

Instruments of Change

The Rise of Identification Designs in Economics

Aleks Shneider

University of Minnesota

October 2026

Why measure the credibility revolution?

- Angrist and Pischke (2010) called it a revolution. Empirical economists started saying exactly where their identification comes from: a policy change, a cutoff, a lottery.
- Nobody doubts it happened. Fewer people have measured how far it reached, and the recent attempts count method words in papers (Currie, Kleven and Zwiers 2020; Goldsmith-Pinkham 2024).
- I go further in three ways. I correct the counts for the classifier's mistakes, using papers I read and labelled myself. I ask whether the designs moved from the top journals down, and which ones did. And I link papers to their authors to see whether designs travel through people.
- The hierarchy is where it gets interesting. Anyone with a dataset can run a DiD, an RDD or an IV regression. An experiment needs funding and a field partner, and a structural model needs years of training. If resources matter, the two kinds should spread differently.

How did the designs of the credibility revolution spread through economics journals, and did they move from the top journals down?

- ① **They spread.** Corrected for classifier error, the share of papers using at least one of seven designs went from 17.5 % in the 1990s to 42.6 % in the 2020s.
- ② **The data-only designs spread down the hierarchy.** In a test I fixed in advance, DiD, RDD and IV converge from the top journals to the rest about 36.5 % per decade faster than RCTs and structural estimation.
- ③ **Designs run in networks, but I cannot show they pass between people.**[†] Authors take up the designs their co-authors and colleagues use. Once every other design serves as a placebo, only RCTs through co-authors stand out: 1 channel of 14, about what chance would give.

- ➊ **Data and measurement:** the corpus, a phrase classifier for seven designs, and hand labels that correct it.
- ➋ **How far did the designs spread?** Prevalence by decade, corrected for classifier error.
- ➌ **Did they spread down the hierarchy?** The comparison of data-only and resource-intensive designs, fixed in advance.
- ➍ **Does it hold up?** Dropping designs, moving thresholds, holding topics fixed, and checking whether the classifier is better at top journals.
- ➎ **Do designs travel with people?** Co-authors, colleagues and advisors, against placebo designs.
- ➏ Limitations and summary. Tables and further checks are in the appendix.

What was fixed in advance, and what is exploratory

Fixed before any label was seen or the run was made

- The two design groups and the test comparing them: A = DiD, RDD, IV; B = RCTs, structural estimation.
- The top-five cut as a robustness check.
- The three conditions for reading the threshold results.
- The rule for reading the topic split.
- The rule for judging each people channel against the placebo designs, written before the checks ran.

Exploratory, marked † on the slides

- Dropping one design at a time; moving event studies into group A.
- The topic split itself.
- Everything about spread through people, including advisors.
- The main weighting of the thinly labelled 1990s, chosen after all four candidates were seen.

- I start from every article OpenAlex lists for these journals, 1990s to 2020s, minus duplicates, editorials, book reviews and other front matter.
- I have the full text of 38.1 % of them, from publishers, repositories and working-paper series. When all I have is the working paper, I adjust it toward its published version, using papers I have in both forms. The texts are then reweighted to look like the whole corpus.
- Dropping methods primers and survey journals leaves 111,105 texts.
- The top tier is the top five plus the journals whose 1990s papers were cited most (38 journals); the other 79 are the rest. Ranking on 1990s citations keeps the tiers from being shaped by the spread itself.
- Finance and accounting (15,661 texts in 24 journals) get their own results and stay out of the tier tests. That leaves 95,444 economics texts.

How a paper gets flagged

- Each design has a list of weighted phrases. A paper is flagged when its weighted count reaches the design's threshold; a phrase counts only a few times, and the reference list is ignored.
- Seven designs: difference-in-differences, regression discontinuity, instrumental variables, randomized trials, event studies, synthetic control and structural estimation. "Any design" means at least one of them.
- Event studies count only in economics journals. In finance and accounting the same words describe the abnormal-return design.
- I chose each threshold to maximize F1 against the hand labels:

	DiD	RDD	IV	RCT	Event study	Synth. control	Structural
threshold	4.0	5.0	4.0	5.0	3.0	4.0	2.5

Difference-in-differences

Strong:
“difference-in-differences”,
“parallel trends”

Supporting: “two-way fixed effects”, “staggered adoption”, “Callaway and Sant’Anna”, “Goodman-Bacon”, “treated and control states”

Regression discontinuity

Strong: “regression discontinuity”, “discontinuity design”, “regression kink”

Supporting: “running variable”, “forcing variable”, “bandwidth”, “rdrobust”, “McCrary”

Instrumental variables

Strong: “instrumental variables”, “two-stage least squares”, “2SLS”

Supporting: “exclusion restriction”, “first stage”, “shift-share”, “weak instruments”

A strong phrase used two or three times flags a paper; supporting phrases add up only when several appear together. Because the reference list is dropped, citing a DiD paper does not count as using DiD.

