Does Empirical Finance Replicate?*

Dmitriy Muravyev
University of Illinois Urbana-Champaign

First draft: August 18, 2026
This version: August 27, 2026

Preliminary draft. Please do not cite or circulate without permission.

Abstract

I attempt to replicate the main result of every empirical paper published in the *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* from 2000 to 2020, first in each paper's original sample and then in later years. An autonomous LLM pipeline rebuilds each result from the paper's text and available databases, and independent audits verify each replication. I obtain 1,584 in-sample replications, 1,005 with enough post-sample data for an out-of-sample test. In sample, key results reproduce their published sign and significance in 75% of papers, and the median coefficient recovers 0.95 of the published one. Out of sample, fewer than half remain significant with their original sign, and the median coefficient falls to 0.45 of its published value. Results with higher original t-statistics reproduce better and decay less, while other paper characteristics matter far less. Published estimates largely reproduce in sample but lose about half their magnitude out of sample.

JEL classification: G00; G10; G30; C18
Keywords: replication; reproducibility; out-of-sample; publication bias; large language models

*Muravyev: Gies College of Business, University of Illinois Urbana-Champaign, Champaign, IL, United States. Email dmuravy2@illinois.edu. I am especially grateful to Andrew Chen for encouraging me to push this project forward and for generously sharing his LLM harness. I thank Hanzhao (Justin) Yu for excellent research assistance. I also thank Vincent Bogousslavsky, Ivan Ivanov, Neil Pearson, and Josh Pollet for helpful comments and discussions. I used generative AI tools to assist with editing and improving the exposition of the manuscript. All errors are my own.

1 Introduction

Published results in finance are used long after their samples end. A relation estimated on data through 2000 may still be cited, calibrated to, traded on, or used in policy today. Whether that reliance is warranted depends on two questions. Can the result be reproduced from the paper and available data? And if so, does it still hold in the years after the original sample ends? There are reasons to doubt both. Psychology re-ran a hundred of its prominent experiments and fewer than half held up, while economics has struggled to recover published results even from the authors' own data.[1] Finance has additional reasons to worry, as investors learn from published findings and markets and institutions change over time. Harvey (2017) warns that "many of the results being published will fail to hold up in the future," and Hou et al. (2020) argue that most published anomalies fail to replicate. Yet the direct evidence in finance comes almost entirely from stock-return predictors. For the rest of empirical finance, little is known about what reproduces or holds up out of sample.

I study both reproducibility and out-of-sample decay across empirical finance. Finance is especially well suited to the latter because many of its standard databases continue long after a paper's sample ends. I screen every article published in the *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* between 2000 and 2020, 5,430 in all. Together, these journals form the core published record of modern academic finance. For each empirical paper, an autonomous LLM pipeline identifies the main result, the table that carries it, and the table's key coefficients. Without human intervention, it rebuilds the sample from available databases and re-estimates those coefficients in the paper's sample and in later years. I call these the in-sample and out-of-sample replications.[2] Validation passes then check each replication and rerun it when problems appear. I obtain 1,584 in-sample replications, 1,005 with at least five years of post-sample data for an out-of-sample test. Data access rules out many other papers; 16% use confidential or hand-collected data, and 19% require commercial data I do not have. I am not aware of a previous field-wide estimate of these shares in finance.[3] The sample spans the major fields of finance and is about ten times larger than previous large-scale replication efforts in economics.[4]

[1] Open Science Collaboration (2015) replicated 100 psychology studies and reproduced a significant result in 36%. In economics, Dewald et al. (1986) could not recover most results from the authors' own data and code, and Chang and Li (2022) reproduce fewer than half of the macroeconomics papers they attempt.

[2] The replication literature (Nosek et al., 2022) would call the in-sample analysis a *reproduction* and the out-of-sample analysis a *replication*. I use the more descriptive terms in-sample and out-of-sample throughout.

[3] Another 18% of articles are theoretical. Additional papers are excluded when the rebuilt sample cannot adequately match the paper's sample or the available data prove insufficient.

[4] Brodeur et al. (2026) reproduce 110 papers, Herbert et al. (2024) 152, and Chen and Zimmermann (2022) rebuild about 200 published return predictors.

First, published finance is largely reproducible in sample. For each paper, I identify about three key coefficients supporting its main result and measure whether they reproduce their sign and significance (a category match).[5] Averaging within papers, 75% of published results reproduce by this measure. Only 1% come out significant with the opposite sign. Coefficient magnitudes are also close on average. The median replicated coefficient is 0.95 of the published one, so most failures reflect weaker statistical significance rather than smaller estimated effects.[6] Some failures may reflect imperfect data matching rather than a failure of the published result itself. I use the original data when available and close substitutes otherwise, and different database vintages can still make the rebuilt in-sample data differ from those originally used. A failed in-sample replication is therefore not, by itself, evidence that the published result is wrong; the 75% rate is intended to characterize reproducibility across the literature. By the standards of psychology and experimental economics, where roughly half of results repeat, the in-sample reproducibility of finance is high, though far from complete.

Second, published estimates decay sharply in later data. The category-match rate falls from 75% in the original sample to 42% in subsequent years. Some of that decline is expected because post-sample windows are shorter on average. Adjusting post-sample t-statistics to the original sample's precision raises the survival rate to 48%.[7] The decline also appears in coefficient magnitudes, which are not mechanically affected by shorter post-sample windows. The median coefficient ratio falls from 0.95 in sample to 0.45 out of sample, or 0.59 relative to the pipeline's own in-sample estimate. Because the same code recovers the published magnitudes in the original years and the later run changes only the sample period, replication error is unlikely to explain much of the average decline. Published finance is therefore largely reproducible in its original samples, while its estimates retain about half their magnitude afterward.

Third, stronger published results are more likely to reproduce in sample and to persist out of sample. Sorting papers by their original t-statistics shows a strong gradient. The in-sample reproduction rate rises from 56% in the weakest quartile to 89% in the strongest, while the out-of-sample rate rises from 32% to 72%. Some of this difference is mechanical because stronger results can weaken more before losing significance. Coefficient magnitudes confirm the pattern. The weakest results retain only about one-third of their published magnitude

[5]Targets are fixed before replication estimates are produced, using the coefficients the paper itself relies on for its main finding; controls are excluded and a blind re-audit verifies the selection. Replication measures are averaged within paper so that each paper receives equal weight.

[6]Memorized replication code is unlikely to explain the high in-sample rate as most papers predate public replication packages, later papers reproduce less well, and highly cited papers reproduce only slightly better.

[7]The adjustment evaluates each out-of-sample coefficient at the standard error implied by the in-sample replication. Section 2 gives details.

out of sample, compared with nearly two-thirds for the strongest. Weak published results therefore deteriorate on both dimensions. Replication studies in psychology and experimental economics find a similar relation.[8]

Fourth, standard models of publication selection explain only about a fifth of the out-of-sample decay. I apply the empirical-Bayes approach of Chen and Zimmermann (2020) and the selection model of Andrews and Kasy (2019) to the published t-statistics in my sample.[9] Both models predict median coefficient retention of about 0.90, versus 0.45 in the data. For published return predictors, the same procedure recovers the modest publication bias found in prior work. The models capture selection only through the reported statistic, so specification search may generate additional bias they cannot detect.

Fifth, the decay is concentrated at the end of the original sample, which points to data snooping rather than learning or gradual change. Re-estimated one year at a time with the specification frozen, the coefficient ratio is flat within the original sample, drops by 0.28 once the sample ends, and is flat afterward. The ratio does not fall further after publication, which suggests that learning from published findings is not the main channel, and it shows no trend on either side of the boundary, which leaves a limited role for gradual change in markets. The step is largest in asset pricing, 0.41 against 0.12 elsewhere, where noisy returns make it easier to find results that do not hold up. Since selection on the reported t-statistic explains only a fifth of the decline, most of the snooping occurs earlier, in the choice of samples and specifications.

