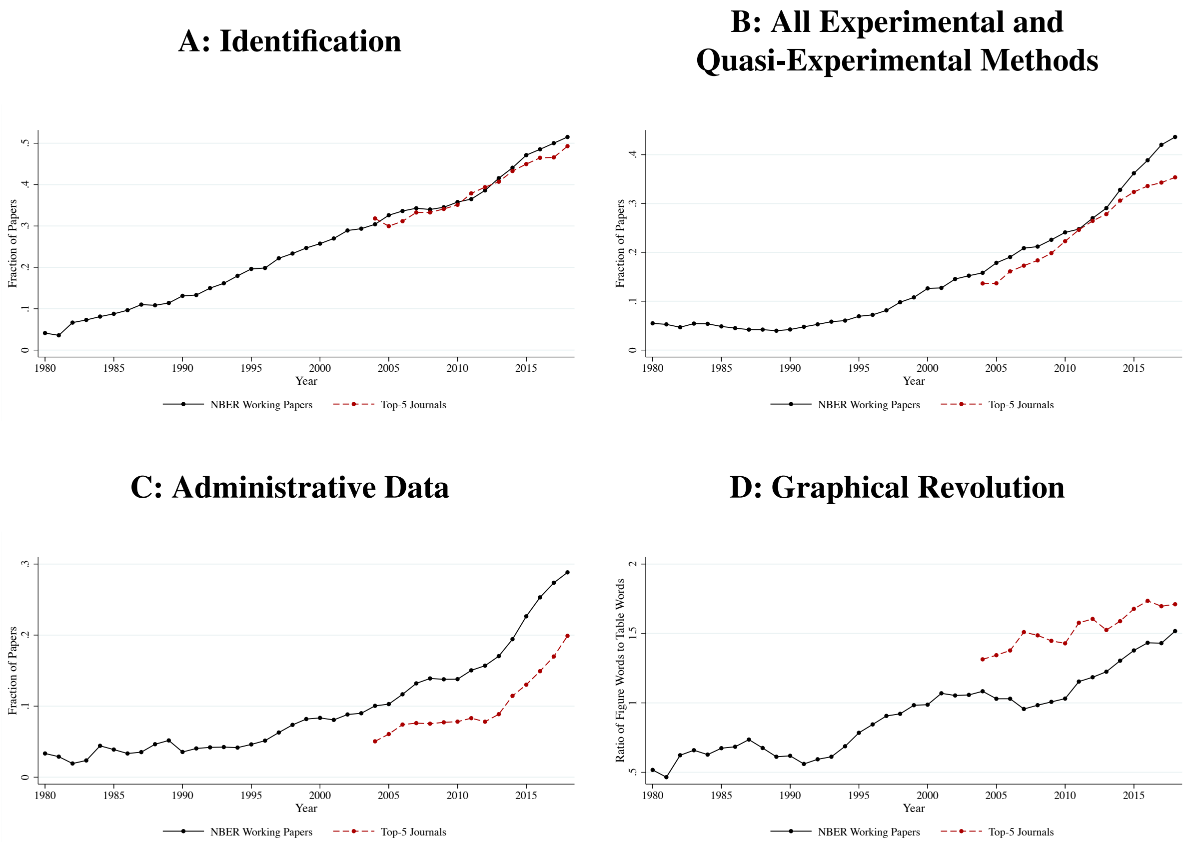
SAEZ, EMMANUEL (2010): "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, 2(3), 180–212.

Figure I: Fraction of Applied Microeconomics Articles in Top-5 Journals



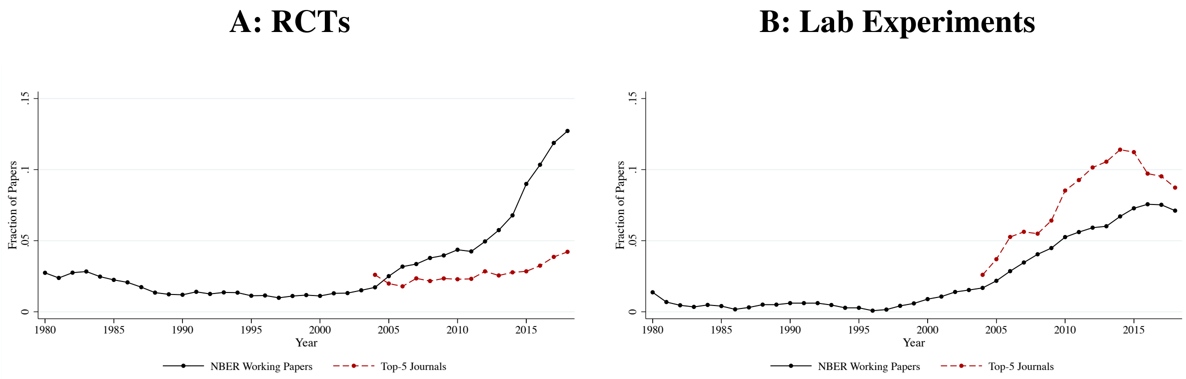

Notes: This figure shows the fraction of papers in top-5 journals that report an applied microeconomics JEL code (left axis) and the total number of papers in the top-5 journals (right axis).