How the classifier does against my labels

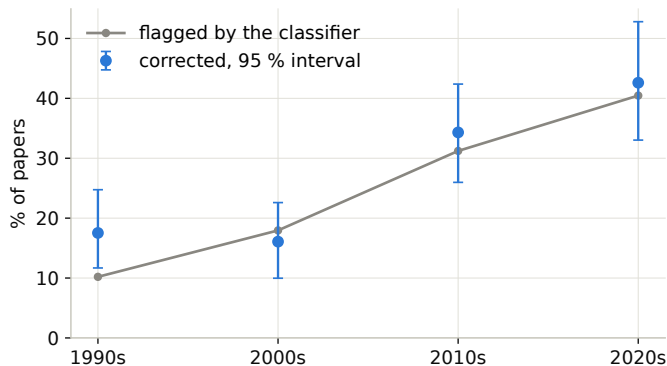
design	precision [interval]	recall [interval]
DiD	0.841 [0.710, 0.945]	0.829 [0.729, 0.932]
RDD	0.805 [0.645, 0.930]	0.809 [0.560, 1.000]
IV (first labelling round)	0.853 (35 positives)	0.829
Event study	0.744 [0.597, 0.889]	0.833 [0.667, 1.000]
RCT	0.743 [0.561, 0.883]	0.538 [0.408, 0.700]
Structural	0.538 [0.366, 0.698]	0.783 [0.584, 1.000]

DiD, RDD, IV and event studies are measured well. The weak spots are RCTs and structural estimation, and those are exactly the comparison group in the main test, so keep this table in mind later. Estimates are weighted to the labelling sample's design, with 95 % bootstrap intervals.

I labelled every paper; an LLM second coder flagged disagreements

- **First round:** 250 papers.
- **Second round:** 339 papers read in full, drawn where the classifier is least certain. They include 14 first-round papers, relabelled to separate the two kinds of event study.
- **Total:** 589 hand labels.
- **Second coder.** An LLM coded the same papers independently. Its agreement with me in the second round: kappa 0.665 to 0.912 across the seven designs, 0.746 for any design (first round: 0.781).
- **Adjudication.** I re-read all 147 disagreements: 109 went my way, 35 the LLM's, and 3 were settled by a written rule. The adjudicated labels are used throughout; the LLM never sets a label by itself.

Papers using an identification design: 17.5 % in the 1990s, 42.6 % in the 2020s



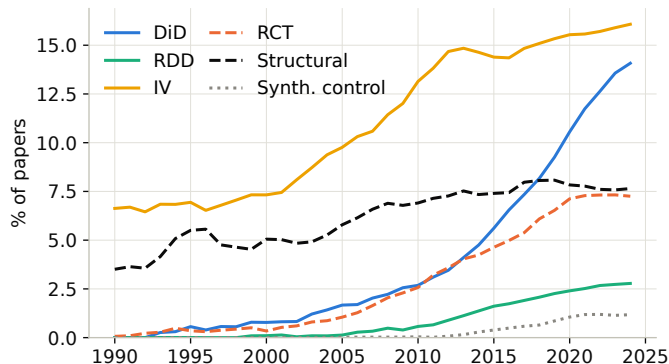
As flagged (analysis sample): 11.0 % $\rightarrow$ 39.0 %, growth $3.54\times$.

Corrected: 17.5 % $\rightarrow$ 42.6 %, growth $2.39\times$ [1.59, 3.78].

The 1990s have few labels, so the correction there depends on how they are weighted: across four weightings, growth runs from $2.25\times$ to $2.44\times$.[†]

The classifier misses more 1990s papers, so the correction mostly raises the starting point.

Every design grew, DiD fastest



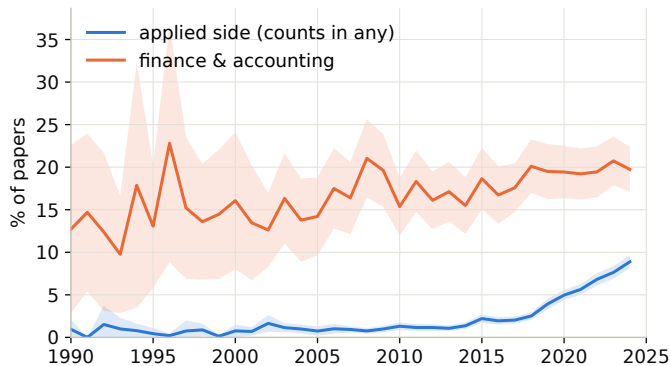
Share of papers flagged, 1990s → 2020s:

- DiD 0.39 % → 12.75 %
- IV 6.93 % → 15.79 %
- structural 4.56 % → 7.72 %
- RCT 0.33 % → 7.29 %
- RDD 0.00 % → 2.63 %
- synthetic control 0.00 % → 1.18 %

As flagged, not corrected;
three-year moving average. Event
studies are on the next slide.