Beyond original strength, observable paper characteristics explain surprisingly little. In sample, the clearest differences appear to reflect how easy a paper is to reconstruct. Asset-pricing papers reproduce at 81%, compared with 69% in corporate finance, despite nearly identical original statistical strength. Earlier papers and papers from top-five institutions also reproduce somewhat better. These differences largely fade when I turn to durability. Out of sample, coefficient decay is broad across fields, with median published-magnitude retention ranging from 0.42 in asset pricing to 0.85 in household finance and other fields. Citations, author affiliation, and team size provide little information about out-of-sample durability. Even jointly, these characteristics explain little of the paper-to-paper variation in decay, with original statistical strength remaining the strongest predictor.

The use of LLMs makes validation central to the design. After the replication protocol was fixed, a script starts a fresh agent for each paper with the same instructions. The agent

[8] Open Science Collaboration (2015), Camerer et al. (2016), and Altmejd et al. (2019).

[9] The original methods study estimates in comparable units. To accommodate heterogeneous empirical designs, I express effects in t-statistic units, orient them so the published effect is positive, and limit the influence of extreme t-statistics. Both models are estimated using only published in-sample statistics and then used to predict shrinkage in later data.

reads the paper, obtains data, writes and runs code, and completes the replication. The model acts as an engineer, translating the paper's description of its data and specification into executable code. It follows the authors' methodology, records every inferred choice, and preserves its first estimate before comparing it with the published table. Every completed replication then passes through independent validation and audit. Auditors read the paper and code, correct errors, and rerun the analysis when necessary. They cannot use closeness to the published estimate as a reason for a change, and their corrections move coefficients toward the published values only slightly more often than away from them. I also personally inspected roughly one hundred completed packages. A separate extraction pass checks every published and replicated number used in the analysis. Category matches, coefficient ratios, and other replication measures are then computed deterministically.

These safeguards do not make any individual replication definitive. A particular paper may still be imperfectly reconstructed and should be interpreted with caution. If the remaining replication errors are idiosyncratic, however, they should largely cancel across 1,584 papers and have little effect on the aggregate conclusions. The asset-pricing subsample provides an external benchmark. Replicated t-statistics track published values with an R^2 of 75%, compared with 82% for the hand-coded predictors of Chen and Zimmermann (2022), while the estimated slope is closer to one.

My first contribution is to the literature on replication and reproducibility. Prior work shows that published findings can fail to reproduce even with the authors' data and code.[10] In finance, Pérignon et al. (2024) rerun research teams' own code on the same data and regenerate the reported results only about half the time. Large-scale exercises that reconstruct published results from standard databases have focused primarily on asset pricing (McLean and Pontiff, 2016; Hou et al., 2020; Chen and Zimmermann, 2022; Jensen et al., 2023). I rebuild results from the paper itself, without the authors' code, across a broad range of empirical designs in finance. This is a stricter test because the code resolves many choices that the paper leaves implicit, yet three-quarters of results reproduce in sign and significance.

My second contribution is to the literature on out-of-sample decay in finance. McLean and Pontiff (2016) show that published return predictors weaken outside their original samples and further after publication, separating statistical bias from investor learning. Subsequent work studies the robustness and persistence of return predictors across samples, periods, and markets.[11] I extend this question from return predictors to empirical finance broadly,

[10]Major replication exercises include Open Science Collaboration (2015); Camerer et al. (2016, 2018); Klein et al. (2018); Nosek et al. (2022) in psychology and experimental economics and Dewald et al. (1986); Chang and Li (2022); Herbert et al. (2024); Brodeur et al. (2026) in economics. Christensen and Miguel (2018) survey reproducibility and credibility in economics. Menkveld et al. (2024) and Huntington-Klein et al. (2021) show that researchers given the same data and question can reach different estimates.

[11]See Harvey et al. (2016); Hou et al. (2020); Chen and Zimmermann (2022); Jensen et al. (2023) on large

preserving each paper's original specification. Published magnitudes decline substantially out of sample in every major field, typically by about half, and the decline is greatest when the original evidence is weakest. The decay of published return predictors is therefore one instance of a field-wide pattern.

My third contribution is methodological. Recent work shows that generative AI can automate substantial parts of economic research, but also that individual outputs remain imperfect and require validation.[12] I show how imperfect agent replications can still support credible aggregate evidence. Independent audits reduce errors in each replication, preserved code and output allow any replication to be inspected and rerun, and with 1,584 papers, remaining errors should largely cancel if they are idiosyncratic. This makes systematic replication feasible at a scale that would be difficult to reach manually.

2 Data and Methodology

The pipeline has four stages. It assembles the papers, rebuilds each main result in the original and later samples, audits the replication, and extracts the numbers used in the analysis. Two rules govern every stage. Each choice is recorded and every reported number can be reproduced from code. The pipeline is never adjusted because an estimate is close to or far from the published result.

2.1 Papers and text

The universe is every research article published in the *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* from 2000 through 2020. Publisher metadata identify 5,455 articles, of which 5,430 remain after excluding biographical essays, presidential addresses, reviews, errata, and comments (Table 1). I obtain the articles directly from the publishers under institutional subscriptions, including full-text XML for the *JFE*. Defining the universe before any replication is attempted avoids the selection inherent in studying a hand-picked set of papers.

collections of return predictors and their robustness, Chen et al. (2025) on post-sample performance of peer-reviewed and mechanically data-mined predictors, Linnainmaa and Roberts (2018) on pre-sample evidence, Welch and Goyal (2008) on out-of-sample return prediction, Jacobs and Müller (2020) on post-publication performance internationally, and Schwert (2003); Chordia et al. (2014); Green et al. (2017) on the erosion of anomalies over time. Chen (2021) studies how publication selection and specification search depend on the strength of the original evidence.

[12] Korinek (2023, 2024) discuss the use of generative AI across the research process. Ludwig et al. (2026) develop a framework for using LLM outputs in empirical work, while Novy-Marx and Velikov (2026) illustrate the scale at which AI can generate finance research. Hu et al. (2025); Brodeur et al. (2025); Kohler et al. (2026) document substantial variation in agent replication performance.

Each article is linked directly to its publisher record, without title matching. I use the printed-issue year from that record as the publication year throughout the paper rather than dates from bibliometric databases, which often reflect advance access or working-paper versions. For the same articles, OpenAlex disagrees with the printed-issue year for roughly 29% of the corpus. In an earlier version of the sample, these dates admitted 47 articles actually published in 2021–2023 and omitted one article published in 2002.[13] The publication year determines both whether a paper enters the 2000–2020 universe and how many post-sample years are available for its out-of-sample test.

The model reads each paper as text, so I convert every PDF to page-by-page text using Mistral OCR before replication begins. The conversion is generally highly accurate, and tables are preserved well enough for the pipeline to recover reported estimates and sample statistics directly from the text. Figures are not used. To guard against the rare conversion error, the verification stage checks the target estimates, signs, and sample sizes against the original PDF and reruns the conversion when needed. The model also receives the paper's internet appendix when one is available.

2.2 Data

Data come from three sources. The first is WRDS, which the pipeline queries directly under my subscription.[14] The second is a local cache of frequently used data, including CRSP, Compustat, and IBES, downloaded in advance to avoid repeatedly querying WRDS. It also contains extracts from other commercial databases. An index describes the available files so that the agent checks local data before downloading anything. The third source is public data available on the internet. These include researcher-maintained series such as Baker and Wurgler's investor sentiment index and Shiller's CAPE data, as well as government and agency datasets. Before treating a dataset as unavailable, the agent searches for a public-use version. Public downloads must be needed for the paper's central result and available without restricted access.[15]

Confidential, hand-collected, and author-owned data are not approximated with superficially similar datasets. A substitute is allowed only when it measures the same economic object and provides substantial coverage of the paper's sample and period. Every substitution is recorded. If no suitable data are available, the replication stops and records the reason.

[13] The discrepancies can be large for advance-access articles, reaching six years for some *RFS* papers, and entire *JFE* volumes are shifted by one year. Moving from bibliometric to publisher dates changed the headline results very little.

[14] The research harness supports authenticated WRDS queries. I thank Andrew Chen for sharing the implementation on which this feature is based.

[15] WRDS sample and trial libraries are never used as replication data.

