Tracking the Credibility Revolution across Fields*

Paul Goldsmith-Pinkham†

April 2, 2026

Abstract

How far has the credibility revolution spread beyond applied microeconomics? I update Currie, Kleven, and Zwiers (2020b) using approximately 44,000 papers—31,500 NBER working papers (1982–2025) and 12,300 articles from eleven top economics and finance journals (2011–2024)—measuring mentions of empirical methods through keyword matching. Three findings emerge. First, finance and macro/other fields differ substantially from applied micro in their mention of credibility revolution methods: as of 2024, 63 percent of applied micro papers mention experimental or quasi-experimental methods, compared to 47 percent in finance and 39 percent in macro/other. The current levels in finance and macro/other are comparable to where applied micro was in 2008–2010, though the long-run trajectories may differ. Second, growth outside applied micro is driven overwhelmingly by difference-in-differences; including DiD raises the share of finance papers mentioning any experimental or quasi-experimental method by roughly 55 percent versus 30 percent for applied micro. Other quasi-experimental methods—instrumental variables, regression discontinuity, experiments—have seen far less growth. Third, I document a striking gap between the methods studied in the *Journal of Econometrics*—where nonparametric estimation and asymptotic theory dominate—and those used by applied researchers, where DiD and identification strategies dominate. Published journal articles confirm these patterns are not artifacts of the NBER sample.

JEL Codes: C18, C81, B41

Keywords: Credibility revolution, difference-in-differences, text analysis, empirical methods, causal inference

*Goldsmith-Pinkham: paul.goldsmith-pinkham@yale.edu. I thank Dana Scott, Pedro Sant'Anna, Nils Enevoldsen, and Esmée Zwiers for helpful comments and suggestions. Full-text search of the papers used in this manuscript are available at `https://paulgp.com/econlit-pipeline/` to allow for alternative keyword searches.

†Yale School of Management and NBER

How far has the credibility revolution spread? Angrist and Pischke (2010) documented a sea change in how economists approach empirical work—a shift toward transparent research designs, explicit identification strategies, and credible causal inference. Currie, Kleven, and Zwiers (2020b) showed that this shift was accelerating through the late 2010s, at least in applied microeconomics. But that analysis left open a basic question: are finance, macroeconomics, and other fields keeping pace, or has the revolution been narrower than it appears?[1]

I take up this question by extending Currie, Kleven, and Zwiers (2020b)'s approach to a much larger corpus. Using keyword matching on the full text of approximately 44,000 economics papers—31,500 NBER working papers (1982–2025) and 12,300 articles from eleven top journals (2011–2024)—I track mentions of empirical methods across fields and over time. The expanded sample adds finance and macro/other fields, which were omitted from the original analysis, and supplements working papers with published journal articles. Because the analysis measures keyword *mentions* rather than verified *use*, the trends should be interpreted as tracking the diffusion of methodological language—a proxy for, but not identical to, actual method adoption.

The answer is clear: mentions of credibility revolution methods have spread unevenly across fields. I organize the findings around three main results.

First, finance and macro/other differ substantially from applied micro on most measures. As of 2024, 63 percent of applied micro papers mention experimental or quasi-experimental methods, compared to 47 percent in finance and 39 percent in macro/other (Table 3). In identification language, the current levels in finance and macro/other are comparable to where applied micro was in 2008–2010. The gap has shown little sign of closing.

Second, the credibility revolution outside applied micro has been—to a first approximation—a difference-in-differences revolution. Including DiD in the methods measure raises the finance share by roughly 55 percent versus 30 percent for applied micro. Other quasi-experimental tools—instrumental variables, regression discontinuity, RCTs—have seen far less growth in finance and macro. This reliance on a single method is striking given the recent econometrics literature highlighting sensitivities in DiD designs (Roth 2022; De Chaisemartin and d'Haultfoeuille 2020; Callaway, Goodman-Bacon, and Sant'Anna 2024).

Third, I document a pronounced gap between the methods studied in the *Journal of Econometrics*—where nonparametric estimation, bootstrap methods, and asymptotic theory dominate—and those used by applied researchers, where DiD and identification strategies are the dominant tools. The tools powering the credibility revolution and the theoretical literature developing new estimators occupy largely separate methodological spaces.

Two features of the analysis strengthen confidence in these patterns. Published articles from top journals show trends that closely mirror the NBER data, with slightly higher rates of credibility revolution methods—consistent with a publication selection effect favoring methodologically rigorous papers. And a validation exercise using LLM-based classification confirms that keyword matching achieves 80–92 percent agreement rates for most method categories with more sophisti-

[1]Throughout this paper, I use "macro/other" to refer to the NBER field grouping that includes macroeconomics alongside several other programs; see Table 2 for the full composition.

cated approaches at near-zero computational cost, though agreement is lower for broader categories like identification strategy and structural models.

The paper proceeds as follows. Section 1 describes the data and methods. Section 2 presents the NBER working paper results, documenting trends across fields, programs, and methods. Section 3 extends the analysis to published articles from top journals. Section 4 examines the gap between econometric theory and applied practice. Section 5 concludes.

1 Data and Methods

I measure mentions of empirical methods over time following the approach in Currie, Kleven, and Zwiers (2020b): searching the full text of papers for keywords and regular expressions that capture the language of the credibility revolution (e.g. "threats to identification" or "identification strategy").[2]

NBER Working Papers. I collect the full text of approximately 31,500 NBER working papers from the NBER website, covering papers 1000 through the most recent available (1982–2025). Unlike Currie, Kleven, and Zwiers (2020b), who focus exclusively on "applied micro" papers, I include all papers in the NBER working paper series. Each paper is associated with one or more of nineteen NBER research programs, which I use for field classification.

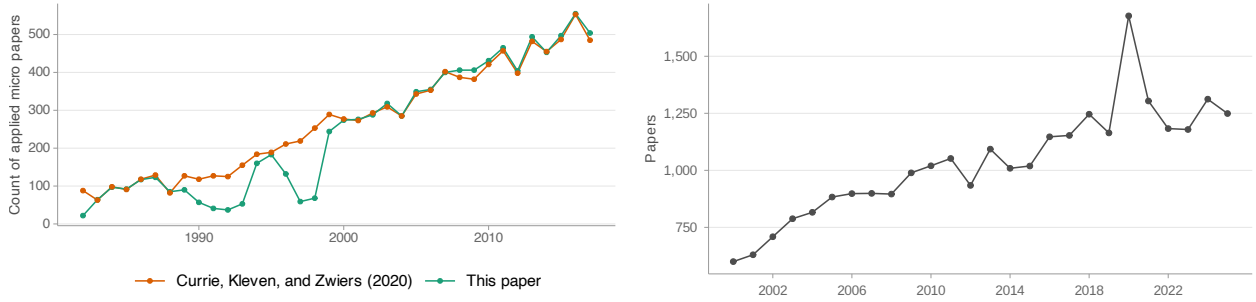
Top Journal Articles. I supplement the NBER data with articles from eleven leading economics and finance journals, covering 2011–2024: three general-interest economics journals (AER, QJE, JPE), the four American Economic Journals (Applied, Policy, Macro, Micro), three top finance journals (Journal of Finance, Review of Financial Studies, Journal of Financial Economics), and the Journal of Econometrics. I extract full text from published PDFs using PyMuPDF. For AER, I filter out Papers and Proceedings (P&P) articles using DOI patterns: issue-based identification for 2011–2014 (when P&P appeared in a designated issue of the AER) and DOI prefix matching for 2015–2017 (when P&P received distinct DOI prefixes); from 2018 onward, P&P papers moved to a separate journal. I exclude the Review of Economic Studies (zero text extraction coverage) and Econometrica (near-zero text coverage for 2011–2014, partial thereafter) from the main analysis; Appendix K documents text extraction rates by journal and year. In total, the journal sample comprises approximately 12,300 articles.

Text Processing. For each paper, I extract the full text, strip out the references section—identified by looking for section headers followed by high concentrations of "Journal" mentions—and apply the keyword search. I use the same keywords and regular expressions as Currie, Kleven, and Zwiers (2020b), with appropriate case sensitivity for each category. The full list is in the Appendix. Full-text search of the papers used in this manuscript are available at `https://paulgp.com/econlit-pipeline/` to allow for alternative keyword searches.

[2] See the Appendix for the full set of keywords. I follow the same method as Currie, Kleven, and Zwiers (2020b).

Validation. I validate keyword matching against two external benchmarks. First, I compare keyword flags to the hand-coded method labels in Brodeur, Cook, and Heyes (2020), matching 357 papers across nine journals (2011–2020) by title. Treating Brodeur et al.'s labels as ground truth, keywords achieve high recall—99% for DiD and IV, 95% for RD—meaning they rarely miss a paper that uses a given method. Precision is lower (69–74% for DiD, IV, and RD), reflecting that keywords also flag papers that mention a method without using it as a primary research design. For RCTs, precision is 67% and recall is 83%, reflecting the diverse terminology economists use for experimental designs. Second, I classify a stratified sample of 750 papers using two independent LLMs (Claude Haiku 4.5 and Qwen 3.5-122B). Both LLMs produce nearly identical positive rates for every method category, and agreement with keywords runs 80–92% for most categories—though agreement is lower for identification strategy (66%) and structural models (69%). Full results appear in Appendix A.

Field Classification. For journal articles, I classify papers into fields using a two-step procedure. First, field-specific journals are directly classified: AEJ Applied and AEJ Policy map to "Applied Micro," AEJ Macro to "Macro," AEJ Micro to "Micro Theory," the three finance journals (JF, JFE, RFS) to "Finance," and the Journal of Econometrics to "Econometrics." Second, for the general-interest journals (AER, QJE, JPE), I use JEL codes when available. Each paper's JEL code first letters determine its field: D, J, L, H, I, Q, R, or K codes map to "Applied Micro"; G codes to "Finance"; E or F codes to "Macro"; and C codes to "Econometrics." When a paper has JEL codes spanning multiple fields, I assign it to the first matching field in the priority order listed above—applied micro takes precedence, then finance, then macro, then econometrics. This reflects the fact that many papers cross boundaries (e.g., a paper with both D and G codes is likely applied work using financial data). Papers without JEL codes—primarily from QJE and JPE, which do not report them—default to "General Econ." In the AER, 87% of papers are classified as Applied Micro via JEL codes, 5.5% default to General Econ, and the remainder split across Finance, Macro, and Econometrics.



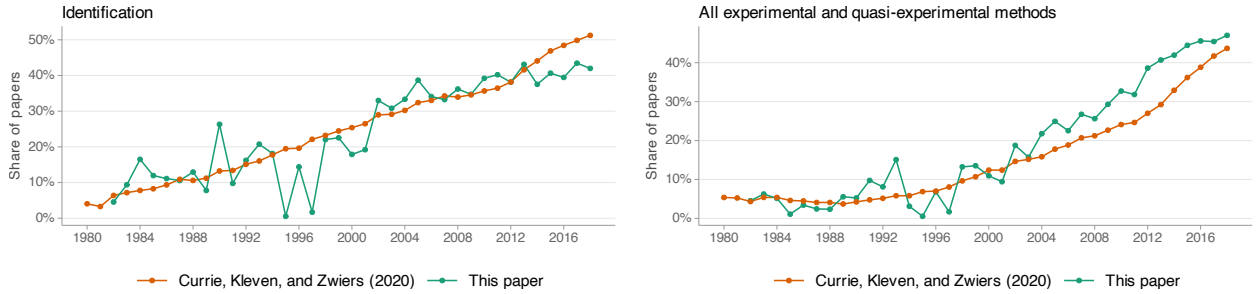
(i) Comparison of sample size to Currie, Kleven, and Zwiers (2020b) in "applied micro"

(ii) Total papers in final sample over time

Figure 1: NBER Working Paper Counts over Time. Data for Currie, Kleven, and Zwiers (2020b) is measured in Appendix Figure B.I. in their paper. My sample ends in early 2025.

As Currie, Kleven, and Zwiers (2020b) note in their replication package (Currie, Kleven, and

Zwiers 2020a), PDF-to-text conversion introduces errors. To see how this affects my sample, I compare paper counts over time in the "applied micro" setting to Currie, Kleven, and Zwiers (2020b) in Figure 1(i). My sample has more gaps in the 1990s—reflecting data processing errors for PDFs in that period—but coverage is close in the early 1980s and from 1999 onwards. Figure 2 provides a more direct check: I compare two headline estimates from Currie, Kleven, and Zwiers (2020b) to mine—the share of papers referencing identification and the share referencing any experimental or quasi-experimental method (RCTs, lab experiments, difference-in-differences, regression discontinuity, instrumental variables, event studies, or bunching). My estimates track well except in the late 1990s. I therefore focus on 2000 onwards for all results, leaving a sample of 24,702 papers. Figure 1(ii) plots the sample over time.



(i) Comparison of identification measure to Currie, Kleven, and Zwiers (2020b) in "applied micro"

(ii) Comparison of all experimental and quasi-experimental measure to Currie, Kleven, and Zwiers (2020b) in "applied micro"

Figure 2: Validation of measurement with Currie, Kleven, and Zwiers (2020b). Data for Currie, Kleven, and Zwiers (2020b) is taken from Figure 2 Panel A and B. I plot the raw (annual) measure, while the Currie, Kleven, and Zwiers (2020b) data is a rolling five-year mean; the smoothing explains the slight visual discrepancy between the two series.

Each NBER working paper can be submitted to one or more of nineteen programs, and 55 percent list more than one.[3] Table 1 reports the breakdown. The three largest programs are Economic Fluctuations and Growth (macroeconomics), Public Economics (applied micro, as classified by Currie, Kleven, and Zwiers (2020b)), and Labor Studies (also applied micro). Development Economics is smaller in part because it began only in 2012.

To compare across programs, I extend Currie, Kleven, and Zwiers (2020b)'s classification. I define "finance" as Asset Pricing and Corporate Finance, and "macro/other" as the remaining programs: Development of the American Economy (the economic history group), Economic Fluctuations and Growth, International Finance and Macroeconomics, Law and Economics, Monetary Economics, and Productivity, Innovation, and Entrepreneurship. Table 2 defines these groupings. One difference from Currie, Kleven, and Zwiers (2020b) is worth noting: they define applied micro as papers that *solely* list applied micro programs, making it a paper-specific label. I instead use non-exclusive labels—if a paper is in both finance and applied micro, it counts in both categories. The overlap

[3] 45 percent have one program, 32 percent have two, 15 percent have three, 5 percent have four, and 2 percent have five.

NBER Program	Number of Papers
Applied Micro	
Labor Studies	5,970
Public Economics	5,896
Economics of Health	3,641
International Trade and Investment	2,466
Children and Families	2,193
Industrial Organization	2,160
Economics of Education	2,105
Development Economics	1,955
Political Economy	1,869
Environment and Energy Economics	1,724
Economics of Aging	1,698
Finance	
Asset Pricing	2,985
Corporate Finance	2,785
Macro/Others	
Economic Fluctuations and Growth	5,645
International Finance and Macroeconomics	3,107
Monetary Economics	2,924
Productivity, Innovation, and Entrepreneurship	2,785
Development of the American Economy	1,675
Law and Economics	1,385

Table 1: NBER Working Paper Series counts by program

Field Group	Number of Papers
Applied Micro	18,288
Macro/Others	5,111
Finance	1,758
Finance + Macro/Others	1,692

Table 2: Breakdown of papers by field groupings

matters but is not extreme: 44 percent of papers are exclusively applied micro, seven percent exclusively finance, and 19 percent exclusively macro/other. The most common cross-field pairings are applied micro and macro/other (19%) and finance and macro/other (6%). Appendix D reports all main results under an exclusive classification (papers assigned to a single field); the cross-field gaps are qualitatively similar and, if anything, slightly wider.