DiD passed structural estimation around 2018 and is closing in on IV.

“Event study” names two designs, and only the applied one took off



Economics journals (the applied design): 0.63 % → 6.85 % of papers.

Finance and accounting (abnormal returns): 14.70 % → 19.75 %. This design was already routine in the 1990s.

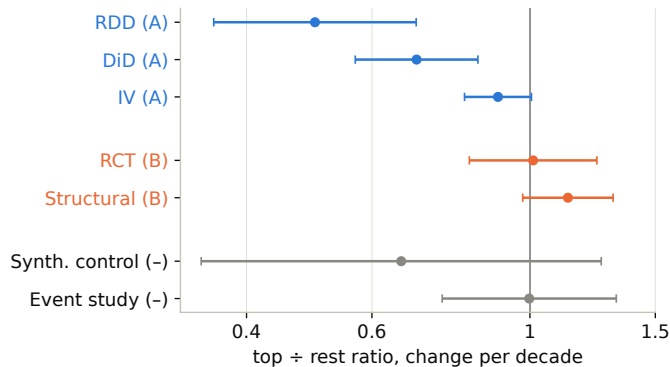
Pooled, the two series would hide the applied design's rise, so “any design” counts only the applied side. Precision: applied 0.867, finance 0.625.

Papers also say more about identification: $2.28\times$ more per word

- Identification phrases per 10,000 words of body text, all seven designs, 1990s $\rightarrow$ 2020s: 6.35 $\rightarrow$ 14.51.
- Phrases per paper grew $3.30\times$, and the median paper got $1.59\times$ longer.
- Of the per-paper growth, 61 % comes from density and 39 % from length.

More papers use the designs, and each one spends more of its text on them.

Lower-tier journals caught up only on the designs that need just data



Change per decade in each design's top-to-rest ratio (top = tiers 1–2, rest = tier 3), with 95 % intervals. Below one: the rest is catching up.

Blue (group A) needs only data. Orange (group B) needs funding, a field partner or a structural model. I fixed the groups before seeing any label; event studies and synthetic control are in neither.

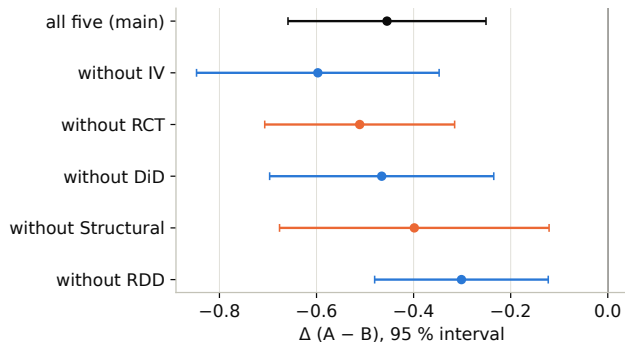
DiD 0.693, RDD 0.499, IV 0.902; RCT 1.010, structural 1.131.

The test fixed in advance: group A catches up 36.5 % faster per decade

$$\Delta = \bar{\beta}_A - \bar{\beta}_B = -0.455 \quad \exp(\Delta) = 0.635$$

- β_m : change per decade in the log top-to-rest ratio of design m (Poisson, journal fixed effects); $\bar{\beta}_A$ and $\bar{\beta}_B$ average over each group.
- Standard error 0.104, clustered by journal with a joint covariance across designs; $p = 1.24 \times 10^{-5}$.
- Randomization inference: $p = 0.000$ in 1000 draws, each relabelling a random set of journals as the top for every design at once.

No single design drives the result, though RDD matters most



Dropping one design[†]: the weakest case is without RDD, $\Delta = -0.301$, randomization $p = 0.006$.

Top five vs the rest (planned check):
 $\Delta = -0.448$, clustered
 $p = 5.92 \times 10^{-5}$, randomization
 $p = 0.142$. Low power by construction: only 5 journals can be the top.

Event studies in group A[†]:
 $\Delta = -0.358$, randomization $p = 0.001$.

Within topic[†]: $\Delta = -0.348$,
randomization $p = 0.011$.

The result is not a shift in research topics[†]

sample and adjustment	Δ	rand. p
all texts (the main test)	−0.455	0.000
texts with a JEL topic	−0.412	0.003
within topic (reweighted)	−0.348	0.011
primary topic only	−0.373	0.004
broad topic groups only	−0.405	0.003
regression adjustment	−0.367	0.000

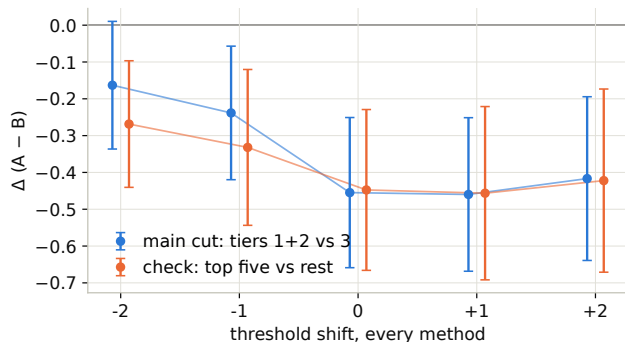
Regression adjustment: clustered p , no randomization.