The pipeline therefore uses data that an outside researcher with comparable subscriptions and public access could obtain.

This data boundary removes most papers from the replication sample. Of the 5,430 research articles, about 18% are theoretical or have no suitable empirical table, 16% require confidential, hand-collected, or author-owned data, and 19% require commercial data outside my subscriptions. Another 11% fail for other reasons, most often because an attempted data substitute does not reproduce the paper's sample closely enough. Other cases involve unsupported estimators, public data that cannot be accessed programmatically, or replications that run but fail to recover the paper's sample or estimates. Papers whose computation cannot be completed within the time budget described below also enter this category, including many studies that require very large transaction-level datasets (e.g., TAQ).

Each excluded paper has a recorded reason, and Table 1 is constructed directly from these classifications. The pipeline runs end to end for the remaining 1,967 papers, about one-third of the original universe.

Data substitution is common but concentrated in a few recurring cases. Of the 1,584 papers, 562 use at least one substituted or proxied core dataset. A typical paper uses three core sources, and 15% of these sources are substitutes or proxies. The most common cases involve international equity data, proprietary bond data, author-assembled event samples, and intraday market data. For example, Datastream and Worldscope are often replaced by Compustat Global or CRSP, while proprietary bond and ratings data are replaced by FISD, TRACE, or public Treasury data. Other recurring substitutions use SDC, ExecuComp, or ISS to reconstruct event samples, or daily CRSP measures when the original intraday construction cannot be rebuilt.[16] Papers that use a substituted core source reproduce somewhat worse in sample, as Section 3 shows.

No single missing database explains the papers that cannot be attempted. Among the roughly 1,000 papers stopped for lack of commercial data (Table 1), the largest cluster requires international market data from Datastream, Worldscope, or Bloomberg, affecting about 200 papers. Other recurring gaps include ISS Voting Analytics, Morningstar data beyond CRSP coverage, news sources such as Factiva and RavenPack, Ancerno institutional trades, and pre-TRACE bond prices. Beyond these, the missing commercial data are highly dispersed across more than a hundred products, most of which block only a handful of pa-

[16] Counts of affected papers: Datastream or Worldscope, 67; proprietary bond and ratings files such as the Lehman/Barclays indices and Bloomberg or Moody's issue data, 43; author-collected event and governance samples such as SEO and spin-off lists, CEO turnovers, and proxy-statement peer groups, about 60; intraday ISSM or TAQ constructions, 30. Less frequent cases replace Morningstar or Lipper fund data with the CRSP mutual fund database and multi-vendor hedge fund samples with TASS alone (about 20), and older analyst or holdings databases such as First Call, Zacks, and CDA/Spectrum with I/B/E/S or Thomson 13F (8).

pers. Confidential and author-collected data are a larger barrier than any individual vendor, stopping about 860 papers. Some apparent data failures instead reflect computational scale. In particular, some intraday TAQ and large SDC-based studies could in principle be attempted with the available subscriptions but exceed the pipeline's two-hour computation budget described below.

2.3 Replication

For each paper, an LLM agent first determines which central empirical results can be rebuilt with the available data. It then selects the feasible table that best represents the paper's main claim and identifies the key coefficients and sample statistics needed to evaluate the replication. If no central result can be rebuilt, the attempt stops and the reason is recorded. Before producing any estimates, the agent writes a detailed methodology covering the sample, variables, merges, estimation method, fixed effects, and standard errors. Each choice is tied to the relevant passage in the paper, and the agent follows the authors' design even where it regards a choice as inferior, recording the concern rather than correcting it. This rule is enforced even when it costs closeness. In one paper, for example, the text states repeatedly that portfolios are value-weighted; the value-weighted replication is far from the published coefficients, while equal-weighted portfolios come fairly close. The replication keeps the value weights because that is what the paper says, and the discrepancy is recorded rather than resolved by choosing the specification that matches. Choices the paper leaves unresolved are marked as inferred and receive additional scrutiny during validation. Appendix A reproduces the full instructions given to the agent.

Production replications run autonomously from this point.[17] A Python script launches a fresh agent with the paper, the common research environment, and the fixed replication instructions. The agent then reads the paper, obtains the data, writes and runs the code, and completes the replication; the production replication stage runs end to end without paper-specific human intervention. The lack of human intervention during replication does not mean the resulting package is accepted automatically: completed packages then enter the independent validation and audit stages described below.

Each paper is handled in a fresh session, so the agent has no conversational history from earlier replications and receives no paper-specific instructions beyond the common protocol. It can use the shared tools, data indexes, variable recipes, and procedural lessons accumulated across replications, such as WRDS download utilities and general lessons about merges and

[17] I developed the replication protocol through a small set of guided pilot replications, revising the instructions based on observed failures and feedback from the models. After the protocol was finalized, production replications were launched automatically and completed without paper-specific human intervention.

data construction, but not conclusions or guidance from earlier papers.[18]

Each replication has a computation budget of two hours for building the sample and producing the first estimate; the agent must estimate the runtime in advance and stop if a reasonably optimized implementation cannot fit. The budget binds rarely, stopping 14 papers, almost all of them transaction-level microstructure studies. The agent then builds the sample in a single reproducible script and checks it against the paper. It compares observation counts at the available levels of aggregation and the summary statistics reported by the authors. If the rebuilt sample differs from the original by more than 30%, the agent may try at most three substantively different constructions. Each change must be justified by the paper or by an explicitly recorded inference. If none produces an adequate match, the replication stops.

Once the sample is fixed, the agent runs the target specification and permanently saves the resulting estimates and code before looking at the published coefficients. Revisions are allowed only afterward and at most three times. Each revision must correct an assumption previously marked as inferred or follow from a specific passage in the paper that the first implementation missed. Moving closer to the published estimate is never a valid reason for a change. These rules prevent the pipeline from searching across plausible implementations until it finds the published number.

The replication stops when the key coefficients reproduce in sign and statistical significance, when essential data are unavailable, or when the prespecified sample or specification attempts are exhausted. Successes and failures are treated alike. The complete record is preserved, including the code, constructed data, estimates, data sources, inferred choices, and each revision. Published numbers are stored separately as comparison values and can never enter fields containing replication estimates. The agent's own assessment of replication quality is also recorded but is not used in the analysis. The resulting package then passes through the three validation stages described below.

For papers with enough later data, the out-of-sample test begins only after the in-sample replication is finalized. The same data construction and specification are applied to subsequent years. The agent cannot alter the specification in response to the out-of-sample results. The purpose is to ask whether the published result survives in new data, not whether a modified version of it can be made to survive.

[18] All replications run in a common Docker environment adapted from Andrew Chen's *Ralph Wiggum* research harness, with the required programming and data-analysis software installed and authenticated access to WRDS. Replications use the most capable Claude and GPT models available at the time at high reasoning effort. The most recent runs use Claude Opus 5 and GPT-5.6 Sol. Pilot replications at medium effort produced materially weaker implementations.

2.4 Verification and audit

Every completed replication is first verified by an independent agent from a different model family, and I personally inspected about one hundred replications during the project.[19] The verifier is instructed to improve the replication rather than defend the existing one. It reads the paper, methodology, and code independently and treats the original agent's notes as evidence rather than as conclusions. It begins with choices marked as inferred and then checks the sample construction, where many consequential errors arise. Common problems include duplicate observations after merges, incorrect match rates, lag or weighting errors, and changes in data coverage relative to the paper's original sample. When it finds a problem, the verifier corrects the code and reruns both the in-sample and out-of-sample estimates. It also expands the replication to the full target table when the original attempt covered only the key coefficients. The verified package records what reproduces, what remains partial, and whether remaining differences reflect implementation choices or unavailable data. Failed and partial replications are retained alongside successful ones.

A separate extraction pass determines which numbers enter the analysis. It does not rerun the replication. Instead, it reads the paper and the verified output and matches each replicated statistic to the corresponding published statistic in the target table. For each match it records the coefficient and the available measure of inference in common units. It then identifies the one to four key statistics that support the paper's central claim. Every published value is checked directly against the paper rather than taken from the replication report. This safeguard was added after an early audit found that the archived code reproduced its own results reliably, while the reporting layer sometimes contained stale or mistranscribed published values that made the replication appear closer than it was.[20] Any package with an unresolved discrepancy that could change a result's classification is returned for further verification and excluded if the discrepancy cannot be resolved. From this point onward, all replication measures are computed mechanically from the extracted numbers. No model judges how close a replication is.