Throughout the analysis, field and program labels are non-exclusive: a paper contributes to every program to which it is submitted. The full corpus includes approximately 31,500 NBER papers and 12,300 journal articles. I focus on 2000 onwards for most results, leaving a sample of approximately 24,700 NBER papers (after excluding papers classified as "Other"). The full historical sample is used when comparing with Currie, Kleven, and Zwiers (2020b).

Table 3 provides a snapshot of the headline numbers. For each field, I report the share of papers mentioning identification, experimental or quasi-experimental methods (with and without DiD), and DiD specifically, both for 2016–2024 and for the 2000–2015 average. These numbers summarize the main patterns documented in the paper.

	2016–2024					2000–2015				
Field	*N*	Ident.	Exp./QE	DiD	Excl. DiD	*N*	Ident.	Exp./QE	DiD	Excl. DiD
Applied Micro	8,265	40.2%	58.3%	25.3%	45.8%	9,067	33.4%	42.9%	11.8%	37.1%
Finance	586	22.7%	35.8%	19.8%	23.2%	1,121	15.1%	22.5%	10.9%	14.2%
Macro/Others	2,514	25.1%	29.7%	12.5%	21.8%	4,047	17.6%	22.0%	6.3%	17.7%

Table 3: Summary of credibility revolution measures by field. Shares are computed from NBER working papers. "Exp./Quasi-exp." includes DiD, event studies, IV, RD, RCTs, lab experiments, and bunching. "Exp./QE excl. DiD" excludes difference-in-differences and event studies.

2 Results from NBER Working Papers

2.1 Overall trends

Figure 3 presents the updated version of Currie, Kleven, and Zwiers (2020b)'s Figure 2, now covering all NBER papers through May 2024.[4] Each panel shows field-specific trends as colored lines, with the overall aggregate as a dashed black line.

Nearly all trends continue in the direction Currie, Kleven, and Zwiers (2020b) documented. The share of papers explicitly mentioning identification has risen overall, with growth slowing markedly since 2016 (panel a).[5] The share mentioning any experimental or quasi-experimental method, by contrast, has continued to rise even after 2016 (panel b). This means identification language has saturated while mentions of specific methods keep growing. Administrative data (panel c) has also continued its upward trend.

But the aggregate trends mask substantial heterogeneity. Figure 3 previews the paper's central finding: mentions of credibility revolution methods have spread unevenly, with persistent gaps between applied micro on the one hand and finance and macro/other on the other. The next two subsections document these gaps in detail and decompose them by method and program.

Graphical revolution trends (figures-to-tables ratio) are reported in Appendix B.

2.2 Comparison across fields

Figure 3 splits each variable by the three field groupings defined in Table 2. The gaps are large and persistent. For identification, experimental and quasi-experimental methods, and administrative data, applied micro is well above both finance and macro/other. Applied micro's identification share has grown more slowly since 2017, reaching 46 percent by 2024, and remains 13–17 percentage points above finance and macro/other. For experimental and quasi-experimental methods (panel b), applied micro reaches 63 percent by 2024, while finance stands at 47 percent and macro/other at 39 percent (Table 3). All time-series figures use two-year moving averages, which smooth year-to-year fluctuations; averages for 2016–2024 are reported in Table 3.

[4] Currie, Kleven, and Zwiers (2020b) use a five-year moving average; I present two-year moving averages throughout.

[5] The keywords are listed in the Appendix, as well as in the Appendix of Currie, Kleven, and Zwiers (2020b). For example, matches for identification flag phrases like "identification assumption" and "causal identification."

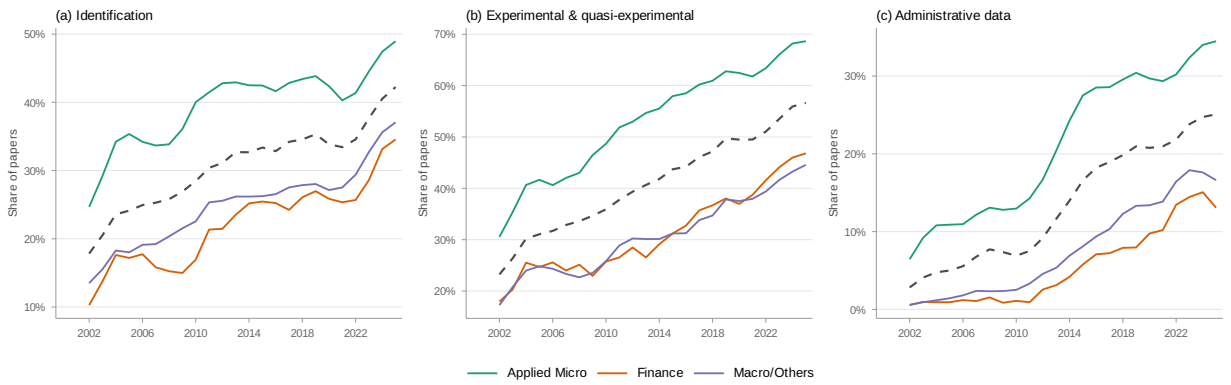


Figure 3: Credibility revolution trends in NBER working papers (two-year moving averages). Colored lines show field-specific trends; dashed black line shows the overall aggregate. See Table 2 for field definitions and the Appendix for keyword definitions.

To put these gaps in context, it helps to ask where finance and macro/other stand *today* relative to applied micro in the *past*. In identification, the current levels in finance and macro/other are comparable to where applied micro was in 2008–2010. In experimental and quasi-experimental methods, finance is comparable to applied micro circa 2011–2012 and macro/other to applied micro circa 2008. In administrative data, they are comparable to applied micro circa 2013. Whether this reflects a lag that will eventually close or different long-run equilibria—driven by differences in available data and the nature of the research questions—is an important open question.

Figure 4 presents method-specific trends by field. I start with difference-in-differences (panel a), which includes event studies. All three fields show steep growth, with applied micro leading. Finance is close behind—partly because the term "event study" captures financial event studies (abnormal return studies) that differ methodologically from DiD-style event studies. Appendix C decomposes this measure into strict DiD language and event study mentions, and identifies papers combining "event study" with "abnormal return" as a proxy for financial event studies. While financial event studies account for a meaningful share of finance's event study mentions, they do not explain the full convergence: finance's strict DiD share has also grown substantially.

Panel (b) tells a different story: synthetic controls. In Appendix Figure A.V of Currie, Kleven, and Zwiers (2020b), synthetic control was growing rapidly as of 2018 among applied micro papers. That growth continued through 2020 but has since leveled off. Even when including "synthetic DiD" methods (Arkhangelsky et al. 2021) and related variants, the combined measure shows that synthetic control mentions have plateaued—a notable contrast to DiD's continued rise. Appendix H decomposes classic synthetic control from synthetic DiD mentions.

I also examine Bartik and shift-share instruments (Goldsmith-Pinkham, Sorkin, and Swift 2020; Borusyak, Hull, and Jaravel 2022; Adão, Kolesár, and Morales 2019) (panel c). Since 2013, this method has grown rapidly across all fields, though with some decline in macro/other and finance after 2021. As of 2024, 2–4 percent of papers mention Bartik or shift-share. To see what this means in context, panel (d) plots the share mentioning instrumental variables at all. That share has stayed

roughly constant over time—at about 25 percent for applied micro, and roughly 18 percent for both finance and macro/other. A back-of-the-envelope calculation suggests that roughly 14–24 percent of papers mentioning IV also mention Bartik or shift-share instruments, with the share higher in applied micro than in finance.

Turning from instruments to experiments and regression discontinuity: in panel (e), applied micro leads in RCT mentions, with 20 percent of papers by 2024. Finance and macro have also grown, but more slowly. The divergence is striking because all three fields started from a similar base as of 2003—this means the gap is entirely a post-2003 phenomenon. In panel (f), applied micro leads finance and macro/other by about 7–8 percentage points in regression discontinuity mentions as of 2024, but the share has flattened across all fields over the past eight years. RD appears to have reached a natural ceiling.

What accounts for the gap between applied micro and the other fields? Some papers may be pure theory or observational work. Another possibility—already measured by Currie, Kleven, and Zwiers (2020b)—is structural estimation (keywords include "structural model," "structural general equilibrium model," and "GMM"). In Figure 5(i), macro/other and finance have a 7.5–10 percentage point higher share of structural estimation mentions, consistent with the greater prevalence of structural models in these fields (though applied micro includes Industrial Organization, which also relies heavily on structural methods). Because the broad measure includes GMM and MLE—which appear in many non-structural contexts—Appendix F also reports a narrow measure excluding these terms. The cross-field pattern is qualitatively similar under both definitions.

More revealing is Figure 5(ii), which isolates papers that mention structural estimation *without* also mentioning experimental or quasi-experimental methods. Here the gap widens: finance and macro/other papers are roughly twice as likely to fall in this category as applied micro papers as of 2024. This means that when applied micro papers use structural models, they typically pair them with complementary research designs—a pattern far less common in finance and macro. Structural estimation thus accounts for part of the cross-field gap: finance and macro have more papers that rely on structural models alone, without the identification strategies that characterize the credibility revolution. More broadly, part of the gap reflects denominator composition: finance and macro/other have a larger share of purely theoretical papers. Conditioning on papers that mention at least one empirical method or data source narrows the gap—the experimental/quasi-experimental share rises to 76 percent for applied micro, 65 percent for finance, and 61 percent for macro/other in 2024—but the gap clearly persists (Appendix G).

2.3 Breakdown across programs

The field-level averages mask important within-field variation. Within finance, asset pricing and corporate finance produce very different types of empirical work. The same is true within applied micro and macro/other. To see what is driving the field-level trends, I disaggregate to individual NBER programs.

Figure 6 plots the share of papers mentioning identification and experimental/quasi-experimental

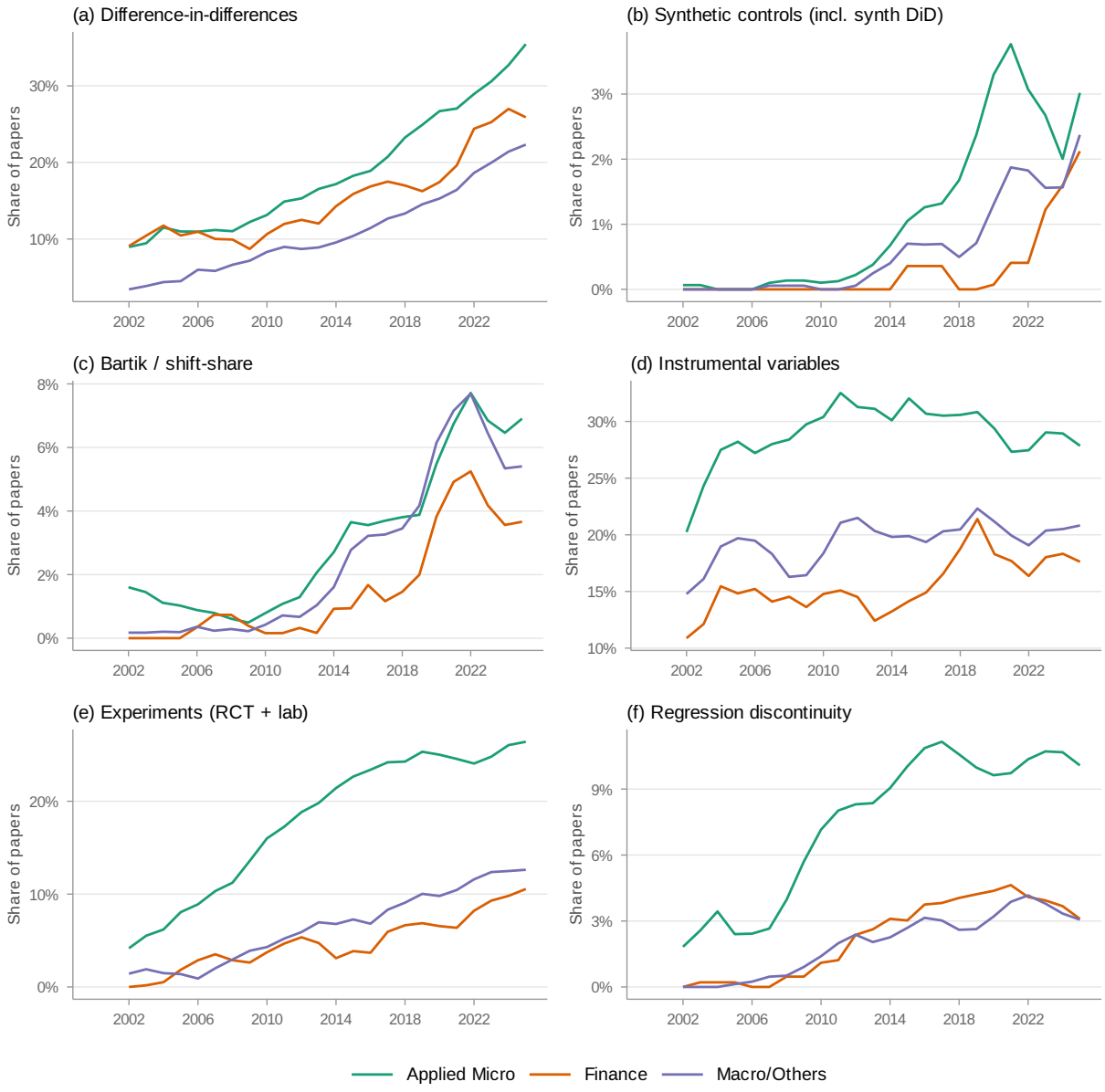
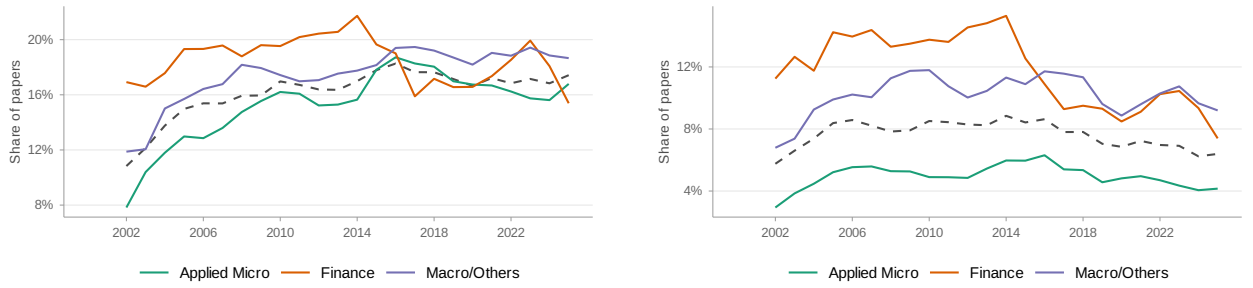


Figure 4: Method-specific trends by field (two-year moving averages). Note: y-axis ranges differ across panels to accommodate different prevalence levels. See Table 2 for field definitions and the Appendix for keyword definitions.