Topics come from JEL codes; 70.4 % of texts have one.

Within topic, each tier and decade is reweighted to the same topic mix. By the reading rule set before the run, the result survives.

Of the change from all texts to within topic, the smaller sample accounts for −0.043 and the topic mix for −0.064. Holding topics fixed shrinks the gap by about a quarter, and it stays clearly negative.

Lower thresholds weaken the result by adding noise



Moving every threshold down or up together, the test passes at three of five settings and fails at the two lowest.

Three conditions, set in advance, for the papers the lowest settings add:

- 1 mostly false positives: precision 0.114;
- 2 no tier trend of their own: pooled β 0.090, which holds narrowly (margin 0.0095);
- 3 no cleaner at the top than in the rest: gap 0.048.

The classifier is not more accurate at top journals

- **Flagged papers** (precision): top 70.0 % vs rest 78.7 %, a difference of -8.7 pp (SE 9.4 pp).
- **Unflagged papers** (share correctly left out): top 89.7 % vs rest 92.1 %, -2.4 pp (SE 6.5 pp).
- **Logit** of a correct call on top, with design fixed effects (164 draws): coefficient -0.262 (SE 0.526), $p = 0.618$.
- Correcting any design for precision by tier moves the top-to-rest ratio from 1.519 to 1.677 [1.316, 2.154]. A gap that is constant over time cancels from the trend.

Better measurement at the top cannot explain why the gap closes.

Do designs travel with people?[†] A model of each author's first use

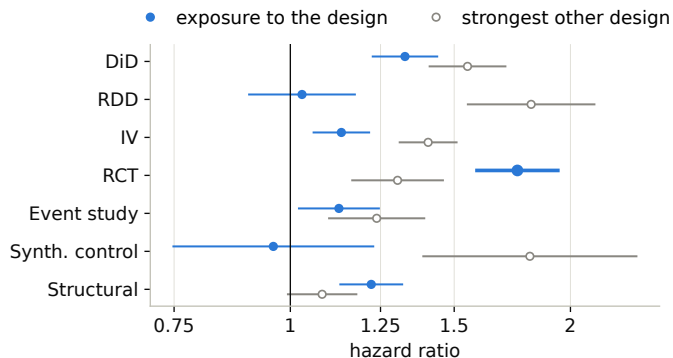
- I follow authors with at least three labelled papers, year by year, until they first use a design. Each design gets its own model.
- Three kinds of exposure, all measured before the year: a past co-author had used the design; some share of the author's colleagues had; the author's PhD advisor (from RePEc Genealogy) had.
- The models control for year, cohort, career age, topic, past journal tier, past output and the share of the author's papers I can label.

	IV	DiD	structural	RCT	event study	RDD	synth. control
first uses	7,457	5,297	5,170	3,189	2,871	1,634	573

Separating a design-specific effect from shared taste

- People who work together tend to like the same things, so a raw correlation proves little.
- First, I drop first uses co-written with an already exposed co-author: that co-author may simply have done that part of the paper.
- Second, every other design is a placebo. Exposure to all seven designs goes into the same model. If a co-author's DiD predicts my first RCT as well as their RCT does, what travels is a taste for identification, not the design.
- Third, for colleagues, the effect has to hold within an employer, not only across employers.
- The rule was written before these checks ran. A channel qualifies only if its own design beats the strongest other one, still does once the author's own past designs are controlled, and, for colleagues, holds within employer.

Co-authors' designs predict adoption, but so do their other designs[†]

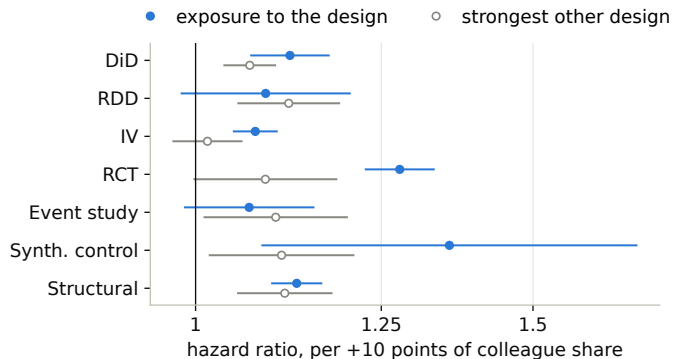


Hazard ratios, one model per design: a past co-author used this design (filled) or the strongest of the six others (hollow). Thick: the channel qualifies.