Because individual agents still make mistakes, the completed packages went through

[19] Verification and later audits were run mainly with OpenAI models, with Claude Opus used for a calibration subsample. Using different model families for replication and verification reduces the chance that both share the same blind spot. Table A2 shows that replication outcomes do not differ materially with the model family used for validation.

[20] In that audit, rerunning the archived code reproduced the archived replication estimates in every case examined. About half of the packages contained some reporting discrepancy, and four contained a published value whose transcription changed the sign-and-significance category. The extraction pass therefore treats the paper's own table as the source of record and logs discrepancies including mistranscribed published values, stale replication numbers, incorrectly paired specifications, and incorrect statements about what replicated.

several additional audit waves, culminating in a blind re-audit of every replication by an agent that had not produced it. The waves were sequential because more capable models became available during the project, with each wave re-examining the packages produced by the previous one.[21]

The auditor first asks whether the archived code faithfully implements the paper's method for the chosen target, using appropriate data and sample construction, and whether the archived estimates were actually produced by that code. It reads the paper, methodology, code, and data structure and records every discrepancy before seeing the replication results. That list is then frozen. Any later correction must address an item already on the list, and closeness to the published estimate cannot justify a change. A faithful implementation that differs sharply from the paper remains a valid replication, while an implementation that matches the paper for the wrong reason is corrected. The auditor also reruns the code, verifies claims that required data are unavailable, and checks whether the reconstructed sample is adequate to test the paper's claim. Material problems are corrected and rerun. Skipped out-of-sample tests are likewise reviewed, and the test is added whenever the data permit it.

The exhaustive audit changed the code or sample construction in most packages, but usually changed the key estimates only modestly. Among roughly 1,500 packages, 862 required a substantial rebuild, 457 a narrower correction, and 108 no change. [22] When estimates did change materially, they moved toward the published result about as often as away from it. Papers were dropped when the audit concluded, based only on data availability and construction, that the available data could not meaningfully test the paper's claim.[23] A final cross-family check addressed differences between auditors. OpenAI models classified data as insufficient much more often than Claude models, so the Claude-audited packages with the strongest warning signs were re-examined by the other model family, leading to sixteen additional exclusions.

A separate outcome-blind pass measures how completely each paper documents its em-

[21] Two earlier targeted waves reviewed 95 papers flagged for methodological problems and 175 papers with unusually poor replication outcomes. The second deliberately searched only for false failures, cases in which the replication made the published result look worse than the available data warranted. Suspiciously good matches were left for the exhaustive audit.

[22] These counts cover the packages that existed at the time of the exhaustive audit. About 400 papers that I had initially missed were replicated afterwards and passed through the same replication, verification, and blind-audit stages in an accelerated run; they are flagged in the data and are included in all results below.

[23] The dropped papers reproduced substantially worse in sample than the retained papers, with category matches of 59% and 75%, respectively, but no differently out of sample. This pattern is consistent with inadequate data adding noise rather than the audit selectively removing inconvenient results. Packages that had been repaired in earlier audit waves were also more likely to need no further change in the exhaustive audit than packages never repaired before, 21% versus 6%.

pirical design and how faithfully the final code follows it. Using a fixed inventory of 22 methodological choices, an agent first records what the code does and freezes that record before reading the paper. It then records whether each choice is stated explicitly, stated only in part, delegated to an external source, or omitted. Deterministic code compares the two records and classifies each choice as exact, consistent with the paper, a data substitution, an approximation, a contradiction, or an omission. These comparisons produce measures of documentation completeness, design complexity, and implementation fidelity. A blind second extraction of a 5% sample checks for drift. These measures enter the analysis of what predicts reproduction, including the indicator for a substitute core data source in Table 5 and the splits in Table A2.

2.5 Sample and measures

The analysis sample applies a small set of screens so that each replication provides a meaningful test of the published result. Each paper enters only once and must have an in-sample benchmark, a rebuilt sample that reasonably matches the paper's original sample and period, adequate data, and no unresolved integrity issue. Papers whose central claim is a null result are excluded from the headline analysis because there is no estimated effect whose magnitude can decay. These screens leave 1,584 in-sample replications. The out-of-sample analysis additionally requires that the new sample begin after the published sample ends and span at least five years, leaving 1,005 papers with valid tests on the same key estimates used in sample. A review of the out-of-sample legs found 82 papers whose later sample measures a key variable differently from the original sample, for example because a data source or linkage changes at the boundary. The 11 papers whose dependent variable is a different object out of sample are excluded from the sample entirely, and 39 papers with a material switch in a key regressor, sample, or linkage keep their in-sample replication but are excluded from all out-of-sample analysis. Table 1 reports the full selection funnel, and Figure 1 shows coverage by year.

A separate extraction pass records the published and replicated values for each key estimate. It checks the relevant cells directly against the paper and the verified replication output and puts both sides in common units. When a paper does not report a t-statistic directly, the pass reconstructs it from the reported standard error, p-value, confidence interval, or other available inference statistic. This pass only identifies and records the correct numbers. It does not judge whether a replication is close. Once the values are recorded, every measure is computed deterministically, without LLM judgment. Measures are first constructed for each key estimate and then averaged within paper, so that a paper with

many key coefficients receives the same weight as a paper with one. Table A1 summarizes the definitions.

The first measure is category match. Each published and replicated estimate is classified as negative and significant, insignificant, or positive and significant, using the 5% level or the paper's stated one-sided convention when applicable. A key estimate matches when the two values fall in the same category. The coefficient ratio measures magnitude. It divides the replicated coefficient by the published coefficient after orienting the published sign to be positive, so that one denotes the same magnitude, a value between zero and one a smaller effect in the same direction, and a negative value a reversal. Ratios are winsorized at the pooled 1st and 99th percentiles and averaged within paper, and tables report the median across papers.[24] I summarize the strength of the original evidence with a synthetic t-statistic, defined as the median of the oriented t-statistics of a paper's key estimates and winsorized at the same percentiles. Its published value is fixed independently of the replication outcome and is used to sort papers by original strength in Table 3. It is a descriptive index of the evidence supporting the paper's central result, not a test statistic for the paper as a whole. Out-of-sample measures use the same key estimates identified from the published paper and restrict attention to those that are significant in the original result, so that there is an effect whose survival can be evaluated. An out-of-sample estimate need not have reproduced successfully in sample to enter this comparison.

Out-of-sample windows are usually shorter than the original samples, so later estimates are less precise even when the underlying effect is unchanged. The median out-of-sample sample contains 58% as many observations as the published sample (Table 2). I therefore report an adjusted significance measure that holds precision fixed at its in-sample level. For each coefficient, I recover the in-sample standard error as $|\hat{\beta}_{IS}/t_{IS}|$ and divide the out-of-sample coefficient by that value. The resulting statistic asks whether the later coefficient would remain significant if it were estimated with the same precision as the in-sample coefficient. Because the standard error comes from the replication itself, the adjustment preserves the paper's original inference procedure.[25]

This adjustment is preferable to rescaling by sample size because the effective number of independent observations is often unclear under clustering or Fama–MacBeth estimation. Where both approaches can be computed, they give very similar adjustments, with median

[24]The in-sample ratio in Table 2 uses the within-paper geometric mean over estimates that retain the published sign, so that halving and doubling receive symmetric treatment. It therefore measures magnitude conditional on preserving direction. The out-of-sample ratios and the ratios used in sorts and regressions use the oriented arithmetic mean, so that reversals enter as negative values.