Figure II: The Credibility Revolution



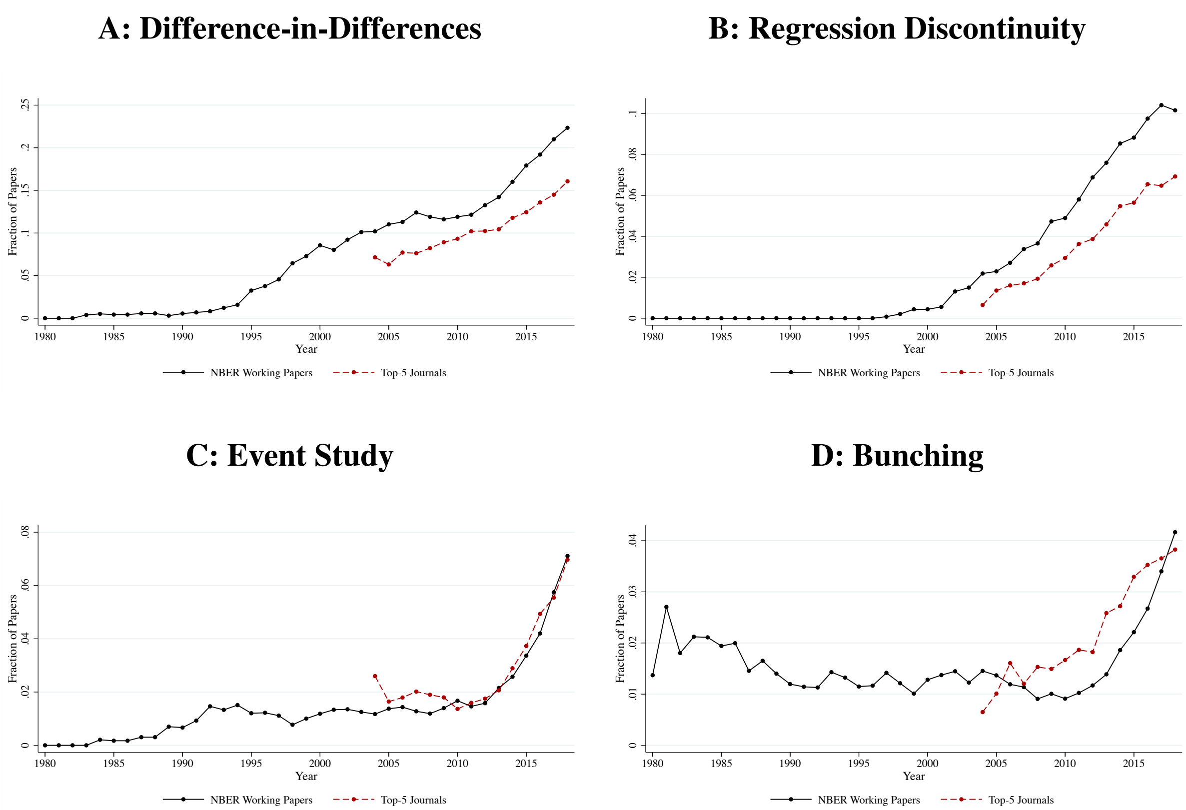

Notes: This figure shows different dimensions of the "credibility revolution" in economics: Identification (Panel A), All Experimental and Quasi-Experimental Methods (Panel B), Administrative Data (Panel C), and the Graphical Revolution (Panel D). Panel D shows the ratio of the number of 'Figure' terms to the number of 'Table' terms mentioned. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure III: Experimental Methods




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

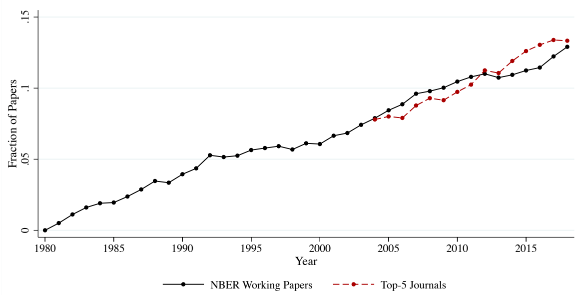
Figure IV: Quasi-Experimental Methods



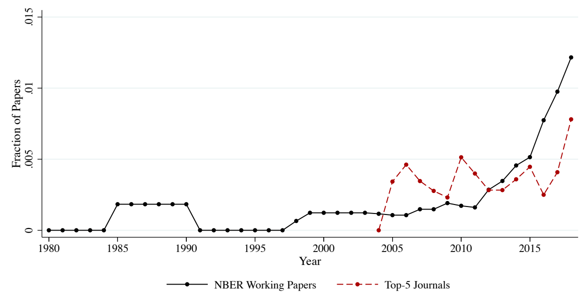

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure V: Precision

**A: Precisely Estimated**



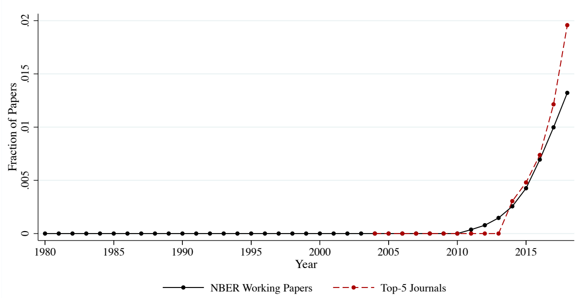

**B: Precisely Estimated Zero**



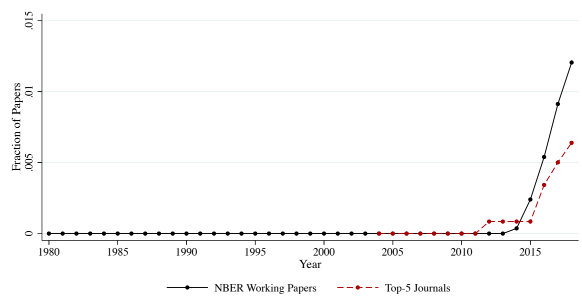

Notes: This figure shows the fraction of papers referring to each term. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure VI: What's New?