Only RCTs qualify: 1.75 for a co-author's RCT against 1.30 for their DiD. Elsewhere another design, often DiD, predicts as well, or the gap is not significant.

Co-authors who run their first RCT only later also predict it (1.54 against 1.76), so part of this is shared taste.

Colleagues' designs predict adoption only across employers[†]



Hazard ratio per 10 points of colleague share, against the strongest other design: RCT 1.28 vs 1.09, IV 1.07 vs 1.01.

With employer fixed effects the colleague effect is gone for every design. Per 10 points of colleague share, the annual adoption rate changes by -0.06 points for DiD, -0.44 for IV and $+0.18$ for RCTs.

Early adopters leaving the risk set can also push the within-employer estimate down, so this is no proof that colleagues do nothing.

Students whose advisor used a design adopt it sooner[†]

design	advisor used it, HR [interval]	<i>p</i>
DiD	1.34 [1.09, 1.65]	0.006
RDD	0.92 [0.62, 1.38]	0.700
IV	1.28 [1.07, 1.53]	0.007
RCT	1.62 [1.24, 2.10]	0.000
event study	1.33 [1.01, 1.74]	0.039
synth. control	2.81 [1.00, 7.90]	0.050
structural	1.44 [1.22, 1.70]	0.000

Advisors come from RePEc Genealogy, which I link to the corpus's authors through RePEc person records and shared DOIs. The model is the main one plus an indicator for a known advisor, with the advisor's use measured before the year.

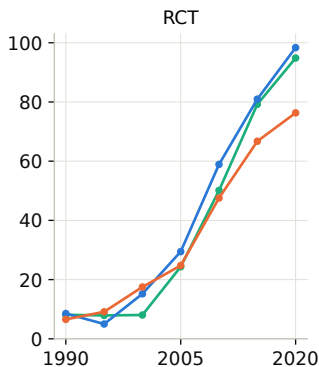
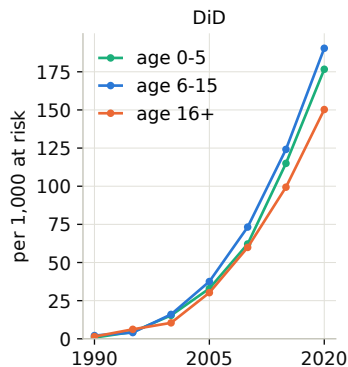
The linked students are few, senior and mostly in the US and Europe, and this channel has no placebo design, so read it as an association.

One channel of 14 qualifies, about what chance would give†

design	co-author			colleagues			within employer	advisor
	own	strongest	other	own	strongest	other		
DiD	1.33	1.55	IV	1.12	1.07	IV	−0.06	1.34
IV	1.13	1.41	DiD	1.07	1.01	DiD	−0.44	1.28
structural	1.22	1.08	IV	1.13	1.11	event study	−0.27	1.44
RCT	1.75	1.30	DiD	1.28	1.09	RDD	+0.18	1.62
RDD	1.03	1.81	DiD	1.09	1.12	DiD	−0.72	0.92
event study	1.13	1.24	synth	1.07	1.10	RDD	−0.64	1.33
synth. control	0.96	1.81	DiD	1.36	1.11	RCT	−0.63	2.81

Hazard ratios of a first paper using the design, from one model per design that holds the exposures to all seven: a past co-author used it; colleague share, per 10 points. *Strongest other*: the largest of the six other designs' ratios, and which design it is. *Within employer*: change in the annual adoption rate, in points, per 10 points of colleague share, with employer fixed effects. Advisor: the main model, no placebo (grey: not significant). Bold: the one qualifying channel. With 14 channels tested at the usual level, about 0.7 would qualify by chance.

Early adopters were mid-career, and central institutions adopted DiD first[†]



First uses per thousand authors at risk, by years since first paper.

- Early adopters within five years of their first paper: 17.9–23.5 %, against 29.1–30.5 % of active authors.
- Central institutions wrote their first DiD paper earlier, net of size: Spearman -0.21 ($p = 0.033$). No such pattern for RDD.

- A phrase list reads words, not designs: structural precision is 0.538, RCT recall 0.538.
- Full text covers 38.1 % of the corpus; the reweighting assumes the rest looks like it on what I observe.
- The 1990s rest on few labels (growth interval $[1.59, 3.78]$), and I chose their weighting after seeing the alternatives.
- The hierarchy result leans on RDD: without it, $\Delta = -0.301$.
- I am the only human labeller; the LLM only flagged disagreements.
- The people results are associations. With every other design as a placebo, 1 channel of 14 stands out, about what chance would give. The advisor channel has no placebo and rests on a small linked sample.