[25]Estimates with $|t| < 0.1$ on either side retain their raw t-statistics because their implied standard errors are dominated by rounding and no plausible adjustment changes their significance category.

factors of 0.78 and 0.81. The main limitation is that changes in residual volatility are treated as changes in precision rather than changes in the underlying effect. I therefore also report the raw out-of-sample results and coefficient ratios, which require no precision adjustment. As a complementary benchmark, I evaluate the in-sample coefficient at the out-of-sample precision, which shows how often an unchanged coefficient would remain significant in the shorter later window. The window-adjusted category match is used when comparing out-of-sample significance across papers in Tables 3 and 5; Table 2 reports the raw out-of-sample match as the headline measure.

The paper characteristics in Tables 4 and 5 come from the publisher record, OpenAlex, and a separate classification pass. Publication year is the printed-issue year from the publisher record. OpenAlex is linked by DOI and provides author counts, affiliations, and citations. Affiliations are taken from the journal article record rather than working-paper versions, and citation counts are used only when OpenAlex identifies a valid journal article.[26] Citation counts are measured per completed year since publication through 2025 and ranked within publication decade, so comparisons are not mechanically driven by paper age. In the citation sorts of Table 4, the first four groups are disjoint, while the top-10% and top-5% rows are nested tails of the distribution, with the top 5% contained in the top 10%. Institutions are ranked by research output in the same journals and period as the replicated papers using the UT Dallas ranking for the *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* over 2000–2020.[27] A paper's institution rank is the best rank among its authors' affiliations. A top-5 indicator equals one if at least one author is affiliated with Chicago, Wharton, Harvard, Stanford, or MIT; this named list, rather than a ranking position, defines the top-school group in the sorts and the regressions. The institution sorts in Table 4 first separate the top-5-school papers and then group the remaining papers by their best-ranked affiliation into ranks up to 20, ranks 21–50, ranks 51–200, and unranked or industry affiliations. Because the first group is defined by the named schools, other highly ranked institutions, such as NYU and Columbia, appear in the ranks-up-to-20 group.

Research field and the behavioral-mechanism indicator come from a separate classification pass run without access to replication outcomes. Using a fixed set of categories and tie-break rules, an agent assigns each paper to one of seven fields based on its main research question and primary dependent variable. The behavioral indicator is separate from field and equals one when beliefs, sentiment, attention, psychological biases, disagreement, or nonstandard preferences form the central proposed mechanism or hypothesis being tested.

[26] If no acceptable citation record is available, the paper remains in the sample and receives the median citation rate for its journal and year, with an imputation flag. Only one main-sample paper is imputed.

[27] Affiliation names from OpenAlex are matched to the ranking by normalized name and school aliases. Research networks such as NBER and CEPR are treated as co-affiliations and never carry a rank.

These characteristics describe papers and authors rather than measure research quality, and the associations reported below are descriptive rather than causal.

3 Results

The results proceed in five steps. I first show that the rebuilt samples closely match the published samples, so the in-sample exercise provides a meaningful benchmark. I then show that published finance is largely reproducible in its original samples, with 75% of key results reproducing in sign and significance and a median coefficient ratio of 0.95, but that the same estimates weaken sharply in later data, retaining only about half their published magnitude. Original statistical strength is the strongest predictor of both reproduction and out-of-sample survival, while most other paper characteristics explain much less. Finally, robustness tests show that reconstruction problems affect in-sample replication where expected but do not account for the broad out-of-sample decay.

3.1 The replication sample

The rebuilt samples closely match the samples in the published papers. Table 2, Panel B, shows that the replication covers 98% of the paper's estimation period on average and the full period for the median paper. The ratio of replicated to reported sample size has a mean of 0.99 and a median of 1.00, with 10th and 90th percentiles of 0.83 and 1.15. Some agreement is imposed by the sample screens in Section 2, but most replications lie well inside those bounds. The in-sample exercise therefore typically applies the paper's specification to roughly the same observations and period rather than to a materially different sample. Each paper is one observation in the analysis, with its key estimates averaged within paper.

Of the 1,584 papers, 1,005 have at least five years of usable post-sample data. Figure 1 shows that these papers span the full publication window and are not concentrated in a few cohorts. Out-of-sample availability is around two-thirds for most publication years and falls for the most recent papers, which have less subsequent data. The median out-of-sample window is 13 years. Where sample sizes are available, the median out-of-sample sample contains 58% as many observations as the published sample. The shorter windows make loss of statistical significance mechanically more likely and motivate the precision adjustment below. The papers with usable out-of-sample tests reproduce in sample about as well as the full sample.

3.2 Reproduction in the original sample

First, published finance is largely reproducible in its original sample. Table 2, Panel A, reports a mean category match of 75%. The median paper reproduces the published sign and significance category for every key estimate. Significant reversals are rare. Only about 1% of key estimates become significant with the opposite sign. Coefficient magnitudes also come back close to those published. Among estimates that retain the published sign, the median in-sample coefficient is 0.95 of the published coefficient. Most failures therefore reflect a loss of statistical significance rather than a large decline in estimated magnitude.

The same result appears in statistical strength. Table 3, Panel A, reports a median published synthetic t-statistic of 3.59 and a median in-sample value of 3.25. Figure 2 shows the relation paper by paper. The fitted relation stays close to the 45-degree line over almost the entire range of published strength, and its local slope is generally near one. Strong published evidence tends to reproduce as strong evidence, while marginal published evidence tends to reproduce as marginal evidence. There is substantial dispersion around the relation, but little systematic shrinkage.

The asset-pricing papers provide an external benchmark for the quality of these machine-built replications. Over the range studied by Chen and Zimmermann (2022), regressing replicated on published t-statistics gives a slope of 0.99 and an R^2 of 75% for the asset-pricing subsample, compared with their slope of 0.88 and R^2 of 82% for hand-coded return predictors. The fit here is somewhat looser, but the slope is closer to one despite covering a much broader set of empirical designs and retaining failed replications. The main in-sample result is therefore not simply that many papers clear a qualitative replication threshold. The pipeline also recovers the magnitude and cross-sectional ordering of published evidence remarkably well.

3.3 Decay out of sample

Second, published estimates weaken sharply in later data. The mean category match falls from 75% in the original sample to 42% out of sample. Shorter windows account for part of this decline. If the in-sample coefficients were unchanged but evaluated at the lower precision of the out-of-sample windows, the match rate would be about 61%. Conversely, evaluating the actual out-of-sample coefficients at the precision of the in-sample replication raises their match rate only from 42% to 48%. Both calculations imply that reduced precision explains some, but not most, of the decline.

Coefficient magnitudes show the same decline without relying on statistical significance. The median out-of-sample coefficient is 0.45 of the published coefficient and 0.59 of the

pipeline's own in-sample estimate. The second comparison is especially useful because it holds the implementation fixed. The same code and data construction recover the result in the original years and produce a much smaller coefficient in later years, which makes reconstruction error an unlikely explanation for the decline. There is substantial heterogeneity across papers, including both reversals and effects that grow, but the typical published coefficient loses about half its magnitude.

Statistical strength falls by a similar amount. Among the papers with out-of-sample tests, the median published synthetic t-statistic is 3.72 and the median window-adjusted out-of-sample value is 1.84. Figure 3 makes the contrast with Figure 2 visible. In sample, replicated strength lies close to the 45-degree line. In later years, with precision held fixed, the fitted relation lies well below it over the range containing most papers. A result that was comfortably significant in its original sample is therefore only marginal on average in subsequent data.

The contrast between Figures 2 and 3 is the central result of the paper. Within each paper, the specification and data construction are fixed across the two runs. They recover the published evidence in the original period and substantially less evidence afterward. The design does not identify why. Markets and institutions may change, published estimates may reflect selection from noisy data, and investors may learn from and trade on published findings. The next results show that the amount of decay is strongly related to one observable feature of the original evidence.

3.4 Original statistical strength and out-of-sample decay

Third, original statistical strength strongly predicts both in-sample reproduction and out-of-sample survival. Table 3 sorts papers into quartiles using their published synthetic t-statistics before either replication outcome is observed. In the weakest quartile, where the typical published t-statistic is 2.5, 56% of key results reproduce in sample. The rate rises almost monotonically to 89% in the strongest quartile, where the typical published t-statistic is 9.1. In later data the gradient is steeper. The window-adjusted out-of-sample match rises from 31% in the weakest quartile to 72% in the strongest.