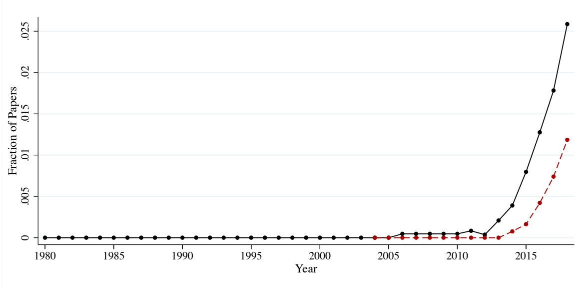
**A: Binscatter**



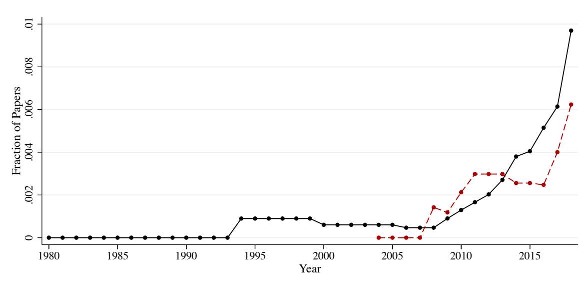

**B: Preanalysis Plan**




**C: Machine Learning**




**D: Text Analysis**

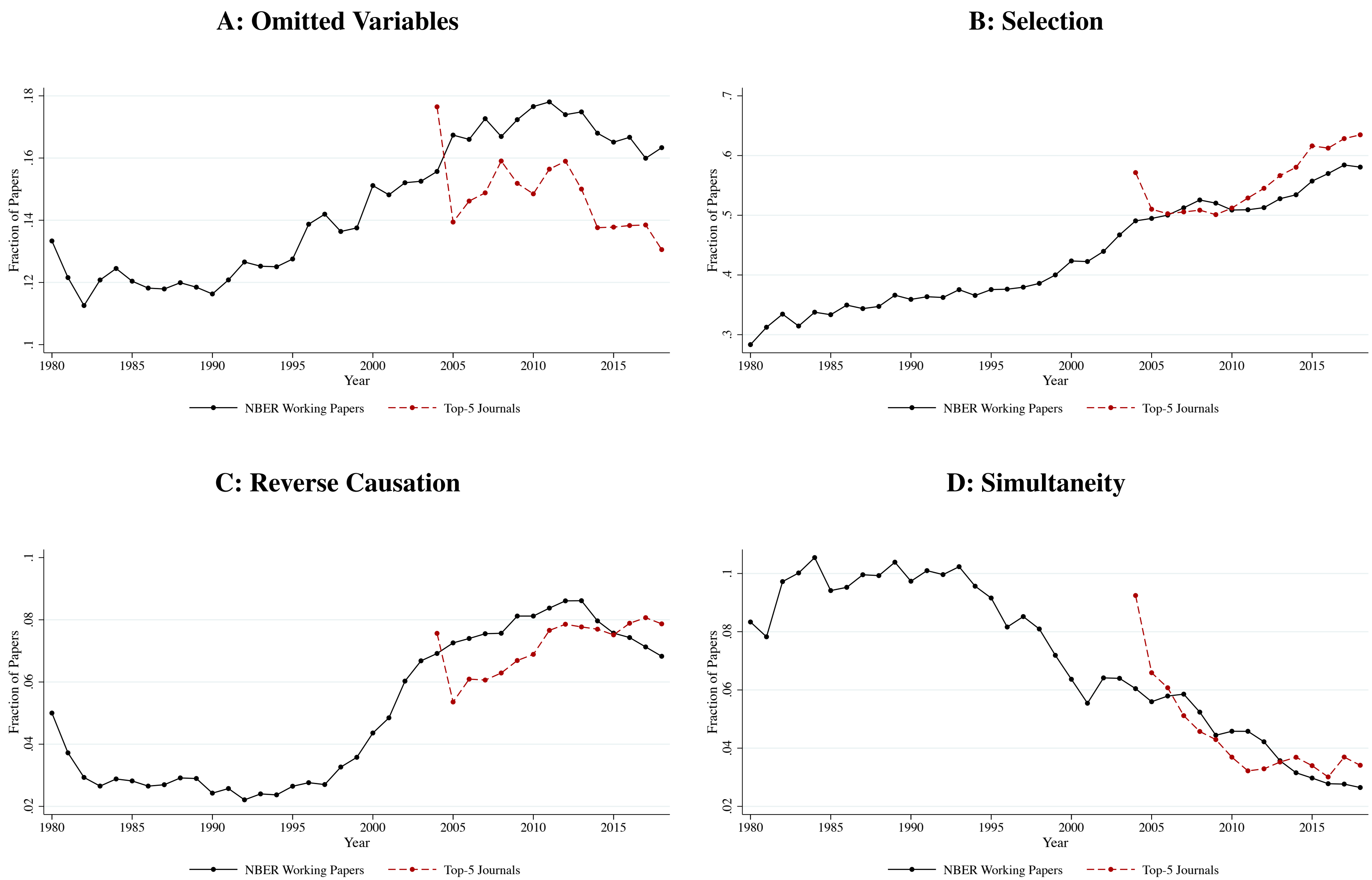



Notes: This figure shows the fraction of papers referring to each method. See Table A.I for a list of terms. The series show 5-year moving averages.