- Corrected for the classifier's mistakes, papers using an identification design went from 17.5 % in the 1990s to 42.6 % in the 2020s, and they spend more of their words on it ($2.28\times$ per word).
- The event-study boom is the applied design. Finance had been running its version all along.
- DiD, RDD and IV reached the rest of the profession faster than RCTs and structural estimation: $\exp(\Delta) = 0.635$ per decade in the test fixed in advance, randomization $p = 0.000$. Dropping any one design or raising the thresholds does not change that, and the classifier is no better at top journals.
- Authors take up the designs their co-authors and colleagues use, but against every other design as a placebo that looks like a shared taste for identification, not a design passed on. Only RCTs through co-authors stand out.[†]

The revolution reached the whole profession where it needed only data. Where it needed money, field access or a model, the top journals' lead did not shrink.

- Measurement and labels
- Prevalence
- The hierarchy: tables and robustness
- Research topics
- Designs and people

Agreement between my labels and the LLM second coder, by design

design	kappa, first round [interval]	kappa, second round [interval]
DiD	0.813 [0.645, 0.932]	0.770 [0.685, 0.848]
RDD	0.746 [0.000, 1.000]	0.846 [0.717, 0.942]
IV	0.836 [0.726, 0.925]	0.781 [0.675, 0.871]
RCT	0.771 [0.519, 0.939]	0.912 [0.818, 0.979]
Event study		0.708 [0.604, 0.808]
Synth. control		0.665 [0.000, 1.000]
Structural	0.585 [0.275, 0.825]	0.726 [0.579, 0.846]
any design	0.781 [0.684, 0.869]	0.746 [0.671, 0.815]

First round: 250 papers; second round: 339 papers. The first round has too few event studies and synthetic-control papers for an informative kappa.

How the disagreements were settled

round	disagreements	my way	the LLM's way	by a written rule
first	43	24	17	2
second	104	85	18	1
both	147	109	35	3

Second round: 280 papers in its main sheet, 45 in a second sheet, and 14 relabelled first-round papers. On the first round alone, precision / recall is 0.853 / 0.829 for IV (35 positives) and 1.000 / 1.000 for synthetic control (1 positive).

Working papers are flagged about as often as their published versions, except for synthetic control

design	ratio r_m [interval]	design	ratio r_m [interval]
DiD	1.020 [0.975, 1.069]	Event study	1.019 [0.950, 1.091]
RDD	1.025 [0.946, 1.114]	Synth. control	1.375 [1.159, 1.708]
IV	1.087 [1.052, 1.127]	Structural	1.042 [0.992, 1.094]
RCT	1.058 [1.000, 1.122]	any design	1.053 [1.033, 1.074]

r_m : the flag rate on a working paper over the rate on its own published version, from papers I have in both forms. A working-paper text is adjusted by r_m before it stands in for the article. IV with a first-stage phrase: 1.088.

Precision by tier differs by design but in no common direction

design	top	rest	labelled papers, top / rest
DiD	0.938	0.792	97 / 158
RDD	0.692	0.900	97 / 158
RCT	0.500	0.875	96 / 158
Structural	0.600	0.500	97 / 158
any design, by decade	0.874	0.792	

Weighted to the labelling sample within each tier. Precision at every threshold is in the paper.

Any design by decade, as flagged and corrected

decade	flagged, unweighted	flagged, analysis sample	corrected [interval]
1990s	10.2 %	11.0 %	17.5 % [11.7, 24.7]
2000s		17.0 %	16.1 % [10.0, 22.6]
2010s		29.1 %	34.3 % [26.0, 42.4]
2020s	40.5 %	39.0 %	42.6 % [33.0, 52.8]
growth	3.97×	3.54×	2.39× [1.59, 3.78]

Analysis sample: reweighted to the corpus, working papers adjusted. Corrected: weighted to the labelling sample's design. With my labels alone growth is 2.28×; with the LLM's alone, 2.63×.

The corrected growth moves little across weightings of the 1990s labels

weighting of the 1990s labels	1990s level [interval]	growth [interval]
combined (main)	17.5 % [11.7, 24.7]	2.39× [1.59, 3.78]
union, cell rates	18.6 % [12.9, 25.8]	2.25× [1.52, 3.56]
literal, cell rates	18.1 % [11.3, 25.8]	2.34× [1.53, 3.97]
literal, draw rates	17.4 % [10.7, 25.3]	2.44× [1.55, 4.14]
original (superseded)	17.6 % [5.3, 28.8]	2.58× [1.41, 8.08]

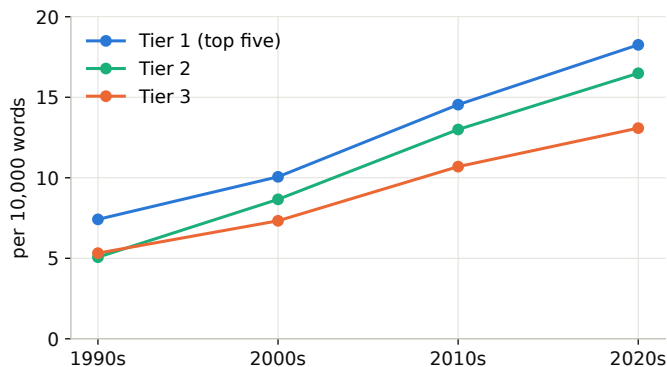
Same labels, bootstrap draws and seed in every row. Because of an error, the original weighting gave the 1990s labels almost no weight. I replaced it, and chose the main weighting, after seeing these numbers.