(i) Structural Models

(ii) Structural Models without mention of experimental or quasi-experimental methods

Figure 5: Panel (a) reports the share of papers that mention structural model estimation. Panel (b) reports the share of papers that mention structural model estimation and do not mention any form of experimental or quasi-experimental methods. Colored lines show field-specific trends; dashed black line shows the overall aggregate. See Table 2 for the breakdown of fields, and the Appendix for definitions on keywords.

methods across all nineteen programs using slope charts. Each line segment connects a program's 2000–2015 share (left) to its 2016–2024 share (right), colored by field. The slope and level of each segment simultaneously reveal where programs stand and how much they have changed.

Despite within-field heterogeneity, the cross-field pattern is strikingly consistent. In panel (a), applied micro programs have higher identification shares than nearly all finance and macro/other programs, with the exceptions of Productivity, Innovation, and Entrepreneurship and Law and Economics. Within finance, there is a large gap between Asset Pricing and Corporate Finance—this means the rise of identification in finance is driven primarily by Corporate Finance. In panel (b), the pattern is similar: applied micro programs lead, with only Law and Economics and Corporate Finance among the non-applied-micro programs reaching comparable levels. Identification mentions grew at broadly similar rates across programs, while growth in experimental and quasi-experimental methods varies sharply: International Finance and Macroeconomics, Economic Fluctuations and Growth, and Asset Pricing have seen markedly less growth than Corporate Finance and Children.

Which methods have driven the growth? Figure 7 presents a heatmap of the change in method share by program between 2000–2015 and 2016–2024. The answer is unambiguous: DiD accounts for most of the growth across programs. The share of papers mentioning instrumental variables has stayed roughly constant. Regression discontinuity has risen only slightly.[6] Experiments have risen across programs, with the largest growth in Law and Economics, Political Economy, and Productivity, Innovation, and Entrepreneurship.

The pattern for finance and macro is clear: growth in credibility revolution methods has been concentrated in DiD. The credibility revolution in these fields has been, to a first approximation, a difference-in-differences revolution. An important caveat is that keyword matching cannot distinguish between papers using traditional two-way fixed effects specifications and those adopting the robust estimators developed in the recent methodological literature (Callaway, Goodman-Bacon,

[6]Most of the growth in RD is concentrated in Public Economics, Economics of Education, and Children. Education is perhaps unsurprising given the prominence of test-score cutoffs as a canonical RD design.

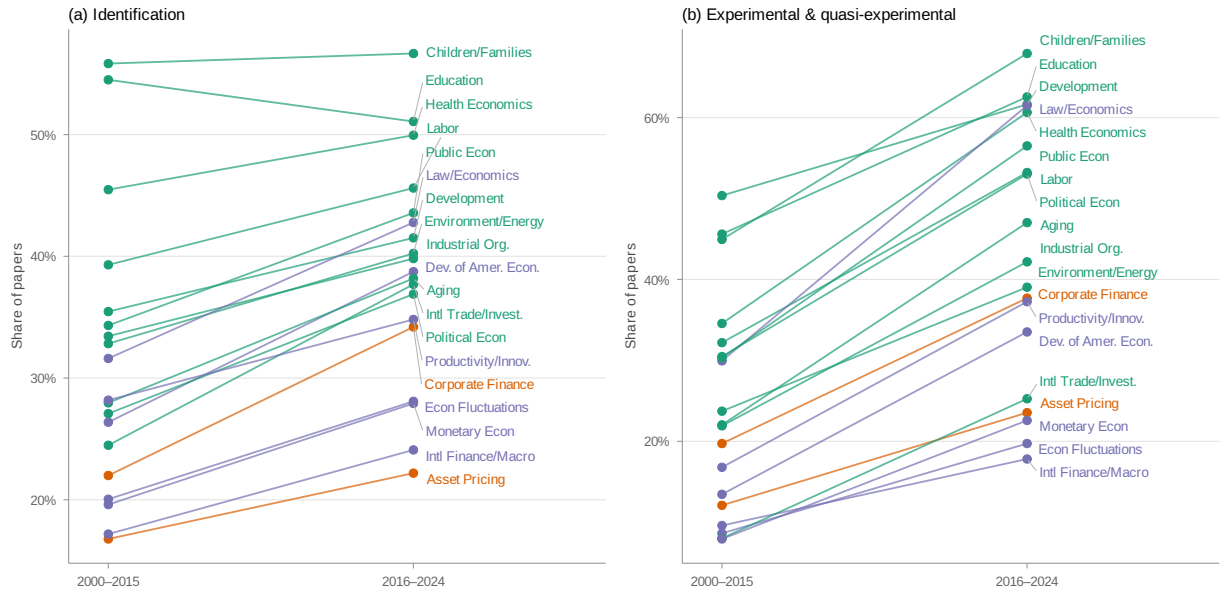


Figure 6: Method mentions across NBER programs. Each line segment connects a program's 2000–2015 share (left) to its 2016–2024 share (right), colored by field. Steeper upward slopes indicate faster growth. See Table 2 for field definitions.

and Sant'Anna 2024; De Chaisemartin and d'Haultfoeuille 2020; Sun and Abraham 2021). This limits what we can say about the *quality* of DiD adoption, even as we measure its *quantity*. It is possible that late-adopting fields have leapfrogged to better implementations—a silver lining of the adoption lag.

Bartik and shift-share mentions have grown across all programs, but most sharply in International Trade and Investment—where nearly 10 percent of papers now mention them—followed by Development of the American Economy and Labor Studies.[7] Synthetic control mentions have also grown, with the largest increases in Law and Economics, Health Economics, and Children.

2.4 The dominance of difference-in-differences across fields

How much does this single method account for the overall growth in experimental and quasi-experimental methods? A simple exercise makes the answer concrete. Figure 8 compares method shares with and without DiD.

Panel (a) breaks down the comparison by field. Over the 2016–2024 period, including DiD raises finance's methods share by roughly 13 percentage points—a 56 percent increase—compared to a similar 13 percentage point increase for applied micro, which represents only a 29 percent increase because applied micro's baseline is much higher. For macro/other, the gap is smaller (about 8 percentage points, a 37 percent increase). Finance's methodological transformation is, to a striking

[7] This likely reflects the use of shift-share instruments in studying the China Shock (Autor, Dorn, and Hanson 2013), historical migration (Boustan 2010), and the long tradition of Bartik instruments in labor studies (Bartik 1991).

Figure 7: Change in method-specific mentions across NBER programs, 2016–2024 minus 2000–2015. Each cell shows the percentage-point change in share. Blue indicates growth, red indicates decline. See Table 2 for field definitions.

degree, a DiD story.

Panel (b) decomposes the percentage increase by program for 2016–2024. International Finance and Macroeconomics shows the largest increase, followed by Corporate Finance, Health Economics, and Asset Pricing. By contrast, applied micro programs with high overall method shares—such as Development Economics and Education—show relatively small increases from DiD. This reflects their diversified methodological portfolios: these programs rely on a mix of RCTs, RD, and other designs, not just DiD. Importantly, these program-level results are unlikely to be driven by financial event studies—which are methodologically distinct from DiD-style event studies—since the effect is large even for programs like Asset Pricing and Corporate Finance where the abnormal-return event study tradition is strongest. Appendix C decomposes the composite DiD measure into strict DiD language and event study mentions.

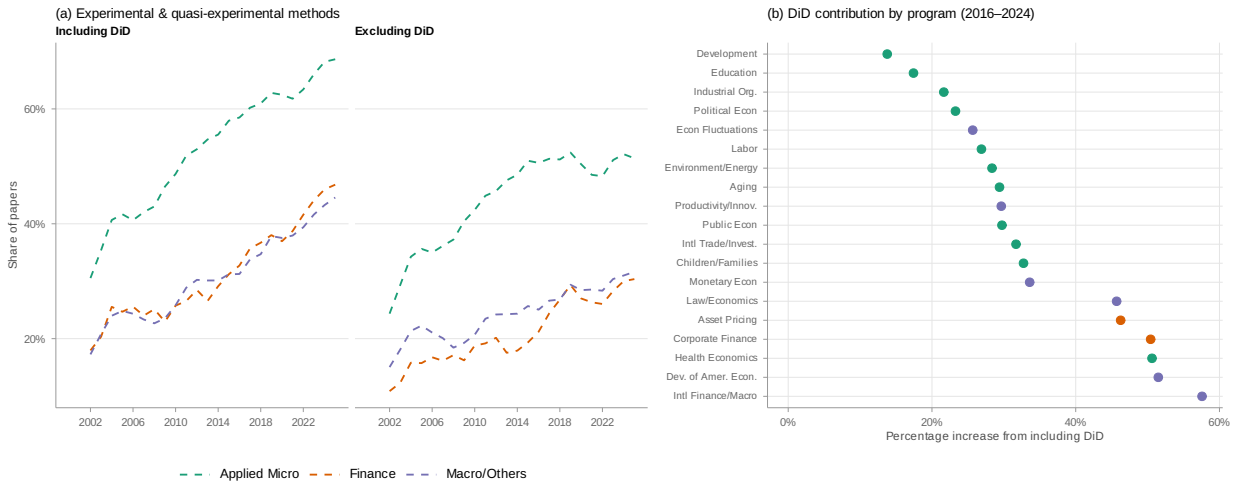


Figure 8: The dominance of difference-in-differences. Panel (a): experimental and quasi-experimental method shares by field, faceted by whether DiD is included or excluded. Panel (b): percentage increase in method share from including DiD, by NBER program (2016–2024). See Table 2 for field definitions.

3 Evidence from Top Journals

The NBER working paper series is a natural laboratory for studying methodological trends, but it has a limitation: NBER affiliates are a selected group, and working papers may differ systematically from published articles. Do the patterns above survive in a different sample? I examine published articles from top economics and finance journals. This analysis serves two purposes: it provides independent confirmation of the NBER trends, and it reveals whether the publication process amplifies or attenuates the credibility revolution's reach.

3.1 Overall trends across top journals

Figure 9 compares identification and DiD mentions across journal fields between 2011–2017 and 2018–2024. The field-level patterns closely mirror the NBER data. Applied micro journals show the highest rates of identification language and experimental/quasi-experimental methods, followed by finance, with macro trailing. The same core asymmetry from Section 2 appears in the published literature. One important caveat: the journal-level field classification for macro is considerably noisier than for the NBER sample. In the NBER data, macro is defined by program affiliation; in the journal data, macro is represented primarily by AEJ Macroeconomics plus JEL-classified papers from the AER. QJE and JPE do not report JEL codes, so macro papers in those journals default to "General Econ" and are excluded from the field comparison. The journal-level macro trends should therefore be interpreted with more caution than the finance or applied micro trends, which benefit from dedicated field journals. Appendix J shows that the cross-field patterns hold when restricting to field-specific journals only.

The DiD pattern (panel b) is notable: applied micro journals show the steepest growth in DiD mentions (from 27% to 41%), while finance shows steady but more modest growth (from 25% to 29%), confirming the NBER finding that the credibility revolution in finance remains substantially behind applied micro. Full time series for all four method categories, including instrumental variables and structural models, are reported in Appendix I.

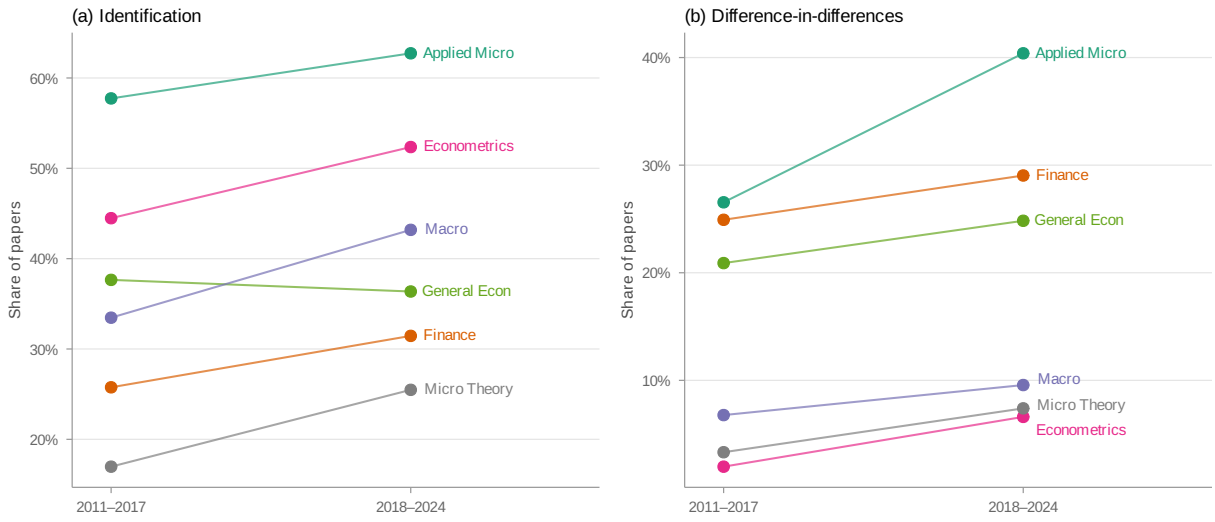


Figure 9: Method mentions across fields in top journals: 2011–2017 vs. 2018–2024. Each line connects a field's early-period share (left) to its late-period share (right). Panel (a) identification, (b) difference-in-differences. See text for journal-to-field mapping.

3.2 Comparison across individual journals

Figure 10 compares mention rates across individual journals. AEJ Applied Economics and AEJ Economic Policy show the highest rates of credibility revolution methods—unsurprising given their

explicit focus on applied empirical work. Among the general-interest journals, AER and QJE show higher rates than JPE, reflecting differences in paper composition. The finance journals show moderate adoption of DiD and identification language but lower rates of RD and experimental methods—echoing the NBER findings at the journal level.

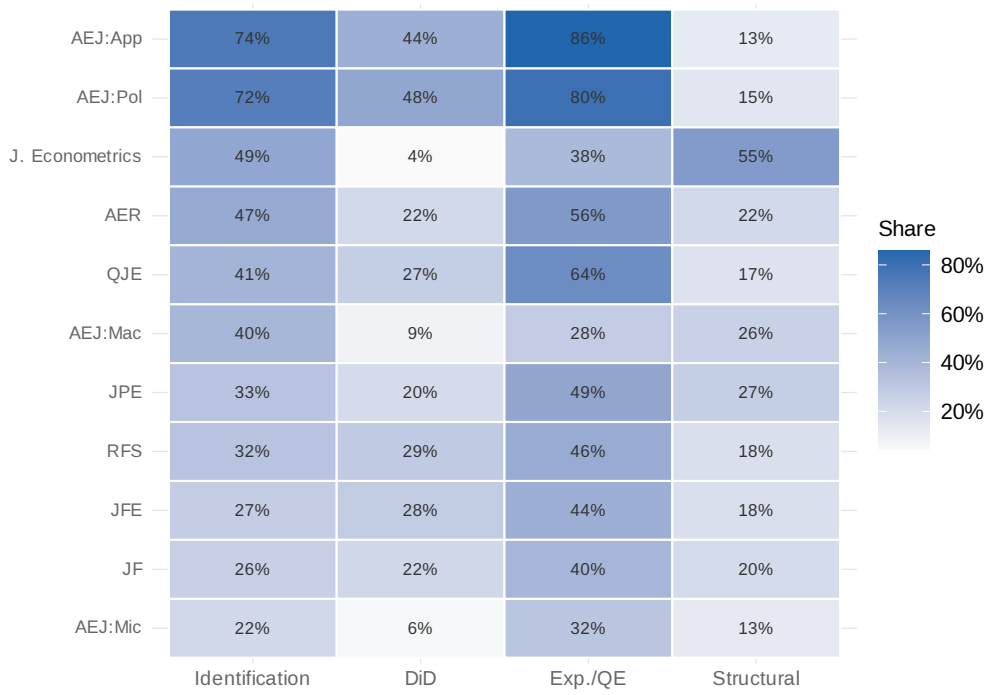


Figure 10: Heatmap of method mentions across individual journals (2011–2024). Color intensity reflects the share of papers mentioning each method category. Journals ordered by identification share.