# Online Appendix (Not for Publication)

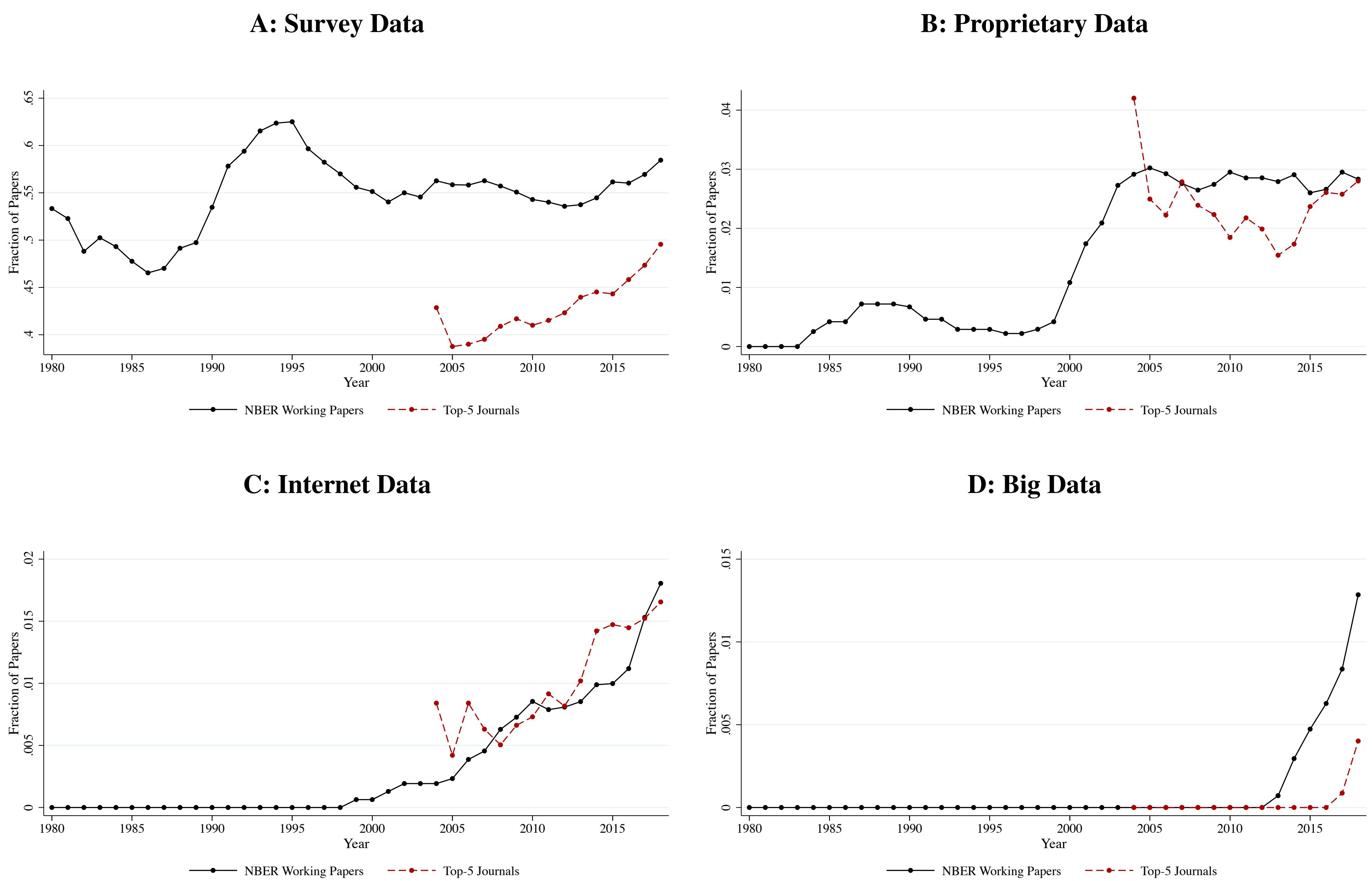
## A Supplementary Figures and Tables

Figure A.I: Identification Concerns



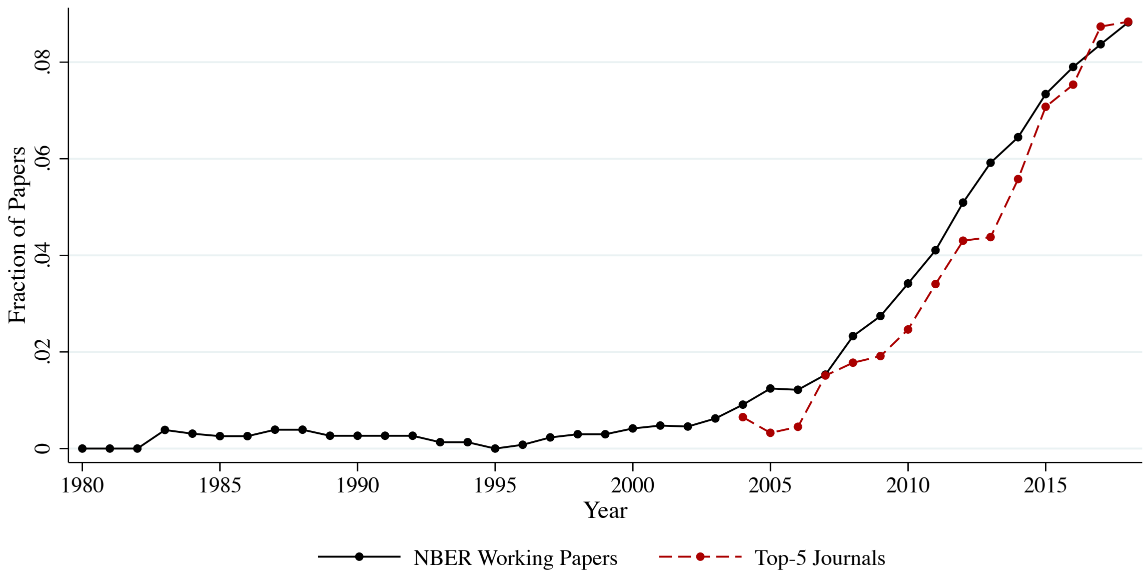

Notes: This figure shows the fraction of papers referring to each term. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.II: Data