All three label sets point the same way

label set	corrected growth [interval]	under the original weighting
adjudicated (main)	2.39× [1.59, 3.78]	2.58×
my labels only	2.28× [1.59, 3.63]	2.37×
LLM only	2.63× [1.70, 4.05]	2.43×

Moving from the original to the main weighting, the LLM-only estimate rises ($2.43\times \rightarrow 2.63\times$) while the adjudicated ($2.58\times \rightarrow 2.39\times$) and my-labels-only ($2.37\times \rightarrow 2.28\times$) estimates fall. All three stay within 2.28–2.63×, and all three show the correction lowering growth.

Identification talk rose in every tier



Identification phrases per 10,000 words of body text, all texts, by tier.

Across all tiers: 6.35 $\rightarrow$ 14.51 per 10,000 words. Per paper, talk grew 3.30 $\times$: 61 % from density, 39 % from longer papers (median length 1.59 $\times$).

Tier trend ratios per decade, main cut and top-five cut

design	tiers 1–2 vs tier 3 (main)	top five vs the rest (check)
DiD	0.693 [0.569, 0.845]	0.785 [0.474, 1.300]
RDD	0.499 [0.360, 0.692]	0.619 [0.437, 0.878]
IV	0.902 [0.809, 1.005]	0.999 [0.876, 1.139]
RCT	1.010 [0.822, 1.242]	1.091 [0.901, 1.321]
Structural	1.131 [0.977, 1.309]	1.387 [1.125, 1.709]
Event study	0.998 [0.753, 1.322]	0.844 [0.632, 1.126]
Synth. control	0.660 [0.345, 1.259]	0.788 [0.395, 1.574]

Change per decade in the top-to-rest ratio (Poisson, journal fixed effects); below one, the rest catches up. Grey: in neither group.

The test under both cuts, and with event studies in group A

version	Δ	$\exp(\Delta)$	SE	clustered p	randomization p
main: tiers 1–2 vs tier 3	−0.455	0.635	0.104	1.24×10^{-5}	0.000
top five vs the rest	−0.448	0.639	0.111	5.92×10^{-5}	0.142
event study in A, main cut [†]	−0.358	0.699	0.104	0.000578	0.001
event study in A, top five [†]	−0.430	0.651	0.094	4.45×10^{-6}	0.146

Randomization draws: 1000 (main), 977 (top five). Per decade, group A catches up 36.5 % faster under the main cut and 36.1 % under the top-five cut.

Dropping one design, and moving all thresholds together

design dropped [†]	Δ	rand. p
none (main)	-0.455	0.000
DiD	-0.466	0.001
RDD	-0.301	0.006
IV	-0.597	0.000
RCT	-0.511	0.000
structural	-0.399	0.023

threshold shift	Δ	rand. p	passes
-2	-0.163	0.093	no
-1	-0.238	0.016	no
0	-0.455	0.000	yes
+1	-0.460	0.000	yes
+2	-0.417	0.003	yes

Right: every design's threshold moved by the same number of steps; zero is the chosen thresholds.

The papers added by lower thresholds, design by design

design	precision of the added papers	their own tier trend, β
DiD	0.271	-0.084
RDD	0.000	0.308
IV	0.071	0.047

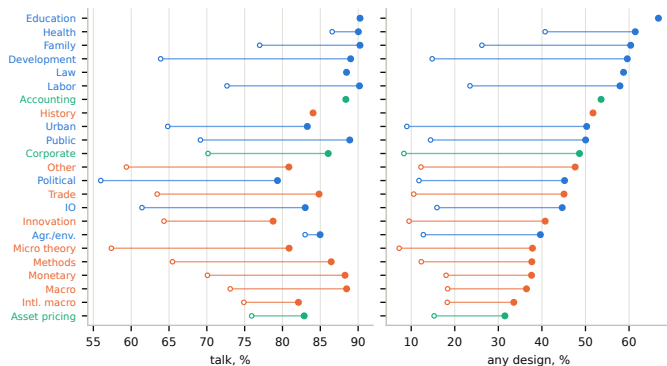
The three conditions, set in advance: (i) pooled precision 0.114; (ii) pooled β 0.090, which holds narrowly (margin 0.0095); (iii) precision gap, top minus rest, 0.048. The added papers are those the two lowest settings flag and the chosen thresholds do not.