3.3 NBER working papers vs. published articles

Could the NBER trends be artifacts of the working paper selection process? Figure 11 overlays the NBER and journal time series for key methods, matching by field. The trends are strikingly similar. Published articles show slightly higher rates of most credibility revolution methods—consistent with a selection effect where papers using transparent research designs are more likely to clear the bar at top journals. The convergence across these two independent samples strengthens confidence that the patterns in Section 2 reflect real changes in how economists do empirical work.

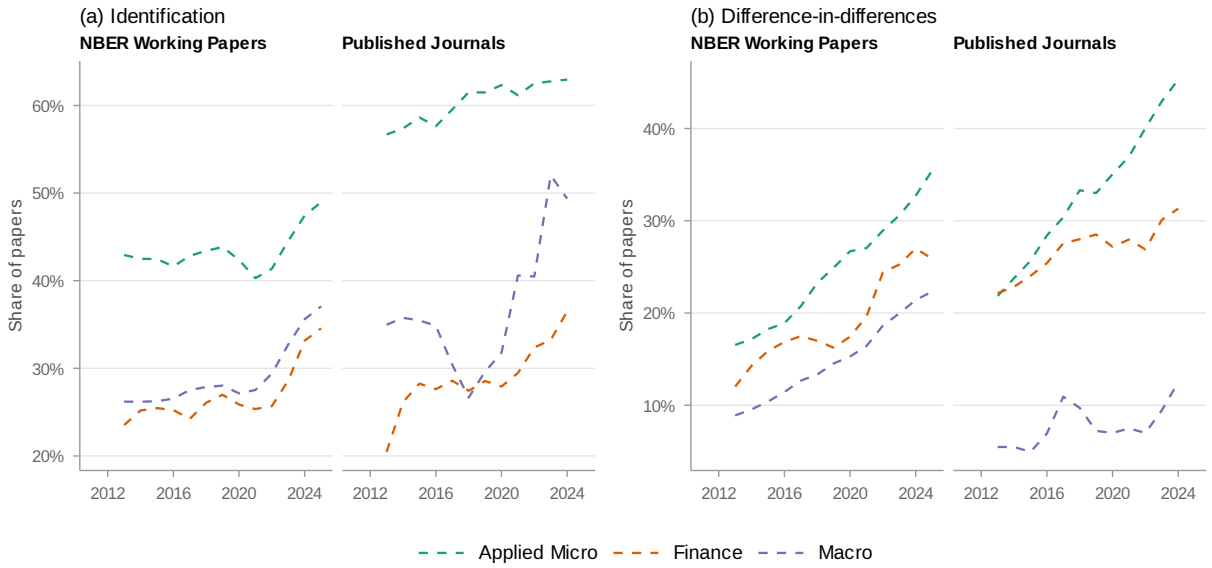


Figure 11: NBER working papers vs. published journal articles: time series by field (2011–2024). Each panel compares NBER working papers (left facet) with published journals (right facet) for identification (left) and difference-in-differences (right).

4 Econometric Theory and Applied Practice

Having established that the credibility revolution has spread unevenly across applied fields, I now turn to a deeper question. The credibility revolution depends on tools developed by econometricians—DiD estimators, IV techniques, RD designs, synthetic control methods—all of which required substantial theoretical development. If the revolution's reach has been uneven across *applied* fields, what about the field that supplies its theoretical infrastructure?

To answer this, I compare term prevalence in the *Journal of Econometrics* (2,318 papers, 2011–2024) with that in applied economics and finance journals. I begin with a natural question: of the methods that applied researchers care about, what share of *Journal of Econometrics* papers touch on them? Panel (a) of Figure 12 shows the answer. Most credibility revolution methods—DiD, event studies, RD, RCTs, administrative data, synthetic control, Bartik instruments, binscatter, and heterogeneous treatment effects—appear far less frequently in the *Journal of Econometrics* than in applied journals. DiD appears in approximately 19% of applied journal papers but under 4% of *Journal of Econometrics* papers; event studies show a similar gap. The exceptions are identification language and instrumental variables, where the *Journal of Econometrics* matches or exceeds applied journals—reflecting the theoretical literature on identification and IV estimation that is a core focus of the journal.

This raises a natural follow-up: if the *Journal of Econometrics* is not primarily publishing on applied methods, what *does* it publish? Panel (b) takes a data-driven approach. I start by constructing keyword lists for twenty candidate topic areas in econometric theory, drawn from the major sections of standard graduate econometrics textbooks and the *Journal of Econometrics*' own

subject classifications: nonparametric and semiparametric estimation, time series models (VAR, GARCH, cointegration, unit root), Bayesian methods, bootstrap and resampling, machine learning and regularization, panel data methods, limited dependent variable models, quantile regression, kernel methods, forecasting, robust inference, weak identification, simulation/Monte Carlo, and asymptotic theory.[8] I then compute the prevalence of each category in the *Journal of Econometrics* and in applied journals, rank by *Journal of Econometrics* prevalence, and show the top fifteen alongside the corresponding rates in applied journals. Composite categories and narrow variants are excluded to avoid double-counting.[9] Asymptotic theory and Monte Carlo simulation top the list—appearing in 86% and 65% of papers respectively—but these reflect the standard toolkit for deriving and validating estimators; applied papers rely on asymptotic theory implicitly even when they do not use the term. The more informative contrasts involve substantive methods: nonparametric estimation (58%), time series models (54%), structural/GMM/MLE methods (54%), and Bayesian methods all appear at far higher rates in the *Journal of Econometrics* than in applied journals. These are the estimation and inference techniques that form the theoretical infrastructure of econometrics—important in their own right, but distant from the day-to-day practice of most applied economists. The gap between the two panels illustrates how the *Journal of Econometrics* and applied journals occupy largely separate methodological spaces.

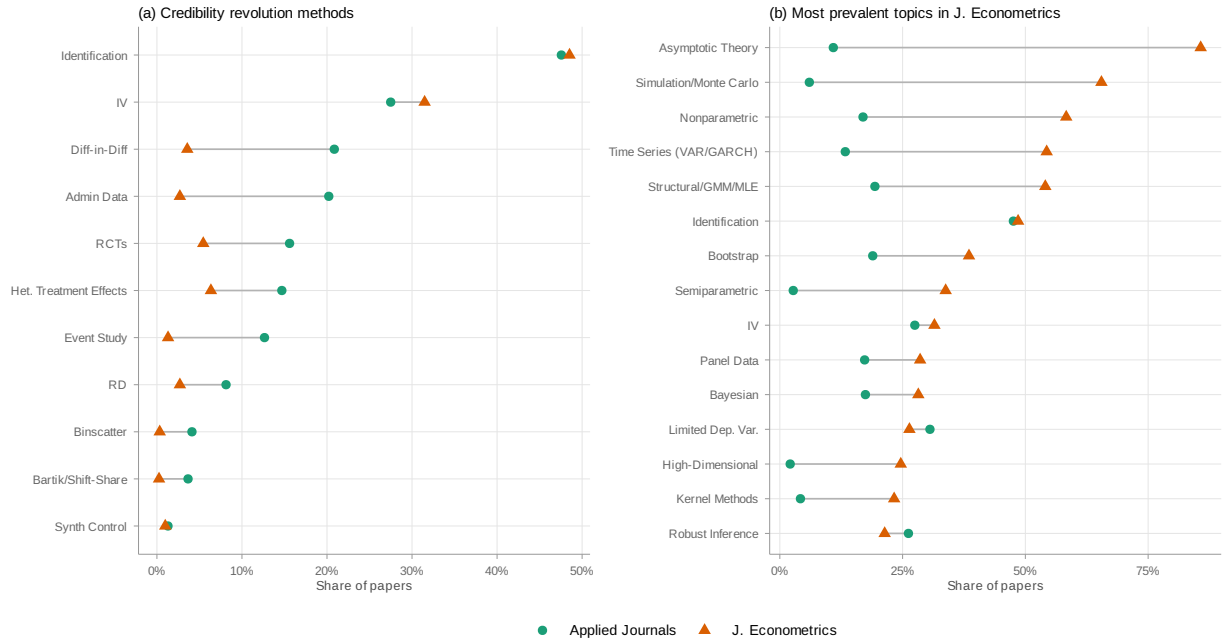


Figure 12: Comparison of term prevalence: *Journal of Econometrics* vs. applied economics journals (2011–2024). Panel (a) shows credibility revolution methods; panel (b) shows the fifteen most prevalent topics in the *Journal of Econometrics*, ranked by share.

[8]The full list of trigger phrases for each category appears in the Appendix. The core applied categories follow Currie, Kleven, and Zwiers (2020b); the econometrics categories are new to this paper.

[9]I exclude composites like "all experiments" (which unions DiD, RCTs, etc.) and narrow variants like "synth DiD" (subsumed by "synth control").

Figure 13 makes the full picture concrete through a heatmap of key terms across all journals. The *Journal of Econometrics* has a strikingly different methodological profile from every other journal in the sample—high rates of asymptotic theory and simulation (reflecting the standard tools for proving and testing estimators), along with nonparametric estimation, time series models, and Bayesian methods, but low rates of DiD and event studies. Appendix L shows that credibility revolution methods have grown within the *Journal of Econometrics* over the sample period, suggesting the gap may be narrowing.

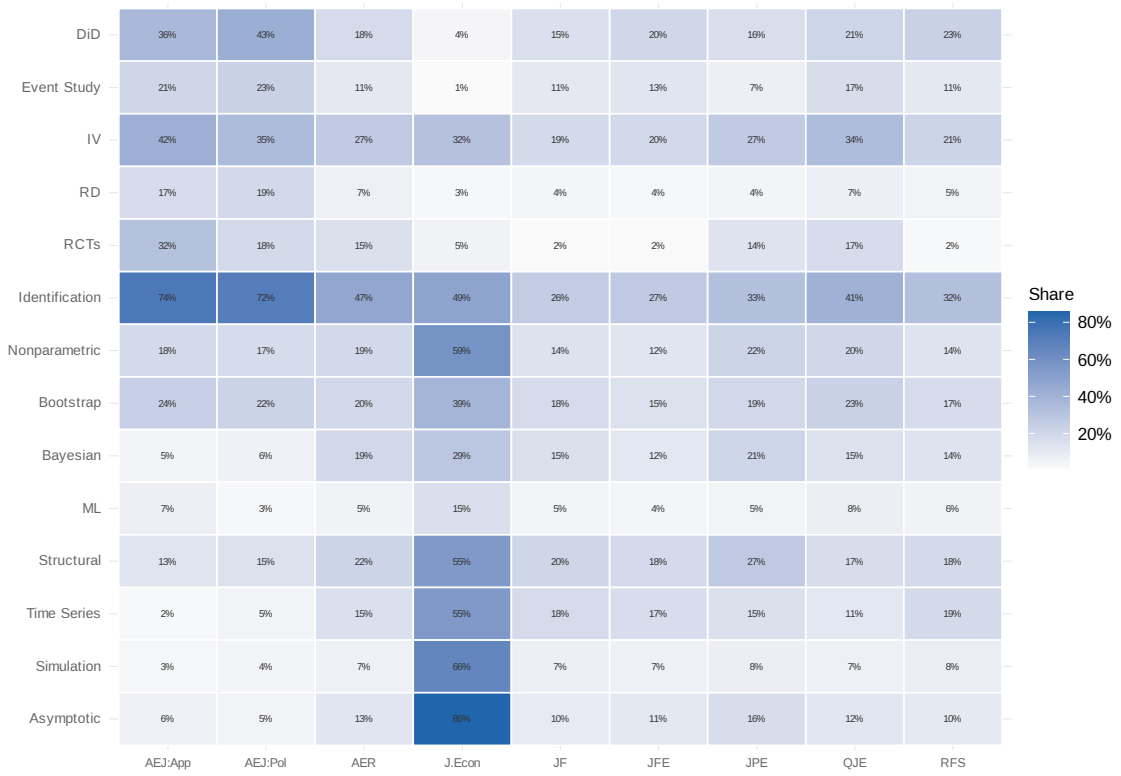


Figure 13: Heatmap of method term prevalence across journals (2011–2024). Color intensity reflects the share of papers mentioning each term.

Three caveats are important.

First, the *Journal of Econometrics* has been at its most influential when it engages directly with the credibility revolution's tools. The literatures on heterogeneous treatment effects (De Chaisemartin and d'Haultfoeuille 2020; Callaway, Goodman-Bacon, and Sant'Anna 2024), staggered DiD (Roth 2022; Rambachan and Roth 2023), and machine learning for causal inference have reshaped applied practice—though many of these contributions appeared in other outlets, not the *Journal of Econometrics* itself. Contributions of this type that do appear in the journal represent a relatively small slice of what it publishes. Panel (b) of Figure 12 shows that the bulk of the journal focuses on nonparametric estimation, time series models, and Bayesian methods—problems important in their own right but distant from the day-to-day practice of applied economics. This suggests an opportunity: the journal's outsized impact on applied work, when it chooses to engage with applied

tools, makes it a natural venue for bridging the gap documented here.

Second, the gap could reflect productive intellectual specialization rather than misalignment, with applied and theoretical work advancing on parallel tracks that occasionally intersect. Methods journals may be developing tools that will eventually diffuse to practice—as happened with the DiD robustness literature, which moved from econometric theory to widespread applied adoption in under five years.

Third, the cross-field differences documented in Sections 2 and 3 should not be read as implying that all fields *should* converge to the applied micro toolkit. Many questions in macroeconomics are fundamentally about general equilibrium, and the applied micro toolkit—built around partial equilibrium and local treatment effects—may not be the right tool for every setting. The same is true in asset pricing, where the object of interest is often an equilibrium price rather than a treatment effect. The more relevant distinction is between fields where quasi-experimental methods are feasible but underused—corporate finance, for example, has abundant natural experiments—and fields where the questions themselves call for different approaches. Nakamura and Steinsson (2018) offer a thoughtful example of how credibility revolution thinking can be adapted to macroeconomic settings without simply importing the applied micro playbook.

Why does this gap matter? Because the rare instances where the two literatures *do* intersect have been extraordinarily productive. The DiD robustness literature—Callaway, Goodman-Bacon, and Sant'Anna (2024), De Chaisemartin and d'Haultfoeuille (2020), Sun and Abraham (2021)—moved from the pages of econometrics journals to widespread applied adoption in under five years, fundamentally changing how researchers implement one of the most common research designs. The treatment effects literature in the *Journal of Econometrics* has had similar impact. These examples demonstrate that the gap is not inherent: when econometricians engage with the tools applied researchers actually use, the payoff for both sides is large. The gap documented here thus represents an opportunity, not just a description.

This mirrors the cross-field patterns documented in Sections 2 and 3 at a different level of analysis. Just as finance and macro differ from applied micro in their adoption of credibility revolution methods, the econometrics literature differs from applied practice in its methodological focus. Tracking the diffusion of new methods from econometric theory to applied practice—and identifying which theoretical innovations eventually reshape practice—is a natural direction for future work.