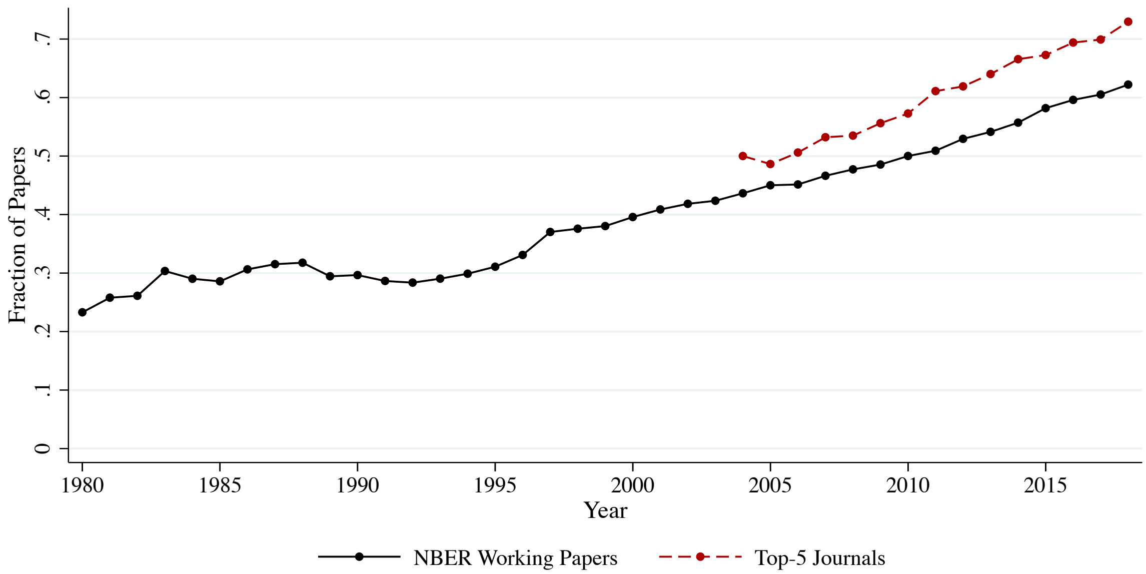

Notes: This figure shows the fraction of papers referring to each type of data. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.III: External Validity



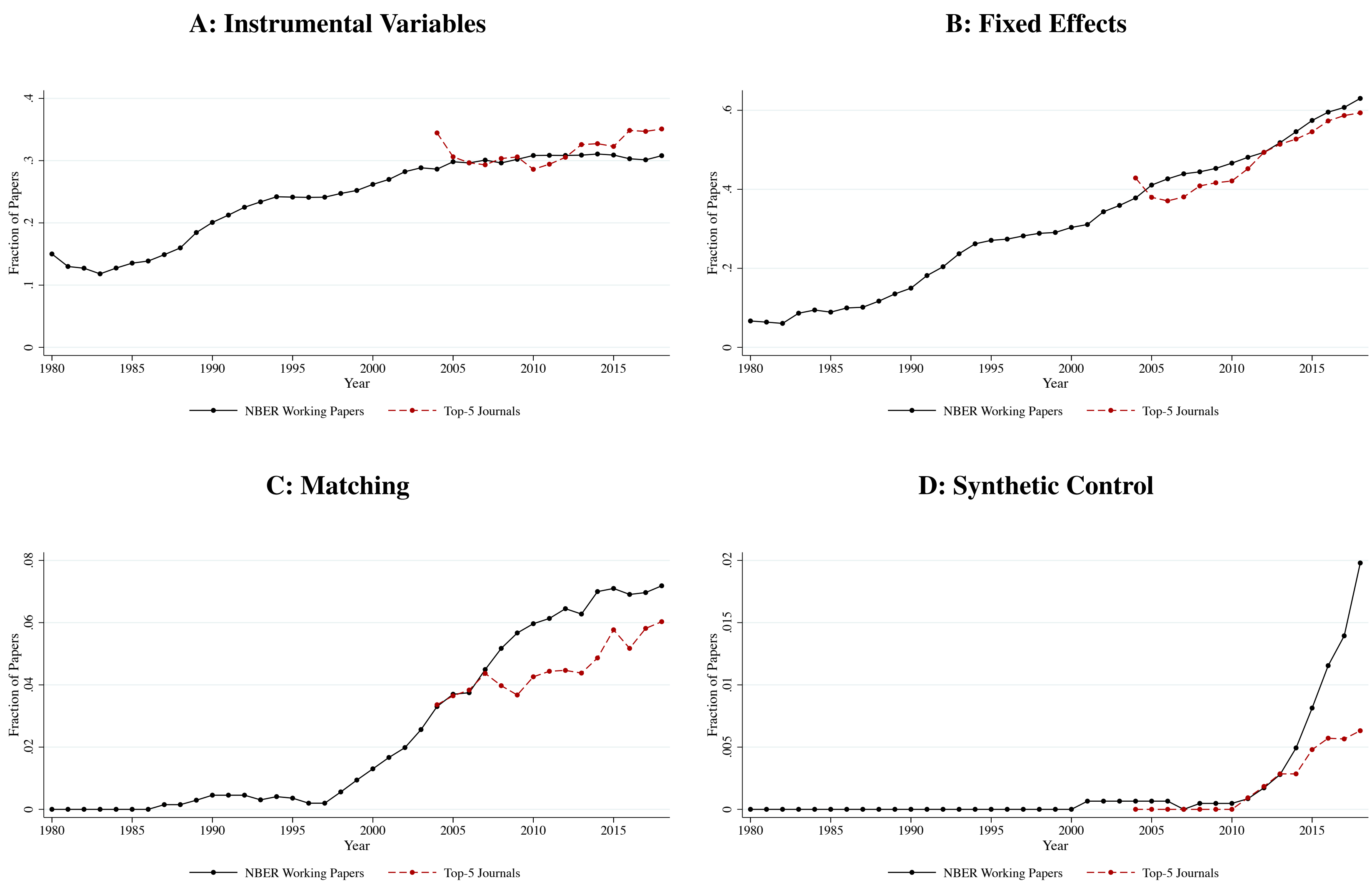

Notes: This figure shows the fraction of papers referring to external validity. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.IV: Mechanisms