Top-to-rest level ratios, as flagged and corrected for precision by tier

design	as flagged	corrected [interval]	precision, top	precision, rest
DiD	2.001	2.657 [1.976, 4.168]	0.938	0.706
RDD	2.555	1.990 [1.179, 2.948]	0.692	0.889
IV	1.509	1.681 [1.365, 2.263]	0.929	0.833
RCT	1.462	0.835 [0.406, 1.300]	0.500	0.875
Event study	1.918	2.398 [1.918, 3.837]	1.000	0.800
Structural	1.788	2.038 [1.143, 3.731]	0.600	0.526
any design	1.519	1.677 [1.316, 2.154]	0.874	0.792

A level gap that is constant over time cancels from the trend. Synthetic control has too few labelled positives at the top to correct.

Designs by research topic†



Topics from JEL codes; 70.4 % of texts have one. Hollow: 1990s; filled: 2020s. As flagged, not corrected; topics sorted by their 2020s share of any design.

Within topic: with each tier and decade reweighted to the same topic mix, $\Delta = -0.348$, randomization $p = 0.011$; on the same papers without reweighting, $\Delta = -0.412$.

Topic coverage, and any design by research area[†]

texts with a JEL topic	share
all	70.4 %
tiers 1–2	80.3 %
tier 3	68.4 %
tier 3, 1990s	25.6 %

any design, as flagged	1990s	2020s
applied micro	17.5 %	48.6 %
macro, theory, other	13.9 %	39.3 %
finance	12.1 %	40.1 %

JEL sources, in order of preference: EconLit, the JEL line printed on the paper, the RePEc article record, the linked RePEc working paper. A paper counts in every area its codes reach.

The classifier's precision does not differ by research area; recall is lowest in finance

area	precision, any design	recall, any design
applied micro	0.830	0.877
macro, theory, other	0.730	0.833
finance	0.763	0.584
no JEL topic	0.829	0.712

Weighted to the labelling sample, adjudicated labels. Equal precision across the three areas: $p = 0.531$ (largest gap 0.099); equal recall: $p = 0.107$.

Five associations, stable across samples[†]

channel	main	all first uses	published texts	labelled	tiers 1–2
DiD, colleagues	1.14	1.15	1.15	1.18	1.13
IV, co-authors	1.22	1.59	1.22	1.15	1.23
IV, colleagues	1.07	1.07	1.06	1.08	1.07
structural, co-authors	1.23	1.63	1.24	1.13	1.16
structural, colleagues	1.14	1.15	1.14	1.17	1.16

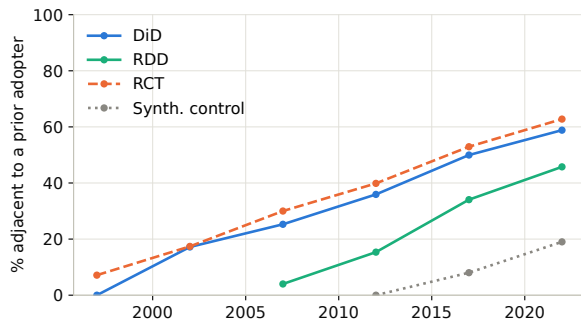
Hazard ratios; colleague shares per 10 points. *All first uses* keeps those written with an exposed co-author; *published texts*: authors whose texts are mostly published versions; *labelled*: authors with most papers labelled; *tiers 1–2*: authors publishing mostly there. The weakest, structural co-authors among well-labelled authors, has $p = 0.050$. These five channels passed an earlier test against one paired placebo design, now withdrawn; none qualifies once every other design is a placebo, so read them as associations.

The timing placebo: co-authors who adopt only later[†]

design	past co-author used it	co-author adopts only later	p , later
DiD	1.49	1.24	0.000
RDD	1.30	1.38	0.000
IV	1.22	1.07	0.022
RCT	1.76	1.54	0.000
Event study	1.28	1.24	0.000
Synth. control	1.10	1.62	0.000
Structural	1.22	1.11	0.004

Both in one model. A co-author who adopts only later cannot have passed the design on, so that ratio measures shared taste, and perhaps the author passing the design to the co-author; transmission predicts a larger ratio for past exposure. It is larger for every design except RDD and synthetic control. The placebo designs on the main slides are the sharper test.

New adopters sit ever closer to earlier ones in the co-author network[†]



Among new adopters with a path to an earlier adopter, the share who had co-authored with one, in five-year windows from **2005–2009** to **2020–2024**:

- DiD 25.3 % → 58.8 %
- RCT 30.0 % → 62.8 %
- RDD 4.0 % → 45.8 %
- synthetic control 0.0 % → 19.0 %

Descriptive: as a design spreads, more of the network has used it. The early RDD and synthetic-control windows have few adopters.