5 Conclusion

The credibility revolution has continued to advance, but the picture is one of uneven progress rather than uniform transformation. Three patterns stand out.

First, credibility revolution methods remain most prevalent in applied microeconomics. Finance and macro/other have made real strides since the early 2000s, but they differ substantially from applied micro on most measures—with current levels comparable to where applied micro was roughly a decade ago. Within finance, growth is concentrated in corporate finance; within macro/other, there is wide variation across programs. Whether these gaps reflect a lag that will close over time

or different long-run equilibria—driven by the nature of the research questions and the available data—is an important open question.

Second, outside applied micro, the credibility revolution has been—to a first approximation—a difference-in-differences revolution. Over the 2016–2024 period, including DiD raises the finance methods share by roughly 55 percent versus 30 percent for applied micro. Other quasi-experimental approaches—regression discontinuity, RCTs, instrumental variables—have seen far less growth in finance and macro. This concentration on a single method is noteworthy given the recent econometrics literature highlighting important sensitivities in DiD designs—most prominently, the work of Callaway, Goodman-Bacon, and Sant'Anna (2024) and De Chaisemartin and d'Haultfoeuille (2020) showing that conventional two-way fixed effects estimators can be severely biased under treatment effect heterogeneity, and the growing literature on robust alternatives (Roth 2022; Roth and Sant'Anna 2023; Rambachan and Roth 2023; Chaisemartin et al. 2022; Sun and Abraham 2021). The rapid diffusion of these methodological refinements reflects the productive interplay between econometric theory and applied practice—and suggests that the concentration on DiD may be less concerning if practitioners are adopting improved estimators alongside the research design itself. Notably, synthetic control methods—which share many properties with DiD—have not experienced comparable growth and may even be declining.

Third, this pattern extends to the boundary between econometric theory and applied practice. The *Journal of Econometrics* is dominated by nonparametric estimation, time series models, and Bayesian methods—alongside the asymptotic theory and simulation tools used to develop and validate estimators. The applied literature runs on DiD, identification strategies, and administrative data. These two literatures occupy largely separate methodological spaces, though the gap may partly reflect productive specialization.

These patterns hold across data sources: NBER working paper trends are confirmed by published articles from eleven top journals, and LLM-based classification validates the keyword approach at 80–92 percent agreement rates for most method categories, though agreement is lower for identification strategy and structural models (Appendix A).

Looking ahead, the dominance of DiD raises a question about the trajectory of the credibility revolution. The revolution's early promise was methodological pluralism—a toolkit of transparent research designs, each suited to different empirical settings. The data show that this pluralism has been more fully realized in applied micro than elsewhere. As finance and macroeconomics continue to adopt credible methods, there is value in diversifying beyond DiD, both to strengthen the robustness of individual studies and to expand the set of questions these fields can credibly address.

What should researchers in finance and macro take from these patterns? The answer depends on the setting. In corporate finance, where natural experiments and policy variation are often available, the gap likely reflects unrealized potential: there are many settings where DiD, RD, or IV could be applied but have not yet been. In asset pricing and general equilibrium macro, the gap may partly reflect the nature of the questions—structural models are sometimes the right tool, not a methodological shortcoming. The useful question is not "why doesn't macro look like applied

micro?" but rather "are there settings in these fields where credible research designs would sharpen inference?"

One limitation of this analysis is that keyword mentions measure the *diffusion of methodological language* but not the *quality of adoption* or *influence* of methods. Validation against LLM classification (Appendix A) shows that keyword precision varies across categories—exceeding 90% for regression discontinuity and lab experiments, but falling below 50% for DiD and event studies, where many mentions reflect discussion rather than use as a primary research design. Cross-field comparisons should therefore be interpreted with caution for categories where precision is lowest, as some of the measured gap may reflect differences in vocabulary rather than uptake. A complementary approach would track citations to foundational credibility revolution papers—Angrist and Krueger (1991), Angrist and Pischke (2009), Imbens and Lemieux (2008)—across fields. If finance and macro cite these works at comparable rates but describe methods differently, the measured gap would partly reflect writing conventions rather than substantive methodological differences. If citation rates also differ, this would reinforce the keyword evidence. This is a natural direction for future work.

Finally, these data cannot distinguish between two interpretations of the cross-field differences: a "lag" in which finance and macro are on the same trajectory as applied micro but further behind, and "different equilibria" in which each field converges to a different steady state determined by its research questions. The fact that applied micro's identification share has plateaued while finance's continues to rise is suggestive, but the data do not settle the question. Resolving this distinction—and understanding what drives it—would require a deeper analysis of the mechanisms behind methodological adoption.

References

Adão, Rodrigo, Michal Kolesár, and Eduardo Morales (2019). "Shift-share designs: Theory and inference". In: *The Quarterly Journal of Economics* 134.4, pp. 1949–2010.

Angrist, Joshua D and Alan B Krueger (1991). "Does compulsory school attendance affect schooling and earnings?" In: *The Quarterly Journal of Economics* 106.4, pp. 979–1014.

Angrist, Joshua D and Jörn-Steffen Pischke (2009). "Mostly Harmless Econometrics: An Empiricist's Companion". In.

— (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". In: *Journal of economic perspectives* 24.2, pp. 3–30.

Anthropic (2025). *Claude Language Models*. URL: https://www.anthropic.com.

Arkhangelsky, Dmitry et al. (2021). "Synthetic difference-in-differences". In: *American Economic Review* 111.12, pp. 4088–4118.

Autor, David H, David Dorn, and Gordon H Hanson (2013). "The China syndrome: Local labor market effects of import competition in the United States". In: *American economic review* 103.6, pp. 2121–2168.

Bartik, Timothy J (1991). "Who benefits from state and local economic development policies?" In.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022). "Quasi-experimental shift-share research designs". In: *The Review of Economic Studies* 89.1, pp. 181–213.

Boustan, Leah Platt (2010). "Was postwar suburbanization "white flight"? Evidence from the black migration". In: *The Quarterly Journal of Economics* 125.1, pp. 417–443.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020). "Methods matter: p-hacking and publication bias in causal analysis in economics". In: *American Economic Review* 110.11, pp. 3634–3660.

— (2024). "Mass reproducibility and replicability: A new hope". In: *American Economic Review* 114.11, pp. 3564–3610.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2024). *Difference-in-differences with a continuous treatment*. Tech. rep. National Bureau of Economic Research.

Chaisemartin, Clément de, Xavier d'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare (2022). "Difference-in-differences estimators for treatments continuously distributed at every period". In: *arXiv preprint arXiv:2201.06898*.

Currie, Janet, Henrik Kleven, and Esmée Zwiers (2020a). *Data and Code for Technology and Big Data Are Changing Economics: Mining Text to Track Methods*. Distributor: Inter-university Consortium for Political and Social Research, Ann Arbor, MI. Nashville, TN. DOI: 10.3886/E120827V1. URL: https://doi.org/10.3886/E120827V1.

— (2020b). "Technology and big data are changing economics: Mining text to track methods". In: *AEA Papers and Proceedings*. Vol. 110. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 42–48.

De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". In: *American economic review* 110.9, pp. 2964–2996.

Garg, Nikhil and Thiemo Fetzer (2025). "Tracking the Credibility Revolution Across Fields Using LLMs". In: *Working Paper*.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020). "Bartik instruments: What, when, why, and how". In: *American Economic Review* 110.8, pp. 2586–2624.

Imbens, Guido W and Thomas Lemieux (2008). "Regression discontinuity designs: A guide to practice". In: *Journal of Econometrics* 142.2, pp. 615–635.

Nakamura, Emi and Jón Steinsson (2018). "Identification in macroeconomics". In: *Journal of Economic Perspectives* 32.3, pp. 59–86.

Rambachan, Ashesh and Jonathan Roth (2023). "A more credible approach to parallel trends". In: *Review of Economic Studies* 90.5, pp. 2555–2591.

Roth, Jonathan (2022). "Pretest with caution: Event-study estimates after testing for parallel trends". In: *American Economic Review: Insights* 4.3, pp. 305–322.

Roth, Jonathan and Pedro HC Sant'Anna (2023). "When is parallel trends sensitive to functional form?" In: *Econometrica* 91.2, pp. 737–747.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2, pp. 175–199.

Table 4: Search Categories and Trigger Phrases. Unless noted otherwise, the outcome is the fraction of papers with at least one phrase match. "Figure" and "Table" categories use average word count per paper.

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
Administrative Data	'administrative data', 'admin data', 'administrative-data', 'admin-data', 'administrative record', 'admin record', administrative regist', 'admin regist', 'register data', 'registry data'	No	Yes	Yes
Big Data	'big data', 'big-data'	No	Yes	Yes
Binscatter	'binscatter', 'bin scatter', 'binned scatter'	No	Yes	No
Bunching	'bunching'	No	Yes	No
Clustering	'cluster'	No	Yes	Yes
Confidence Interval	'confidence interval'	No	Yes	Yes
Data	'data'	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'	Yes	Yes	No
Event Study	'event stud' ' event-stud'	No	Yes	No
External Validity	'external validity', 'external-validity', 'externally valid', 'externally-valid'	No	Yes	No
Figure	'graph', 'figure', 'plot', 'chart'	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'	Yes	No	Yes

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
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'	Yes	No	No
General Equilibrium	'general equilibr', 'general-equilibr'	No	Yes	No

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
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', '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'	No	Yes	No
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'	No	Yes	Yes

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
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'	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'	Yes	Yes	No
Machine Learning	'machine learning', 'lasso', 'random forest'	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'	No	Yes	Yes
Mechanisms	'mechanism'	No	Yes	No
Omitted Variables	'omitted variable'	No	Yes	Yes
Preanalysis Plan	'pre-analysis plan', 'pre analysis plan', 'preanalysis plan'	No	Yes	No
Precisely Estimated	'precisely estimated', 'precisely-estimated'	No	Yes	No

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
Precisely Estimated Zero	'precisely estimated zero', 'precisely-estimated zero'	No	Yes	No
Proprietary Data	'proprietary data', 'confidential data', 'nonpublic data', 'non-public data', 'proprietary-data', 'confidential-data', 'nonpublic-data', 'non-public-data'	No	Yes	Yes
Quasi- and Natural Experiments	'quasi experiment', 'quasi-experiment', 'quasiexperiment', 'natural experiment', 'natural-experiment'	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', 'randomized experiment', 'randomised experiment', 'randomized evaluation', 'randomised evaluation', 'randomized trial', 'randomised trial', 'randomized intervention', 'randomised intervention', 'randomized design', 'randomised design', 'field experiment'	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'	Yes	Yes	No
Reverse Causation	'reverse causa', 'reverse-causa'	No	Yes	Yes

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
Selection	'selection'	No	Yes	Yes
Simultaneity	'simultaneity'	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'	Yes	Yes	No
Survey Data	Sentence structure: we search for instances where the term 'survey' and 'data' are mentioned within two full stops.	No	Yes	Yes
Synthetic Control	'synthetic control'	No	Yes	Yes
Table	'table'	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'	No	Yes	No
Econometrics categories (Section 4)				
Asymptotic Theory	'asymptot', 'large sample', 'convergence rate', 'consistency', 'limiting distribut'	No	Yes	No
Bayesian	'Bayesian', 'posterior distribut', 'prior distribut', 'Markov chain Monte Carlo', 'MCMC'	Mixed	Yes	No
Bootstrap	'bootstrap', 'resampl'	No	Yes	No

Continued on next page

Table 4 – *Continued from previous page*

Category	**Trigger Phrases**	**Case Sens.**	**Wildcard**	**Cond. data**
Forecasting	'forecast accuracy', 'forecast error', 'forecast evaluation', 'forecast combination', 'out of sample forecast', 'predictive regression'	No	Yes	No
High-Dimensional	'high-dimensional', 'high dimensional', 'many regressors', 'many covariates', 'variable selection', 'sparsity', 'sparse model'	No	Yes	No
Kernel Methods	'kernel estimat', 'kernel density', 'kernel regression', 'bandwidth selection', 'kernel smooth'	No	Yes	No
Limited Dep. Var.	'Tobit', 'probit', 'logit', 'censored regression', 'truncated regression', 'sample selection model', 'Heckman', 'discrete choice'	Mixed	Yes	No
Nonparametric	'nonparametric', 'non-parametric', 'non parametric'	No	Yes	No
Panel Data	'panel data', 'longitudinal data', 'fixed effect', 'random effect', 'within estimat', 'between estimat'	No	Yes	No
Quantile Regression	'quantile regression', 'quantile treatment'	No	Yes	No
Regularization	'LASSO', 'ridge regression', 'elastic net', 'regulariz', 'penalized regression', 'shrinkage estimat'	Mixed	Yes	No
Robust Inference	'heteroskedasticity robust', 'HAC', 'Newey-West', 'robust standard error', 'cluster robust', 'wild bootstrap'	Mixed	Yes	No
Semiparametric	'semiparametric', 'semi-parametric', 'semi parametric'	No	Yes	No
Simulation/Monte Carlo	'Monte Carlo', 'MCMC', 'Markov chain Monte', 'Gibbs sampl'	Mixed	Yes	No
Time Series (VAR/GARCH)	'VAR', 'vector autoregress', 'ARMA', 'ARIMA', 'unit root', 'cointegrat', 'GARCH', 'ARCH', 'stationarity', 'impulse response'	Mixed	Yes	No
Treatment Effects	'treatment effect', 'average treatment', 'causal effect'	No	Yes	No
Weak Identification	'weak instrument', 'weak identification', 'Anderson-Rubin', 'Stock-Yogo'	Mixed	Yes	No

A LLM Validation of Keyword Matching

The analysis throughout this paper relies on keyword matching—a simple, scalable approach with obvious limitations. Keywords may produce false positives (papers that mention a method without actually using it) or false negatives (papers that use a method but describe it in non-standard language). Recent work by Garg and Fetzer (2025) uses LLMs for paper classification, while Brodeur, Cook, and Heyes (2024) provide hand-coded benchmarks, raising the question of whether more sophisticated approaches would yield different results.

To assess reliability, I classify a stratified sample of approximately 750 papers using Claude Haiku 4.5 (Anthropic 2025). For each paper, I provide the first 1,500 words and ask the model to identify which methods are actually *used*, as opposed to merely mentioned. I then compare the LLM classification with the keyword flags.

The sample is stratified to ensure adequate representation of each method. I select at least 50 papers flagged by keywords for each of the nine method categories, at least 100 papers with no method flags (to detect false negatives), and fill the remainder with a stratified random sample split equally between NBER and journal papers. Because the sample overrepresents rare methods relative to the population, the agreement rates reported here should be interpreted as method-specific accuracy rather than overall population accuracy. If 80 percent of papers do not mention any method, a high false-negative rate on a rare method could affect population-level trend estimates while barely showing up in these agreement rates. The main text addresses this concern in part by showing that NBER and journal trends are consistent, and by decomposing composite measures (e.g., DiD vs. event studies) to isolate potential sources of terminological conflation.

Table 5 presents the results. For most categories, keyword matching achieves 80–92 percent agreement with LLM classification. Agreement is highest for regression discontinuity and lab experiments—where the terminology is distinctive and unlikely to appear outside its intended context. Agreement is lower for identification (where keyword patterns may flag theoretical discussions) and structural models (where terms like GMM and MLE appear in many contexts beyond structural estimation).