Part of the significance gradient is mechanical because a result with a t-statistic of 9 can weaken considerably without crossing the significance threshold. Coefficient magnitudes show that this is not the whole explanation. The two weakest quartiles retain only about one-third of their published coefficients out of sample, with ratios of 0.36 and 0.34. The strongest quartile retains 0.62. Original strength therefore predicts how much of the effect remains, not merely whether the later estimate stays statistically significant.

Figure 3 shows the relation continuously. For published t-statistics near the significance threshold, the fitted out-of-sample strength changes little with published strength and generally lies below the significance cutoff. Above roughly 4, out-of-sample strength rises much more closely with published strength. The quartile results make the economic magnitude clear. A typical paper in the weakest quartile moves from a published t-statistic of 2.5 to an adjusted out-of-sample value of 0.95. A typical paper in the strongest quartile moves from 9.1 to 5.4. Strong results decay too, but they generally remain strong.

The out-of-sample decline is therefore concentrated among results with weaker original evidence. This pattern is consistent with publication selection. Results close to the significance threshold are more likely to reflect favorable sampling draws and therefore to shrink in new data. The quartile sort does not identify this mechanism, since strong and weak published results may differ in other ways.

3.5 Other characteristics of published results

Beyond original strength, observable paper characteristics explain much less, but several patterns help distinguish why results fail to reproduce from why they decay. Table 4 first reports simple sorts by research field, citations, institution rank, and team size. Research field produces the clearest differences.

Asset-pricing papers reproduce especially well in their original samples. Their in-sample category match is 81% and their median coefficient ratio is 0.97, compared with 69% and 0.91 in corporate finance and 68% and 0.85 in intermediation. This is not because asset-pricing papers begin with stronger published evidence. Their median published synthetic t-statistic is 3.62, compared with 3.48 in corporate finance. A natural explanation is that asset pricing is easier to reconstruct. Much of the field uses common databases, standard portfolio constructions, and recurring variable definitions, leaving fewer consequential choices to infer from the paper.

The ranking changes out of sample. Asset-pricing papers retain only 0.42 of their published coefficient, compared with 0.49 in corporate finance and 0.48 in intermediation. Their window-adjusted category match is also lower than in corporate finance, 46% against 51%. Microstructure illustrates why the coefficient and significance measures are both useful. Its adjusted match is relatively high at 53%, but these papers start with much stronger published evidence, with a median t-statistic of 6.0, and still retain only 0.46 of their published magnitude. Statistical significance can therefore survive substantial economic decay when the original evidence is strong.

The field comparison also shows that decay is not mainly an asset-pricing phenomenon.

Every field loses a substantial share of its published magnitude. Even corporate finance and intermediation retain only about half, while household finance and the residual group retain 0.85. Post-publication trading is a plausible contributor to the larger decline in asset pricing because investors can act directly on many published return predictors. It cannot explain the broader result, however, because similar decay appears in settings where the published relation is not itself a trading signal. Changes in markets, institutions, and behavior remain natural explanations for the common component of the decline.

The other univariate sorts are much weaker. Replication outcomes do not rise monotonically with citations. Papers in the most highly cited 5% do somewhat better, but the broader citation groups show no ordering in either in-sample reproduction or out-of-sample retention. Institution rank shows a modest in-sample pattern. Papers with an author at a top-five school reproduce at 81%, compared with roughly 72–75% across most other groups, but this advantage disappears out of sample. Team size likewise shows no clear monotone relation, although papers with four or more authors perform worse on several measures, including an out-of-sample coefficient ratio of 0.33. These simple differences could reflect field, original statistical strength, publication cohort, or other cross-group characteristics.

Table 5 therefore puts the main characteristics in the same regression. Original statistical strength remains the dominant predictor. The coefficient on log published $|t|$ is large and precisely estimated for all four outcomes. Doubling the published t-statistic is associated with about a 10 percentage-point higher in-sample category match and a 17-point higher window-adjusted out-of-sample match. It is also associated with roughly 0.05 more of the published coefficient in sample and 0.12 more out of sample. The strength gradient from the previous subsection therefore does not arise because strong results come disproportionately from particular fields, cohorts, or institutions.

Controlling for the other paper characteristics leaves the field patterns largely intact. Asset pricing retains a 12 percentage-point advantage in in-sample reproduction and a 0.08 higher coefficient ratio. Out of sample, its adjusted significance rate is no different from other fields, but its coefficient ratio is 0.11 lower, although this difference is not statistically significant. Asset-pricing results therefore remain easier to reconstruct but lose more of their economic magnitude in later data.

Papers built around a behavioral mechanism show a strikingly similar pattern. They are no different in in-sample reproduction and have essentially the same adjusted out-of-sample category match, but retain 0.14 less of their published coefficient. Because the regression includes the asset-pricing indicator separately, this association is not simply an asset-pricing effect. Both patterns are consistent with published findings changing subsequent behavior. Investors may trade on return predictors, while beliefs, attention, or other behavior may re-

spond once a relation becomes known. They are also consistent with a different explanation, namely that asset-pricing and behavioral relations depend more strongly on the economic environment in which they were originally estimated. The data cannot separate these mechanisms. What they do show is that the additional decay appears in coefficient magnitudes rather than in whether an estimate crosses a significance threshold.

Publication year produces almost the opposite pattern. A paper published a decade later has a 4.3-point lower in-sample category match and a 0.08 lower in-sample coefficient ratio, but publication year predicts neither adjusted out-of-sample outcome. Later papers are therefore somewhat harder to reconstruct, but their estimates do not decay faster once reconstructed. This distinction is consistent with the empirical designs and data requirements of finance becoming more difficult to reproduce over time without implying that newer economic relations are less durable. The finding also fits the evidence below that substitutions of core data sources reduce in-sample reproduction without materially changing out-of-sample coefficient decay.

Author and institution characteristics provide weaker signals. A top-five-school affiliation is associated with a 4.6-point higher in-sample match and a 0.10 higher in-sample coefficient ratio, but with no out-of-sample advantage. This could reflect better documentation, more standardized data, or other features that make the original analysis easier to reconstruct, but it provides no evidence that the underlying effects last longer. The star-author indicator is associated with somewhat higher category matches in and out of sample, at the 10% level in sample and the 5% level out of sample, and predicts neither coefficient ratio. I therefore do not interpret author prominence as a reliable signal of economic durability.

Finally, even these regressions leave most of the variation unexplained. The covariates account for 12% of the variation in the in-sample category match and 15% in the adjusted out-of-sample match, but only 3% and 2% in coefficient magnitudes. Published statistical strength is a useful predictor in expectation, and field identifies some additional differences in how effects decay, but paper characteristics provide little ability to predict the fate of an individual result. In particular, prestige and attention provide little information about how much of a published coefficient will survive in later data.

3.6 Robustness to the replication pipeline

The final results ask whether features of the replication process itself could generate the main findings. Table A2, Panel A, splits papers by the model family used in the blind re-audit. The levels are nearly identical. GPT-audited papers have an in-sample match of 75.5% and an adjusted out-of-sample match of 47.6%, compared with 73.7% and 49.1% for Claude-

audited papers. Their median out-of-sample coefficient ratios are 0.45 and 0.47. Panel B similarly shows little relation between how explicitly a paper documents its methodology and whether it reproduces. Papers above and below the median documentation score differ little in sample and only modestly out of sample.

Core data substitution provides a more informative check because it identifies a concrete source of reconstruction error. Panel C shows that data substitution makes the original result materially harder to recover: the in-sample match falls from 77.6% with the original core data to 69.7% with a substitute or proxy, and the coefficient ratios are 0.96 and 0.90. Yet the out-of-sample coefficient ratios are nearly identical at 0.45 and 0.44, and the adjusted category match differs far less than in sample, at 50.0% and 44.2%. Reconstruction error therefore affects the in-sample benchmark but cannot account for the broad decline in later data. Published finance is largely reproducible in its original samples, despite some noise from imperfect reconstruction, and its estimated effects are substantially less durable outside those samples.