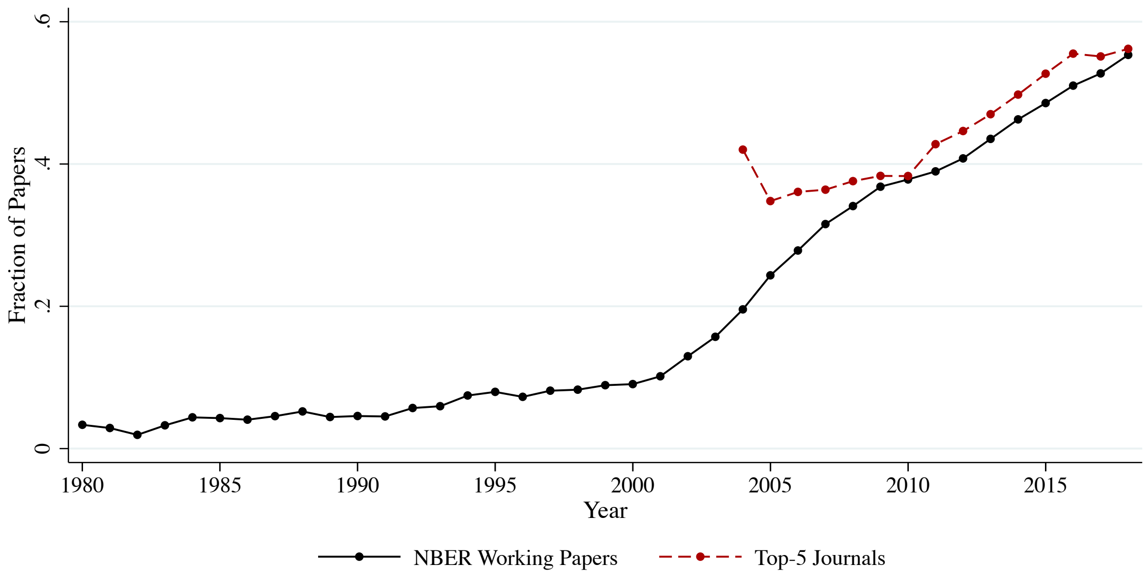

Notes: This figure shows the fraction of papers referring to mechanisms. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.V: Other Empirical Methods



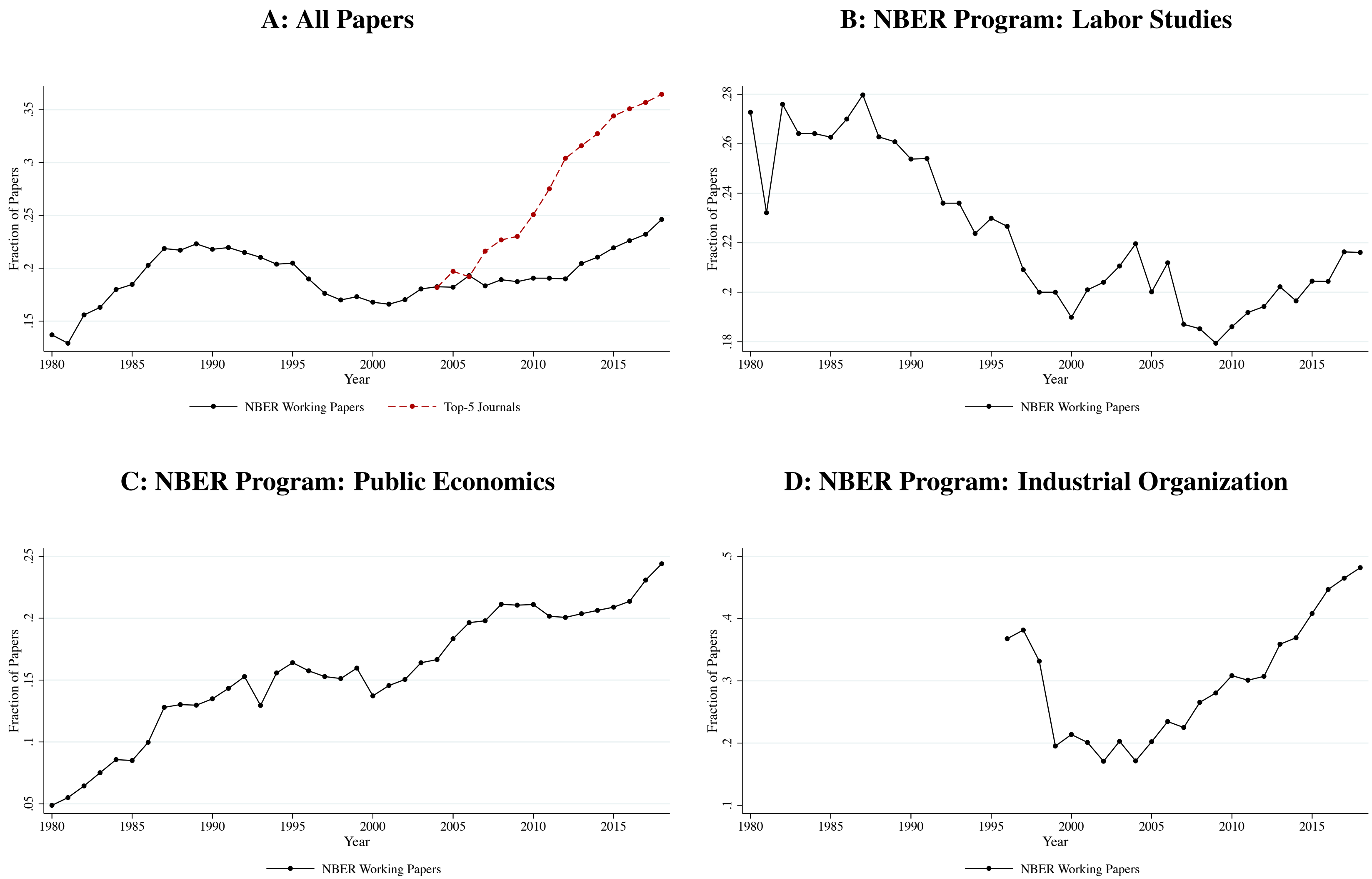

Notes: This figure shows the fraction of papers referring to each method. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.VI: Clustering