Agreement rates also vary by field (Table 6). Agreement tends to be higher in finance and macro, where quasi-experimental methods are rarer and keyword mentions more likely to reflect actual use. Applied micro shows slightly lower agreement, particularly for identification strategy, reflecting the higher base rate and more varied terminology in that field.

Compared to the GPT-4o-mini benchmarks reported in Garg and Fetzer (2025), the keyword approach achieves comparable accuracy at near-zero computational cost (Figure 14). For tracking broad trends over time—the objective of this paper—simple keyword matching is a reliable tool.

Benchmark against hand-coded classifications

As a complementary validation exercise, I benchmark the keyword approach against the hand-coded classifications in Brodeur, Cook, and Heyes (2024), who had research assistants read and classify 1,108 published articles into four categories: DID, IV, RCT, and RDD. Matching papers by title

Table 5: Keyword vs. LLM Classification Agreement

Method	N	Accuracy	Precision	Recall	F1	κ
DiD	750	86.7%	47.1%	80.2%	59.3%	0.520
Event Study	750	86.5%	43.2%	64.3%	51.7%	0.442
IV	750	80.4%	37.6%	93.4%	53.6%	0.439
RD	750	91.6%	43.5%	95.9%	59.9%	0.559
RCT	750	86.8%	40.3%	94.1%	56.4%	0.500
Lab Experiment	750	92.7%	37.5%	85.7%	52.2%	0.489
Identification Strategy	750	65.6%	39.4%	75.0%	51.7%	0.288
Structural Model	750	69.1%	59.1%	36.0%	44.8%	0.250
Administrative Data	750	65.9%	79.4%	39.5%	52.8%	0.305

Notes: Keyword-based classification treated as positive when any pattern matches. LLM classification uses Qwen3.5-122B-A10B-FP8 with temperature 0. Precision and recall measured with keyword as the classifier and LLM as ground truth. Sample of 750 papers stratified by source, field, and method.

Table 6: Keyword vs. LLM Classification Agreement by Field

Method	Overall	Applied Micro	Finance	Macro/Others
DiD	85.7%	83.2%	90.1%	93.6%
Event Study	84.7%	83.2%	79.1%	90.4%
IV	80.8%	80.2%	84.6%	89.4%
RD	91.5%	88.6%	97.8%	96.8%
RCT	90.7%	88.2%	96.7%	100.0%
Lab Experiment	91.3%	88.6%	96.7%	98.9%
Identification Strategy	66.1%	61.7%	69.2%	81.9%
Structural Model	65.7%	69.7%	65.9%	57.4%
Administrative Data	67.7%	64.7%	68.1%	71.3%

Notes: Agreement rates between keyword-based and LLM-based classification, stratified by field. Field classification uses NBER program affiliations for working papers and journal identity for published articles.

yields 501 overlapping articles.

Table 7 reports the results. Keyword matching achieves high recall for IV (95%) and RDD (94%), meaning it detects nearly all papers that trained coders identify as using these methods. Recall for DID is also strong at 84%. RCT recall is also strong at 83%, though precision is lower (67%), reflecting that keyword patterns for experiments cast a wide net. Precision is highest for DID (69%) and RDD (65%), and lower for IV (58%) and RCT (67%).

An important caveat when interpreting these precision numbers: the Brodeur et al. benchmark assigns each paper a single *primary* method, while keyword matching flags *any* method discussed in the paper. These are fundamentally different classification tasks. Among the 114 papers that keywords flag for IV but Brodeur et al. classify under a different primary method, 54 are coded as DID, 39 as RCT, and 21 as RDD. Many of these papers likely do discuss or employ IV—as a robustness check, a complementary strategy for handling noncompliance in an RCT, or a secondary specification—even though it is not the primary research design. Similarly, of the 61 DID "false

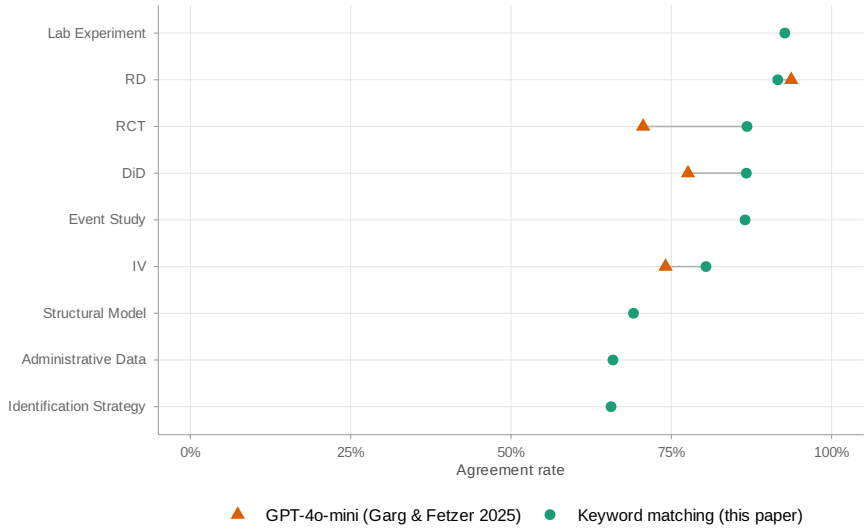


Figure 14: Agreement rates between keyword matching and LLM classification, by method. Where available, GPT-4o-mini accuracy from Garg and Fetzer (2025) is shown for comparison.

positives," 29 are papers whose primary method is IV but that also employ difference-in-differences in some capacity. What looks like low precision against a single-label benchmark may partly reflect the multi-method reality of modern empirical economics.

The pattern of errors differs across methods in a predictable way. For IV, keywords achieve high recall (95%) but moderate precision (58%): nearly every IV paper is detected, but many papers that use IV as a secondary strategy are also flagged. This will overstate the *level* of IV adoption. For RCT, the pattern is somewhat different—lower precision (67%) and high recall (83%)—because the broad keyword patterns for experiments capture many papers that discuss experimental methods without using them as a primary design. Notably, of the 59 RCTs that keywords miss, 24 are flagged for IV (likely intent-to-treat designs with IV for compliance) and 27 for identification strategy—the papers are detected, but under a different methodological label.

These biases in levels need not bias the *time trends* that are the focus of this paper, provided that the error rates are approximately stable over time. A constant false-positive rate for IV would shift the trend line up without changing its slope. The main risk is that terminology conventions themselves evolve—for instance, if "event study" increasingly appears in DiD papers over time, the false-positive rate for event studies would rise, potentially conflating a terminological shift with a methodological one. The DiD decomposition analysis in Appendix C addresses this concern directly by separating strict DiD from event-study language.

Table 8 compares keyword and LLM classification against the same Brodeur et al. benchmark, using a subset of 357 papers where LLM classifications (Qwen 3.5-122B) are also available. Note that Table 7 uses the Brodeur, Cook, and Heyes (2024) strict DiD definition, while Table 8 uses the composite DiD measure (including event studies) to match the main text's classification.

The results reveal a clean trade-off between the two approaches. Keywords achieve near-perfect recall for DiD (99%) and IV (100%)—they almost never miss a paper—but lower precision (68–

Table 7: Keyword Classification vs. Brodeur, Cook, and Heyes (2024) Hand-Coded Benchmark

Method	N_{hand}	N_{kw}	Precision	Recall	F1	Accuracy	κ
DID	164	198	69.2%	83.5%	75.7%	82.4%	0.621
IV	164	270	57.8%	95.1%	71.9%	75.6%	0.526
RCT	108	134	67.2%	83.3%	74.4%	87.6%	0.663
RDD	65	94	64.9%	93.8%	76.7%	92.6%	0.725
Overall	501	696	63.8%	88.6%	74.2%	—	—

Notes: Benchmark against hand-coded classifications from Brodeur, Cook, and Heyes (2024). N_{hand} = papers classified as using this method by Brodeur et al.; N_{kw} = papers flagged by keyword matching. Precision = share of keyword-flagged papers confirmed by hand coding; Recall = share of hand-coded papers detected by keywords. Sample: 501 papers matched by title across both datasets.

69%), reflecting that they also flag papers that mention a method in passing. The LLM achieves higher precision (84–91%) but lower recall for DiD (79%) and especially IV (58%): the LLM is more conservative, flagging only papers where the method is central. For RCT, the LLM dominates on both dimensions (92% precision, 94% recall vs. 81% and 84% for keywords), reflecting that LLMs can recognize the diverse language economists use for experiments. For RDD, both approaches perform well, with the LLM achieving slightly higher F1 (88% vs. 83%).

For the purpose of tracking *diffusion* of methods across fields over time, the mention-detection approach has a natural appeal. A paper that employs DID as its primary identification strategy and IV as a robustness check reflects adoption of *both* methods. The keyword approach captures this; a single-label classifier does not. At the same time, keyword matching cannot distinguish between a paper that *uses* a method and one that merely *cites* it in a literature review, which is the main source of false positives. The time-series implications depend on whether the ratio of use-mentions to cite-mentions is stable over time—a plausible assumption for established methods, though less certain for methods undergoing rapid diffusion.

Table 8: Keyword vs. LLM Classification: Brodeur, Cook, and Heyes (2020) Hand-Coded Benchmark

	Keywords			Qwen 3.5-122B		
Method	Precision	Recall	F1	Precision	Recall	F1
DID	68.0%	99.2%	80.7%	84.4%	78.6%	81.4%
IV	69.1%	100.0%	81.7%	90.3%	57.9%	70.6%
RDD	74.3%	94.5%	83.2%	88.9%	87.3%	88.1%
RCT	81.2%	83.9%	82.5%	91.6%	93.5%	92.6%

Notes: Both approaches benchmarked against hand-coded method labels from Brodeur, Cook, and Heyes (2020). Sample of 357 papers matched by journal, year, and title across nine journals (2011–2020). Brodeur et al. code methods at the test-statistic level; a paper is labeled as using a method if any of its reported estimates uses that method. Keywords detect any mention of a method; the LLM (Qwen 3.5-122B) classifies whether a method is actually *used* as part of the research design. Precision and recall treat the hand-coded labels as ground truth.

B Graphical Revolution

Figure 15 plots the graphical revolution—the ratio of figure mentions to table mentions—overall and by field. The trend continues upward across all fields, consistent with Currie, Kleven, and Zwiers (2020b). Macro/other generally has the highest ratio, though applied micro has tracked above finance in recent years and the ordering is not stable across the full sample period.

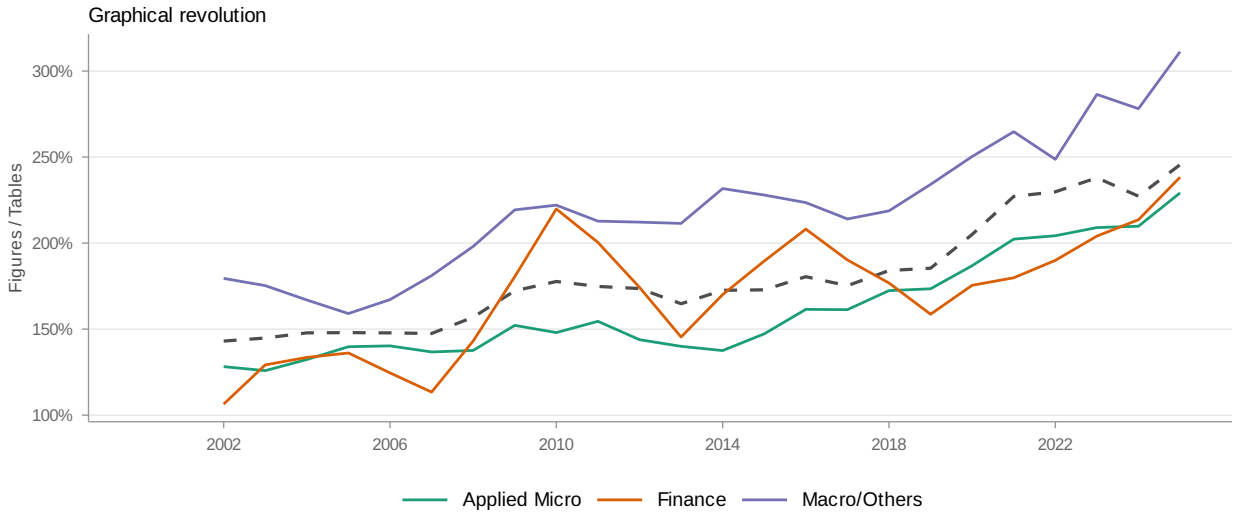


Figure 15: Graphical revolution trends in NBER working papers (two-year moving averages). Colored lines show field-specific trends; dashed black line shows the overall aggregate.

C DiD Decomposition: Strict DiD vs. Event Studies

The main text uses a composite "DiD" measure that flags papers mentioning either "difference-in-differences" (strict DiD language) or "event study." This appendix decomposes the two components.

Figure 16 shows the decomposition by field. Panel (a) plots the strict DiD measure (excluding event studies), while panel (b) plots any event study mention. Panels (c) and (d) focus on two subsets: papers mentioning event studies *without* strict DiD language (panel c), and—within that subset—papers that also mention "abnormal return," a strong indicator of a financial event study rather than a DiD-style event study (panel d). In finance, a substantial share of the "event study only" papers mention abnormal returns, consistent with the concern that the composite DiD measure captures some financial event studies that are methodologically distinct from difference-in-differences designs.

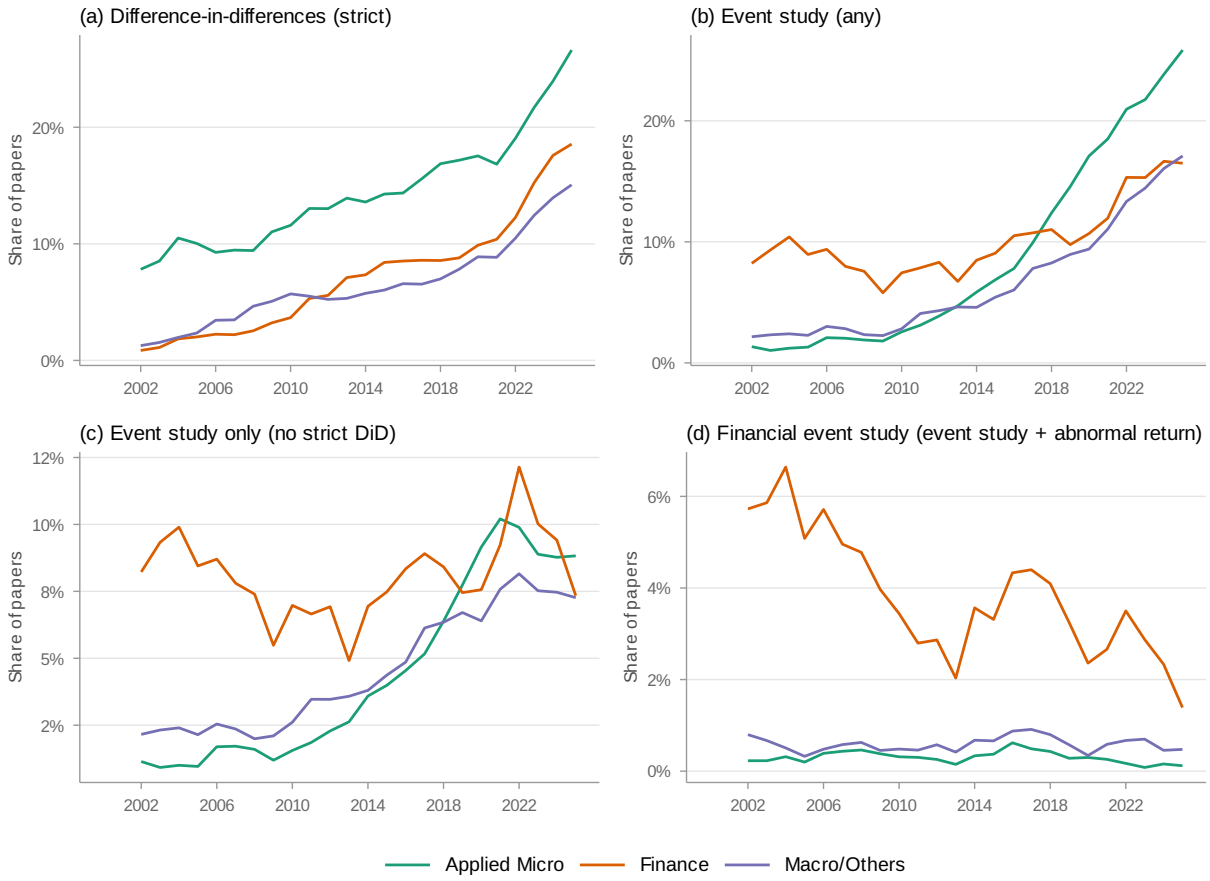


Figure 16: Decomposition of the DiD measure by field. Two-year moving averages. Panel (a): strict DiD language only. Panel (b): any event study mention. Panel (c): event study mentions without strict DiD language. Panel (d): event study mentions with "abnormal return" (financial event study proxy).