4 How Much of the Decay Is Publication Bias?

Publication bias explains only about a fifth of the out-of-sample decay. The strength gradient in Section 3.4 is consistent with publication selection, but the sort cannot say how much decay selection should produce. I therefore fit two standard models of publication selection to the published t-statistics and use them to predict how much each estimate should shrink in fresh data. The models predict that the median paper should retain about 0.90 of its published coefficient. It actually retains 0.45. Selection on reported significance accounts for roughly 18% of the observed decay, and no variant of either model explains more than 22%.

Selection predicts decay because published results are lucky on average. Journals rarely publish insignificant results. An estimate equals the true effect plus sampling noise, so the results that cleared the significance bar carry favorable noise on average. In new data the luck does not repeat, and the estimate falls back toward the true effect. The published t-statistic says how large a fallback to expect. A result that barely cleared the bar often needed luck and should shrink noticeably. A result with a t-statistic of 9 did not need luck and should barely move. The remaining ingredient is the distribution of true effects, and both models recover it from the shape of the published t-statistic distribution. If most true effects were weak, published t-statistics would pile up just above the threshold. A long right tail can only come from genuinely strong effects.

The two models extract different information from the same data. The empirical-Bayes model, in the spirit of Chen and Zimmermann (2020), fits a distribution of true effects to

the published t-statistics above the threshold and shrinks each estimate toward what that distribution makes typical.[28] The selection model, following Andrews and Kasy (2019), estimates the selection process itself from the published estimates that are insignificant, whose scarcity reveals how strongly significance is favored. It finds that an insignificant estimate is about one-fifth as likely to appear among a paper's key results as a significant one.[29]

Applying either model to all of empirical finance requires three modifications, because both were built for collections of return predictors measured in common units. Published coefficients here are in different units across papers, so both models work with t-statistics, which standardize each estimate by its own standard error. All estimates are oriented so the published effect is positive, and the models account for this.[30] And because designs such as persistence regressions naturally produce a few t-statistics in the hundreds, the models are fitted to t-statistics up to 20, with wider and narrower windows as robustness checks. Both models see only the published in-sample t-statistics, so the comparison with out-of-sample outcomes is a genuine prediction test.

Both models predict far too little decay, everywhere. In Figure 4, predicted retention lies between 0.87 and 0.96 across the whole range of published strength, while actual retention starts near 0.31 at the significance threshold and reaches only 0.65 around a t-statistic of 10. The gap is widest exactly where selection should matter most. In the weakest quartile of Table 6, the models predict retention of about 0.90 and the data show 0.35. Even in the strongest quartile the prediction is 0.95 against an actual 0.62. Overall, the two models explain 17.9% and 18.5% of the median decay, with bootstrap confidence intervals of roughly 16% to 22%, and varying the fitting window, the significance cutoff, and the weighting moves the explained share only between 8% and 22%.[31] Weak results do decay more than strong

[28] In the empirical-Bayes model, true effects in t-statistic units follow a Student $t(4)$ distribution with an estimated scale of 2.8, and the likelihood conditions on an estimate being significant and inside the fitting window. In the selection model, true effects follow a Student t distribution with an estimated location of 1.9, a scale of 1.4, and a heavy tail, and the publication probability drops by a constant estimated factor below the significance threshold. Both fits use at most four key estimates per paper and weight papers equally, matching the paper-level aggregation used throughout. Confidence intervals come from 200 bootstrap draws that resample papers and re-estimate the models.

[29] The estimate is 0.18 with a bootstrap interval of 0.16 to 0.22, and it changes little when only one estimate per paper is used. It describes the fitted frequency of insignificant estimates among a published paper's key results rather than a directly observed publication decision. The sample contains only published papers, and standardizing to t-statistics removes the variation in standard errors that Andrews and Kasy (2019) use for identification, so this parameter rests on the model's distributional assumptions.

[30] Because the orientation uses the sign of the published estimate itself, the models observe absolute t-statistics, and the likelihood and predictions are computed for absolute values accordingly. An audit of twelve estimates whose recorded t-statistic sign contradicted the published coefficient found eleven t-statistics transcribed without their sign and one sign inconsistency printed in the original paper; all twelve are corrected here.

[31] The range covers fitting windows of 10 and 20 and an unbounded fit, a 1.645 cutoff that accommodates

ones, as the selection story predicts, but several times more than selection can produce.

The small explained share is not because the method cannot detect selection. Applied to data where the literature knows the answer, the 207 published return predictors of Chen and Zimmermann (2022), the same machinery estimates that bias-corrected returns are 91% of published returns, matching the 8% to 17% shrinkage that Chen and Zimmermann (2020) report and similar estimates elsewhere (Harvey et al., 2016; Jensen et al., 2023). Publication bias in finance is genuinely modest wherever it has been measured. The new fact is on the other side of the comparison: realized decay is several times larger than the bias that any of these models, here or in the prior literature, can find in the published record.

These shares bound the role of selective reporting from below, not from above. The models see selection only on the final reported t-statistic. If authors search across specifications and report the strongest, the published estimate is inflated in ways the models cannot detect. The unexplained residual is correspondingly an upper bound on learning and structural change rather than a clean estimate of either. What the models do establish is one-sided: everything the publication process does through the reported statistic is far too small to explain most of the decay.

The decay documented in Section 3.3 is therefore not an artifact of which results get published. Under any fitted version of the selection story, published effects genuinely weaken in later data, consistent with markets and institutions changing and with investors learning from published findings (McLean and Pontiff, 2016). Existing bias corrections cannot close the gap either. The corrections proposed in the literature shrink published effects by roughly 10% to 15% (Harvey et al., 2016; Chen and Zimmermann, 2020), about the right size for the selection they target but only a small part of the gap to later data. A reader who wants the best estimate of an effect's magnitude today should worry less about publication bias and more about how much the effect has decayed since the sample ended.

5 Why Do Published Results Decay?

Three explanations can account for the decay documented in Section 3.3, and they differ in when the effect should fall. The first is data snooping. Researchers try many samples, specifications, and variable definitions, and the ones that reach the published table are those that worked best in the data at hand. The published estimate then reflects some luck that is specific to the original sample, so the effect should fall as soon as that sample ends, whether

one-sided tests, unweighted estimation, one estimate per paper, and a normal instead of a Student-t effect distribution. Extrapolation beyond the fitting window is immaterial: only 5% of matched estimates lie above it, and restricting the comparison itself to t-statistics up to 20 changes the explained shares by less than 0.2 percentage point.

or not anyone has read the paper. The second is learning. If investors, firms, or regulators act on published findings, the effect should fall further once the paper appears. The third is change in markets and institutions unrelated to the paper, which should erode the effect gradually, with no special role for either the sample end or the publication date. Section 4 measured the narrow form of the first explanation, selection on the reported t-statistic, and found that it accounts for about a fifth of the decay. This section uses the timing of the decay to weigh all three explanations, following the design that McLean and Pontiff (2016) use for return predictors. The size of the drop at the boundary between the original sample and the later years is a direct estimate of how much data snooping inflates published estimates.

For each paper with an out-of-sample test, the pipeline re-estimates the key coefficients one calendar year at a time with the specification unchanged. The same variables, filters, fixed effects, and standard errors are used in every year, and only the sample period varies. Each yearly coefficient is divided by the published coefficient, so that one reproduces the paper and zero means no effect, and the ratios are winsorized and averaged across a paper's key coefficients within each year. Every paper-year then falls into one of three stages, the original sample, the years after the sample ends but before the journal publication year, and the years from publication onward. The main specification regresses the yearly ratio on indicators for the two later stages with paper fixed effects, so that each paper is compared with its own in-sample level, and weights each paper-stage equally so that long windows do not dominate. Standard errors are clustered by paper and by calendar year.[32] The design applies to 836 of the 1,005 papers with an out-of-sample test. The remaining papers cannot be estimated year by year, most often because the specification uses annual data with firm fixed effects or a long-horizon or structural estimator. These excluded papers decay more in the full-sample test, so the estimates below are conservative about the total decline.