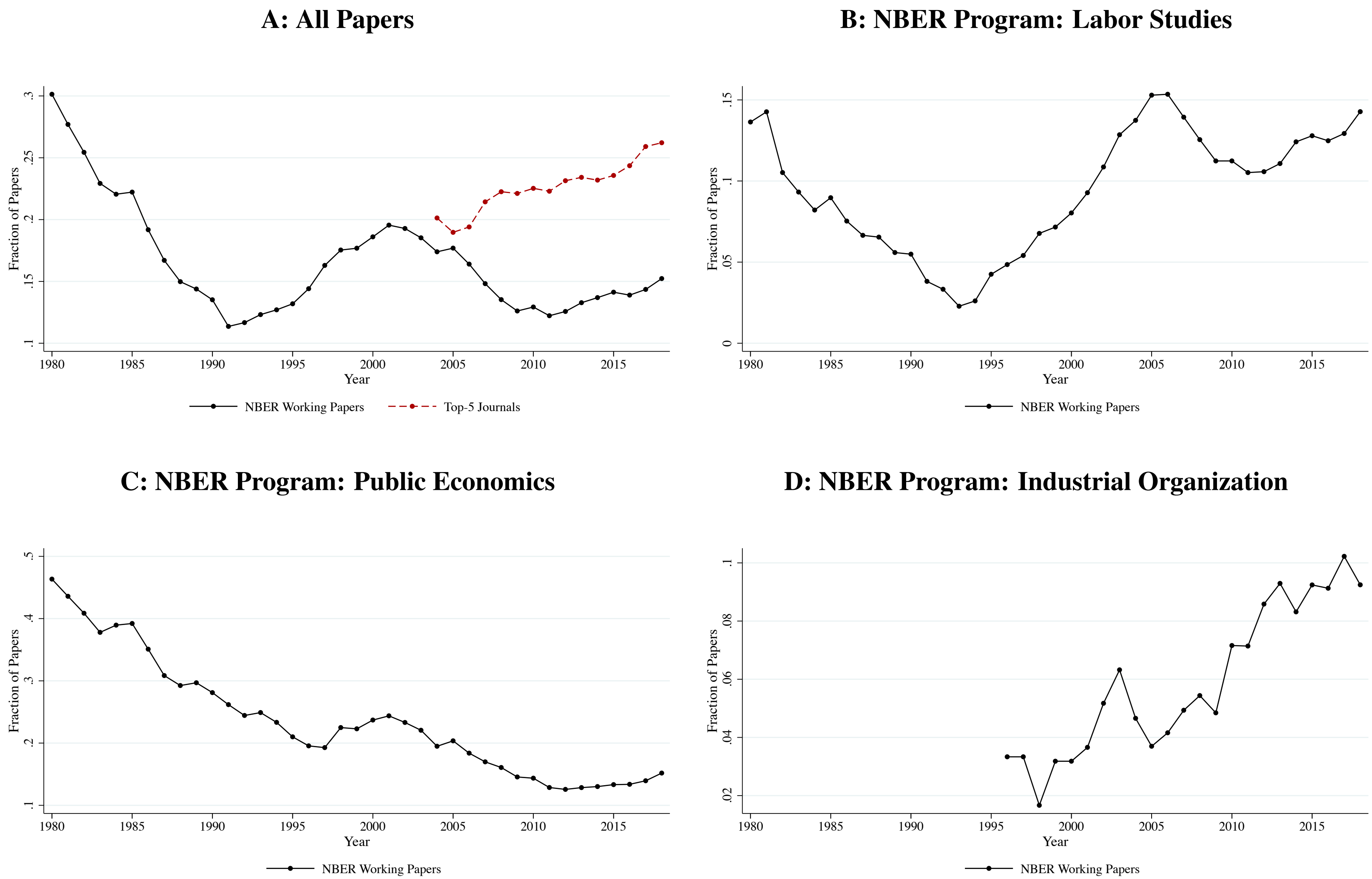

Notes: This figure shows the fraction of papers referring to clustering. See Table A.I for a list of terms. The series show 5-year moving averages.