D Exclusive Field Classification

The main analysis uses non-exclusive field labels: a paper that lists both applied micro and finance programs counts in both categories. This appendix re-runs the main results using an exclusive classification where each paper is assigned to a single field based on its program affiliations. Papers that list programs from multiple fields are excluded.

Under exclusive classification, the sample comprises 18,697 papers (11,828 Applied Micro, 1,758 Finance, 5,111 Macro/Others). Figure 17 replicates the main time-series analysis. The cross-field gaps are qualitatively similar and, if anything, slightly wider under exclusive classification—consistent with cross-listed papers pulling the non-applied-micro averages toward applied micro.

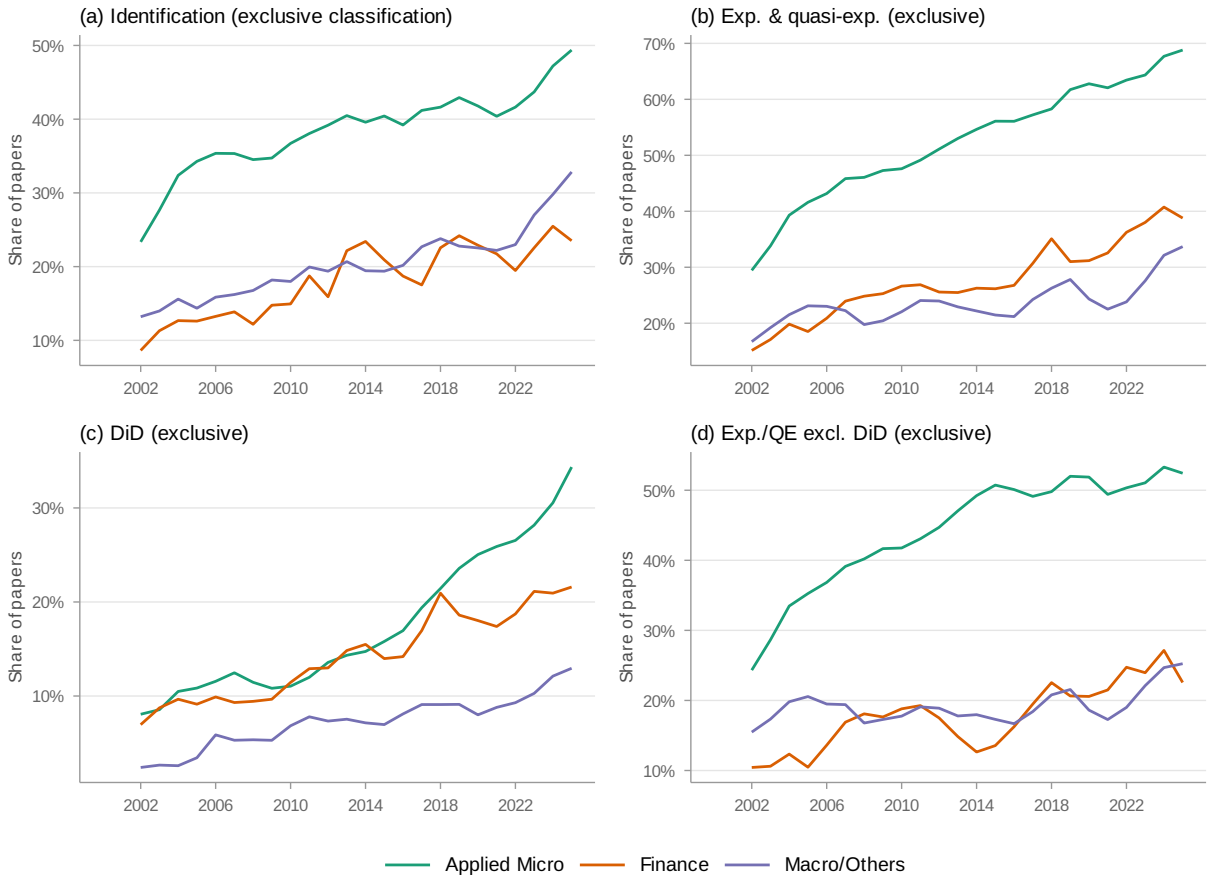


Figure 17: Credibility revolution trends under exclusive field classification. Only papers whose programs all map to a single field are included. Two-year moving averages.

E Rate of Change across Fields

The main text compares *levels* of credibility revolution method mentions across fields. Figure 18 plots the *rate of change*—the first difference of each field's share, smoothed with a three-year moving average. If all fields are converging to the same steady state, the fields with lower levels should show higher growth rates. If fields have different steady states, growth rates should be similar.

The data are more consistent with different long-run equilibria than with simple convergence. Applied micro's growth rate in identification (panel a) was near zero from 2016 through the early 2020s, while finance and macro showed positive growth over the same period—though the most recent years show some convergence in growth rates. For experimental/quasi-experimental methods (panel b), all three fields show positive growth, with finance slightly above macro.

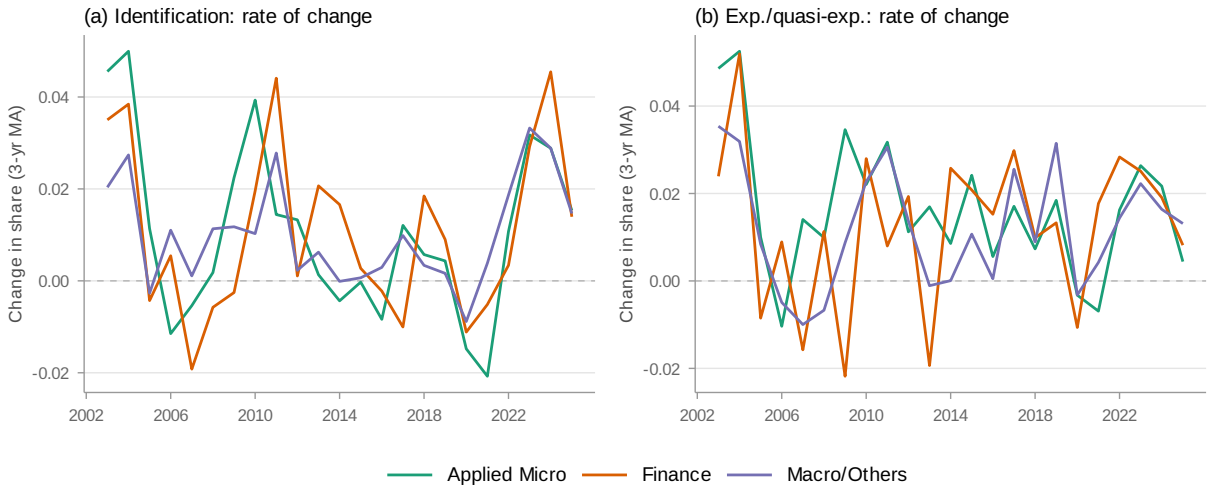


Figure 18: Rate of change in credibility revolution measures by field (three-year moving average of first differences).

F Structural Model Measure: Broad vs. Narrow

The main text uses a broad structural model measure that includes GMM and MLE keywords. Because GMM and MLE are used in many non-structural contexts (e.g., MLE for logit/probit, GMM for moment conditions in reduced-form panel models), this may overstate the share of truly structural papers. Figure 19 compares the broad measure (panels a, c) with a narrow measure that excludes GMM and MLE (panels b, d).

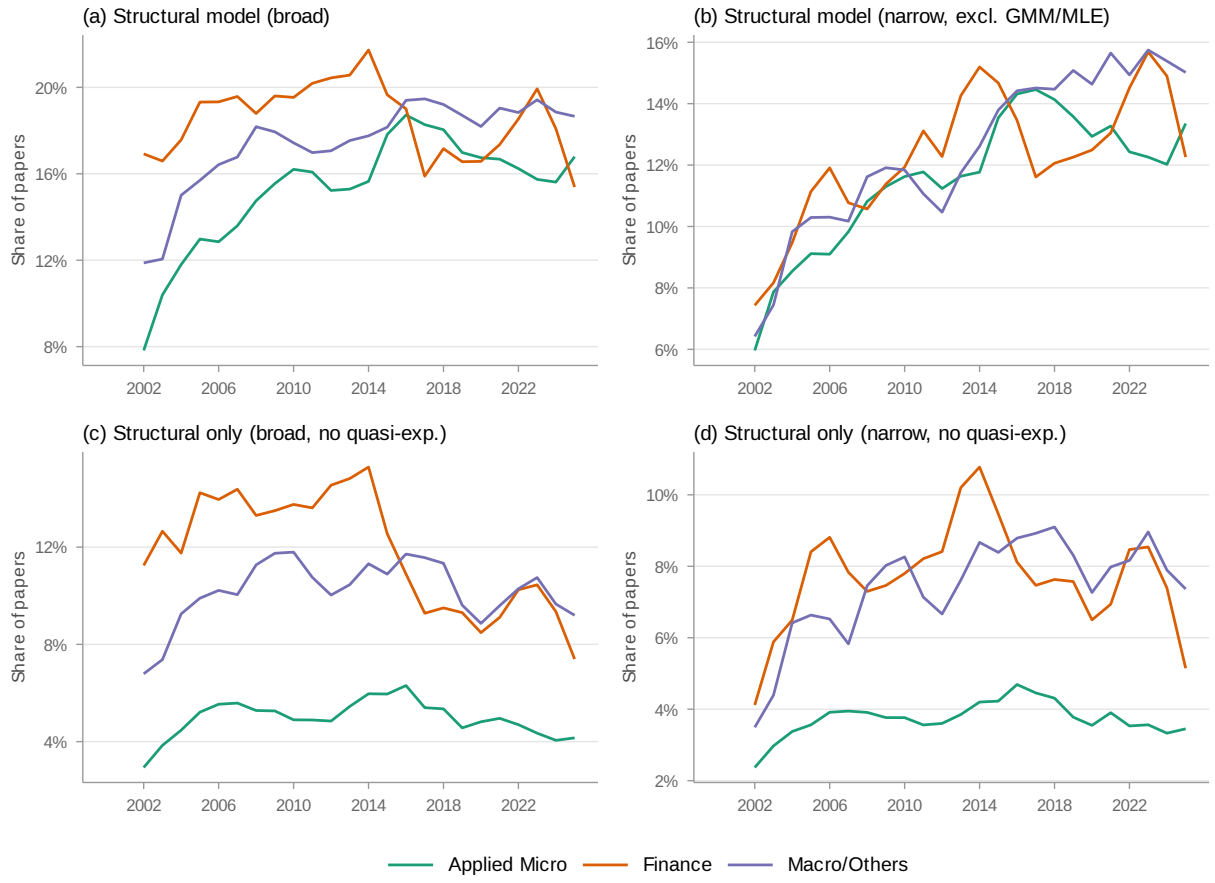


Figure 19: Structural model measures: broad (including GMM/MLE) vs. narrow (excluding GMM/MLE). Panels (a) and (b): all papers. Panels (c) and (d): papers that do not also mention experimental/quasi-experimental methods.

G Denominator Composition

The main analysis computes method shares as a fraction of *all* papers in each field. However, finance and macro/other have a larger share of purely theoretical papers that would never use quasi-experimental methods, potentially inflating the measured cross-field gap. To assess whether denominator composition drives the results, I restrict the sample to "empirical" papers—defined as those that mention at least one of: identification, any experimental or quasi-experimental method, administrative data, survey data, or structural estimation. This is a minimal bar: a paper that discusses none of these is likely theoretical.

In 2024, 82 percent of applied micro papers meet this threshold, compared to 73 percent in finance and 64 percent in macro/other. Figure 20 replicates the method-specific time-series analysis (Figure 4) for both all papers (left column) and the empirical subsample (right column). The cross-field gap narrows—from 63/47/39 percent (unconditional) to 76/65/61 percent (conditional) for experimental and quasi-experimental methods in 2024—but clearly persists across all method

categories. Denominator composition accounts for roughly one-third of the cross-field gap; the remainder reflects differential adoption of credibility revolution methods within empirical work.

Figure 20: Method-specific trends by field: all papers (left) vs. empirical papers only (right). Empirical papers are defined as those mentioning at least one of: identification, experimental/quasi-experimental methods, administrative data, survey data, or structural estimation. Two-year moving averages.

H Synthetic Control and Synthetic DiD

Figure 21 plots synthetic control and synthetic DiD mentions by field in separate facets. The left facet confirms the decline in synthetic control mentions after 2020, which is evident in applied micro and macro/other, though finance shows a slight increase from a near-zero base. The right facet shows that "synthetic DiD" (including "SDID" and "synthetic difference-in-differences") has emerged but remains rare, suggesting that the decline in synthetic control is not fully explained by substitution toward synthetic DiD methods. Note the different y-axis scales.

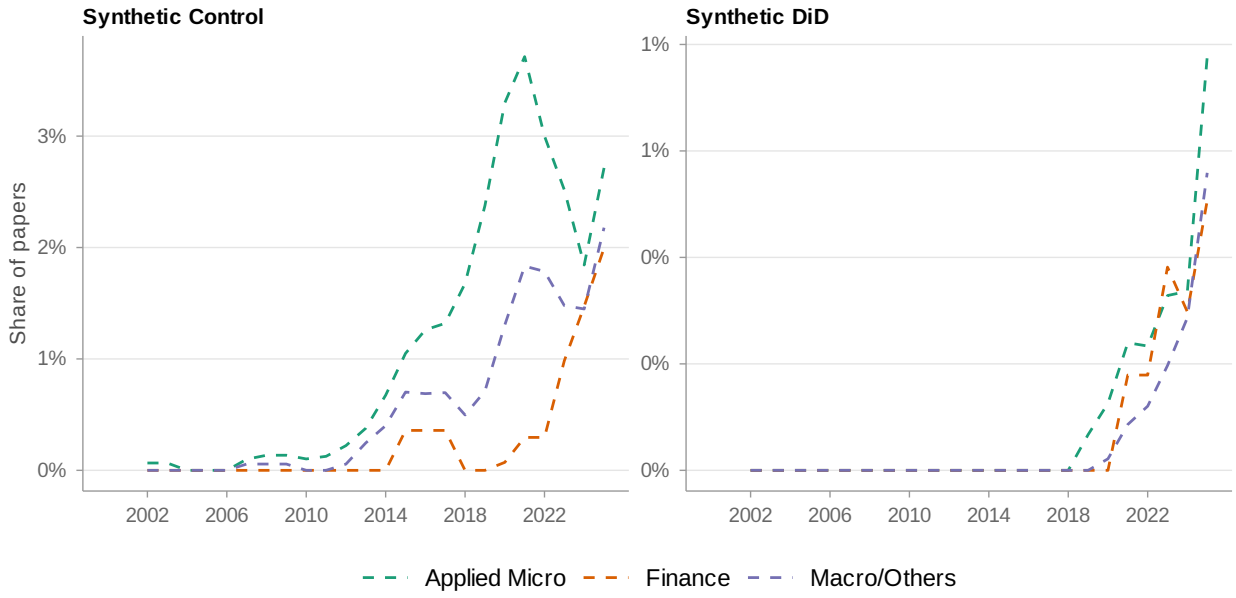


Figure 21: Synthetic control and synthetic DiD mentions by field (two-year moving averages). Note: y-axis scales differ across facets to accommodate different prevalence levels.

I Top Journals: IV and Structural Model Trends

Figure 22 presents the full time series of method mentions across fields in the journal sample (three-year moving averages). The main text shows slope charts for identification and DiD; this figure adds instrumental variables and structural models and shows the annual dynamics. Structural model mentions are markedly higher in finance and macro journals, reinforcing the picture from the working papers that these fields maintain a larger structural modeling tradition alongside the adoption of quasi-experimental methods. IV trends are broadly flat across fields.

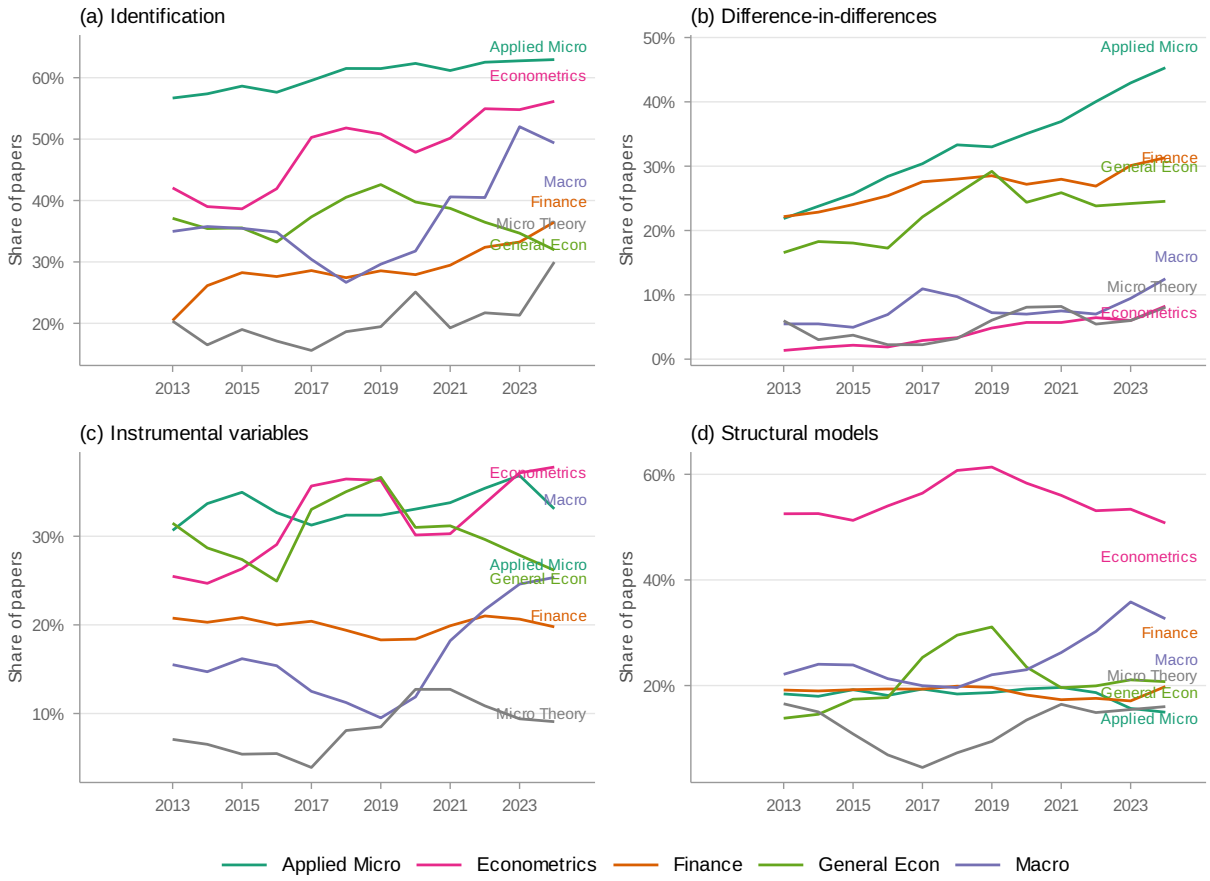


Figure 22: Full time series of method mentions across fields in top journals (2011–2024). Three-year moving averages. Panels (a)–(d): identification, DiD, instrumental variables, structural models. See text for journal-to-field mapping.

J Journal Analysis: Excluding General Economics Journals

The main journal analysis includes general-interest journals (AER, QJE, JPE) classified by JEL codes. Because JEL classification is noisy, this appendix re-runs the journal analysis using only field-specific journals: AEJ Applied and AEJ Policy for applied micro, JF/JFE/RFS for finance, and AEJ Macro for macro. Figure 23 shows that the cross-field patterns hold using only field-specific journals.

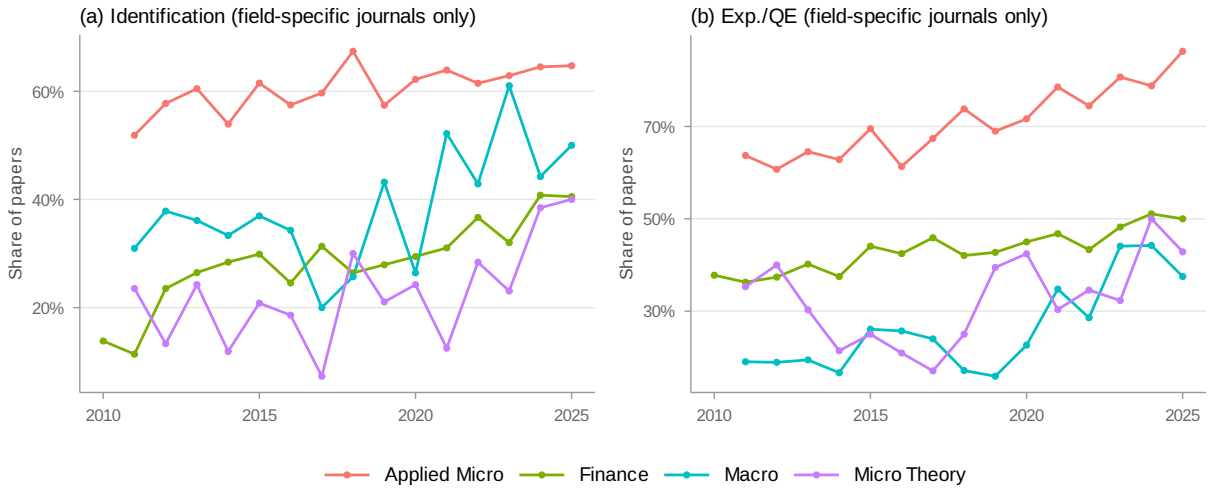


Figure 23: Journal trends excluding general economics journals (field-specific journals only).

K Journal Text Extraction Coverage

Table 9 documents text extraction rates by journal and year. Extracting full text from published journal PDFs is challenging: publisher access restrictions, varied PDF formats, and multi-column layouts all affect extraction quality. Coverage is generally high for journals where PDFs were readily accessible (AEJ journals, AER, finance journals) but incomplete for others. Two journals are excluded from the main analysis due to insufficient coverage: the *Review of Economic Studies* (zero text extraction across all years) and *Econometrica* (near-zero coverage for 2011–2014, partial thereafter). Results including Econometrica appear as a robustness check. Gaps in coverage could bias keyword rates if the missing papers differ systematically from the extracted ones; however, since extraction failures are driven by PDF format rather than paper content, this bias is likely small.

Table 9: Text extraction coverage by journal and year. Each cell shows the number of papers with extracted text out of total papers in the database. Shaded cells indicate coverage below 80%.

Journal	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	
AEJ Applied	30/37	40/40	47/47	30/30	45/45	37/37	31/31	46/46	44/45	50/50	41/41	63/63	55/55	64/64	28/28	651
AEJ Macro	31/32	30/31	36/36	24/24	43/43	30/30	24/24	32/32	40/40	48/48	46/46	56/56	52/52	50/50	30/30	572/
AEJ Micro	38/38	31/31	34/34	44/44	49/49	43/43	41/41	40/40	38/38	34/34	56/56	82/82	65/65	52/52	35/35	682/
AEJ Policy	29/30	37/37	44/44	45/45	46/46	39/39	51/52	48/48	54/54	51/51	57/57	63/64	64/64	64/64	48/48	740/
AER	260/274	260/264	257/259	255/257	261/262	273/274	244/264	112/113	134/134	120/120	115/115	114/115	95/96	111/111	87/87	2698
J. Econometrics	150/150	136/136	135/135	150/150	195/195	129/129	118/118	159/159	165/165	262/262	164/164	143/143	192/192	175/175	161/161	2434/
J. Finance	61/63	93/94	102/105	78/82	100/103	85/130	71/84	87/104	85/103	88/98	81/90	77/86	94/100	95/98	74/89	1271
JFE	139/140	123/125	135/136	114/116	86/87	134/137	112/114	162/164	136/142	141/143	269/273	88/89	80/82	114/120	122/147	1955
JPE	24/36	33/43	31/44	29/46	41/55	33/55	72/97	89/121	66/89	75/97	69/93	86/115	80/115	79/108	73/93	880/
QJE	46/47	56/56	36/45	38/46	39/47	45/54	51/52	32/32	39/40	50/51	48/48	42/42	54/54	38/45	38/48	652
RFS	0/114	96/101	95/100	97/98	109/121	97/104	143/147	132/138	143/145	144/147	135/137	94/94	93/95	85/89	86/111	1549
Excluded from main analysis (robustness only):																
Econometrica	1/70	0/99	0/89	0/83	42/86	54/77	61/82	61/84	58/83	87/111	91/113	91/115	75/100	64/89	58/79	743/
R. Econ. Stud.	0/84	0/47	0/51	0/44	0/53	0/65	0/72	0/75	0/63	0/94	0/98	0/77	0/115	0/113	—	0/

L J. Econometrics: Time Trends in Credibility Revolution Methods

The main text compares term prevalence between the *Journal of Econometrics* and applied journals in cross-section. Figure 24 shows time trends within the *Journal of Econometrics*, separately for theory/estimation terms (panel a) and credibility revolution terms (panel b). While theory terms (nonparametric, bootstrap, Bayesian, machine learning, time series, simulation) remain dominant, credibility revolution methods—particularly DiD, treatment effects, and heterogeneous treatment effects—have grown substantially since 2011, suggesting the gap may be narrowing.

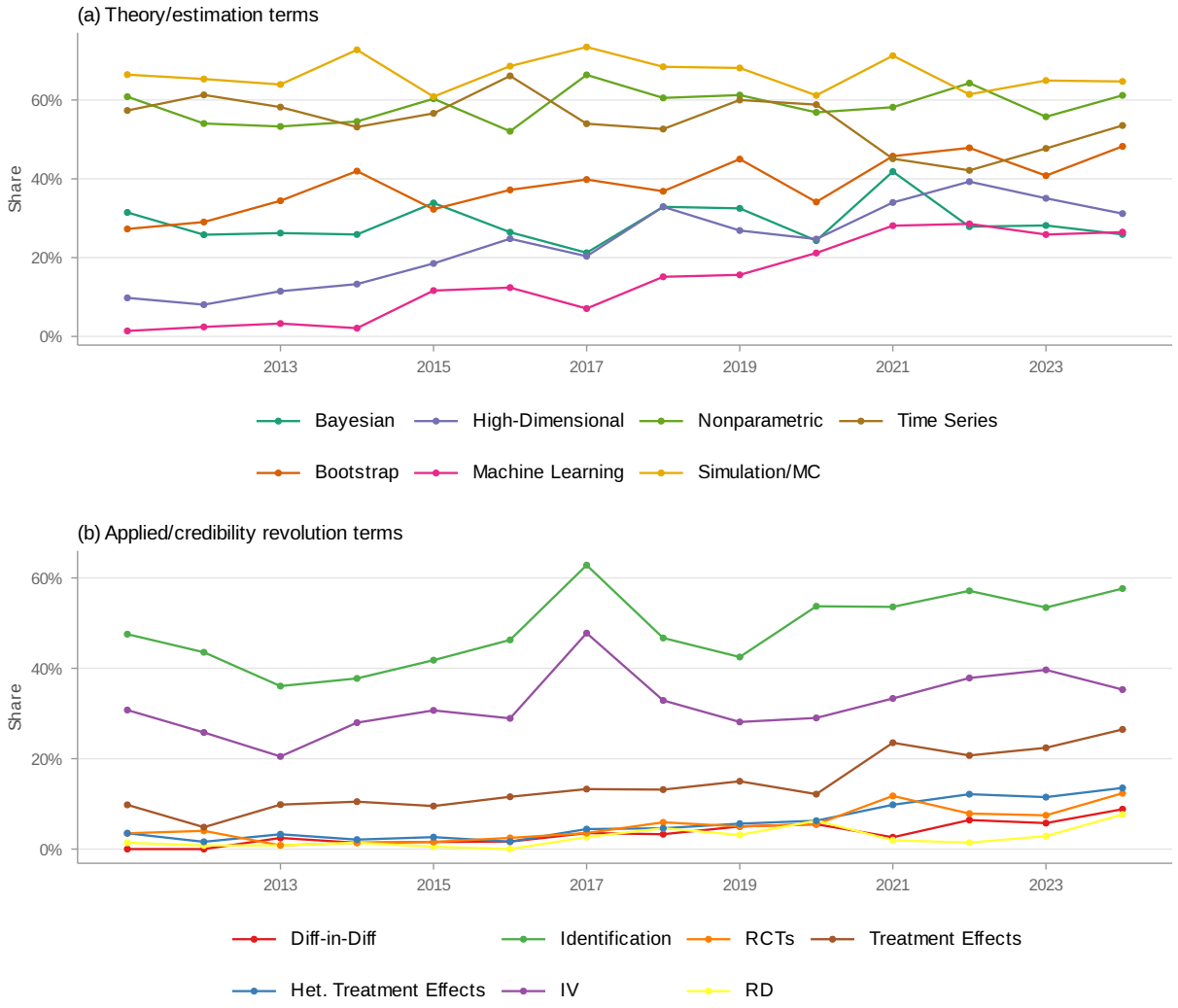


Figure 24: Time trends in the *Journal of Econometrics*, 2011–2024. Panel (a): theory/estimation terms. Panel (b): credibility revolution terms.