Figure A.VII: Structural Methods: Structural Model



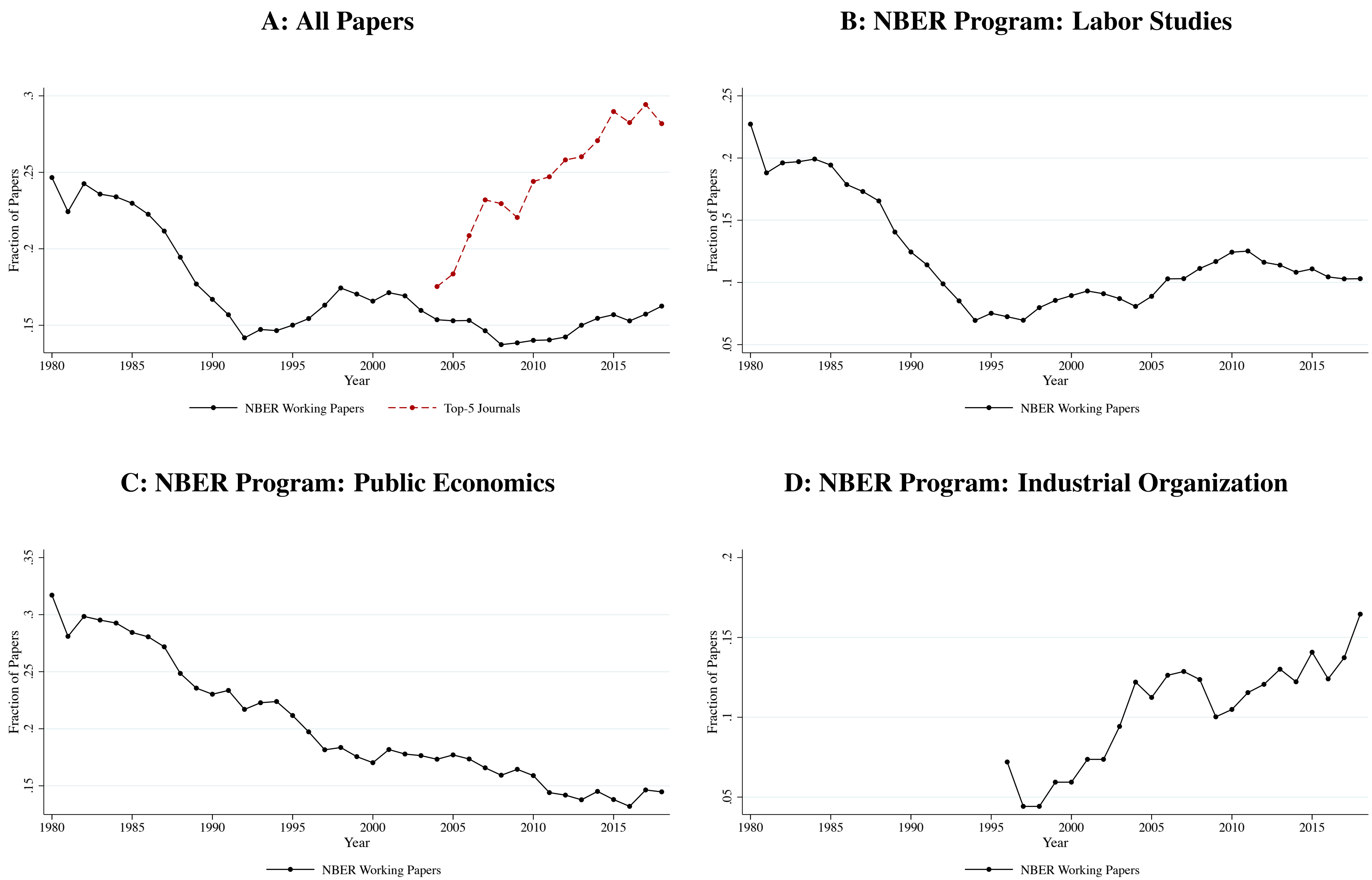

Notes: This figure shows the fraction of papers referring to structural methods. Panel A shows all papers in applied microeconomics, while Panels B-D focus on NBER working papers within specific programs (LS, PE, and IO). The IO series omit the first 5 data points, because of the low number of papers 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.VIII: Structural Methods: General Equilibrium




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.IX: 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 |
| 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 |

| | | | | | |
|---|---|---|---|---|---|
| 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. Eventually we end up with a sample of 4,336 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 2,825 applied microeconomic papers.[8]

### 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

[8] 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.

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.[9]

The more recent of these papers are available in plain text format at the nber.org URL associated with the paper.[10] 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 – 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.[11]

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

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

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

### 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.[12]

#### 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 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

[12] Documentation for PDFMiner is available here.

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.[13] 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.[14] Once we identify the start of the references section, 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

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

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

end in a wildcard[15] 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.

#### 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.

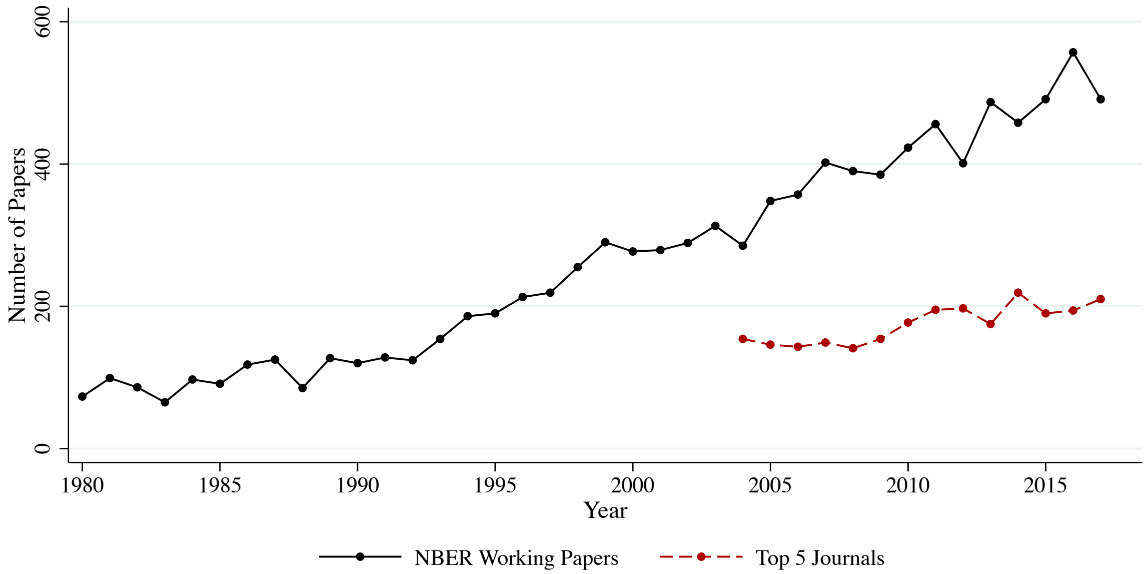
[15] A 'wildcard' refers to a character that can be substituted for zero or more characters in a search string. For example, the phrase 'estimat*', where '*' indicates a wildcard, will match to any of the phrases 'estimation', 'estimate', or 'estimator'.

**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 identifcation 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 | 2,261 | 1,250 |
| E | Macroeconomics and Monetary Economics | 765 | 431 |
| F | International Economics | 416 | 416 |
| G | Financial Economics | 708 | 335 |
| H | Public Economics | 497 | 497 |
| I | Health, Education, and Welfare | 539 | 539 |
| J | Labor and Demographic Economics | 936 | 936 |
| K | Law and Economics | 168 | 168 |
| L | Industrial Organization | 760 | 760 |
| M | Business Administration and Business Economics/Marketing/Accounting/Personnel Economics | 191 | 161 |
| N | Economic History | 210 | 167 |
| O | Economic Development, Innovation, Technological Change, and Growth | 646 | 646 |
| 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 | 253 | 253 |
| Y | Miscellaneous Categories | 0 | 0 |
| Z | Other Special Topics | 165 | 117 |
| N of Papers | | 4,336 | 2,825 |

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 |