Figure 5 shows the central result. Within the original sample the yearly estimates average about 0.85 of the published coefficient, and the level does not trend over the ten years before the sample ends. In the first out-of-sample year the ratio drops to 0.65, and from the following year onward it settles near 0.57, again with no trend over the next ten years. The decay is neither a gradual erosion nor something that waits for publication. It is a step of about a quarter of the published coefficient, located exactly where the data the authors could see give way to data they could not. That location is the signature of data snooping, and the size of the step measures it. Placebo cutoffs placed 5, 8, or 12 years inside the original

[32] Yearly estimates whose standard error exceeds ten times the published magnitude are dropped, 3% of coefficient-years, because a single year cannot identify some specifications. Because the yearly ratios have heavy tails, Table 7, Panel B, repeats the regression as a pooled median regression, and Table A3 in the Appendix adds calendar-year fixed effects and a fixed-effects median regression. All versions agree on the pattern.

sample, or five years into the later period, produce no comparable break (Table A4 in the Appendix).[33]

Table 7 quantifies the step and the publication effect. Once the original sample ends, the mean ratio falls by 0.28 relative to the paper's own in-sample level, and the estimate is precise. Publication changes nothing further. The additional change from the publication year onward is +0.03 with a standard error of 0.07, so the total post-publication decline of 0.26 is entirely the drop that occurred when the sample ended. The pooled median regression in Panel B gives the same numbers, a drop of 0.28 at the sample end and a publication increment of −0.02. I therefore read 0.28 as the share of the typical published coefficient that reflects data snooping rather than a durable effect. The result speaks against learning as the main channel. If investors traded away published anomalies, or if firms and regulators responded to published findings, the effect should have weakened further after publication, and it did not. Post-publication trading is the channel most often proposed for return predictors, and even there the effect has largely gone before the paper appears.

Data snooping is much larger in asset pricing than elsewhere. The step at the sample end is 0.41 for asset-pricing papers and 0.12 for papers in all other fields, and the publication increment is close to zero in both groups. Figure A2 in the Appendix shows the two paths. This ordering is what the mechanism predicts. Returns are noisy, so a given amount of searching across specifications and samples produces a larger favorable draw in return data than in the more persistent accounting, contracting, or institutional variables of other fields. The larger decay of asset-pricing results documented in Section 3.5 is therefore not mainly evidence that markets learn to exploit published predictors. It is mainly evidence that noisy data make it easier to find results that do not hold up.

The strength of the original evidence points the same way. Sorting the papers by the mean published t-statistic of their key coefficients, the step at the sample end is 0.25, 0.49, and 0.27 in the three weaker quartiles and 0.13 in the strongest, where the typical published t-statistic is 10. The publication increment is close to zero in every quartile. Results that clear the significance bar by a wide margin have little room to have been helped by a favorable sample, and they lose the least when the sample ends. Nothing about the publication date should depend on the original t-statistic, and nothing does.

Taken together, Sections 4 and 5 point to data snooping as the main cause of decay, but

[33]The placebo test fits straight lines to each paper's ratios over the five years on either side of a cutoff. At the true boundary the median within-paper difference between the two windows is −0.11 with a t-statistic near four, while every placebo difference is within 0.03 of zero. The jump at exactly the first out-of-sample year is smaller, −0.07, because sample ends fall in mid-year and the first out-of-sample year straddles the boundary for many papers, which is why Figure 5 draws that year as an isolated point. The median version of the figure shows the same shape at a lower level, 0.75 before the boundary and 0.46 after it.

data snooping of a broader kind than publication bias on the final t-statistic. The selection models of Section 4 see only the reported statistic and account for about a fifth of the decline. The step at the sample end is about a quarter of the published coefficient, so most of the snooping operates earlier in the research process, in the choice of samples, specifications, and results that reach the table, where the reported t-statistic cannot reveal it. Slow change in markets and institutions contributes little, since the ratio does not trend on either side of the boundary. The test has limits. It covers the 836 papers whose designs can be estimated one year at a time, it dates publication by the printed-issue year, and it cannot observe the working-paper circulation that precedes publication and that could move markets before the journal date. With those caveats, the evidence places the cause of decay in the original sample rather than in the years after publication, for finance as a whole as McLean and Pontiff (2016) found for return predictors.

6 Conclusion

I replicate the central empirical results of published finance first in their original sample years and then in subsequent data. The in-sample evidence is reassuring. Three quarters of published results reproduce in sign and significance, significant reversals are rare, and the median replicated coefficient is 0.95 of the published one. The out-of-sample evidence is different. The raw category-match rate falls to 42%, or 48% after adjusting for shorter post-sample windows, and the median coefficient falls to 0.45 of its published magnitude. Results with stronger original evidence reproduce better and retain substantially more of their magnitude in later data, while most other observable paper characteristics explain relatively little.

These results separate reproducibility from durability. Recovering a published estimate in its original sample does not imply that the estimate remains equally large in later years. The decline is not confined to asset pricing. Similar magnitude decay appears in corporate finance, intermediation, and household finance, which makes post-publication trading an incomplete explanation for the broader pattern. The design cannot distinguish among publication selection, changes in markets and institutions, and learning after publication. It does show that decay is greatest among results with weaker original evidence and that substantial heterogeneity remains even after conditioning on that strength.

The exercise also shows what large-scale replication with autonomous LLM agents can and cannot do. Individual replications remain imperfect, so the model's output cannot simply be accepted as correct. Recording discretionary choices, preserving code and intermediate results, independently auditing the implementation, and computing replication measures

mechanically makes those errors observable and correctable. This approach complements rather than replaces author replication packages and careful human review, but it makes systematic checking of much larger bodies of empirical work feasible.

Published finance is therefore more reproducible than durable. Most central results can be rebuilt in their original samples, even without the authors' code, but their estimated magnitudes often weaken substantially in later data. That difference matters whenever published estimates are used long after the samples that produced them have ended.

References

Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. Predicting the replicability of social science lab experiments. *PLoS ONE*, 14(12):e0225826, 2019.

Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794, 2019.

Abel Brodeur, David Valenta, Alexandru Marcoci, Juan Pablo Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, Lachlan Deer, Tom Stafford, et al. Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science. I4R Discussion Paper 195, Institute for Replication, 2025.

Abel Brodeur, Derek Mikola, Nikolai Cook, et al. Reproducibility and robustness of economics and political science research. *Nature*, 652:151–156, 2026.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2:637–644, 2018.

Andrew C. Chang and Phillip Li. Is economics research replicable? sixty published papers from thirteen journals say "often not". *Critical Finance Review*, 11(1):185–206, 2022.

Andrew Y. Chen. The limits of p-hacking: Some thought experiments. *Journal of Finance*, 76(5):2447–2480, 2021.

Andrew Y. Chen and Tom Zimmermann. Publication bias and the cross-section of stock returns. *Review of Asset Pricing Studies*, 10(2):249–289, 2020.

Andrew Y. Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264, 2022.

Andrew Y. Chen, Alejandro Lopez-Lira, and Tom Zimmermann. Does peer-reviewed research help predict stock returns? Working paper 4308069, SSRN, 2025.

Tarun Chordia, Avanidhar Subrahmanyam, and Qing Tong. Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, 58(1):41–58, 2014.

Garret Christensen and Edward Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980, 2018.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review*, 76(4):587–603, 1986.

Jeremiah Green, John R. M. Hand, and X. Frank Zhang. The characteristics that provide independent information about average u.s. monthly stock returns. *Review of Financial Studies*, 30(12):4389–4436, 2017.

Campbell R. Harvey. Presidential address: The scientific outlook in financial economics. *Journal of Finance*, 72(4):1399–1440, 2017.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Sylvérie Herbert, Hautahi Kingi, Flavio Stanchi, and Lars Vilhuber. Reproduce to validate: A comprehensive study on the reproducibility of economics research. *Canadian Journal of Economics*, 57(3):961–988, 2024.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020.

Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, and Daniel Kang. REPRO-Bench: Can agentic AI systems assess the reproducibility of social science research? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23616–23626, 2025.

Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and

Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021.

Heiko Jacobs and Sebastian Müller. Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics*, 135(1):213–230, 2020.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518, 2023.

Richard A. Klein et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results. arXiv preprint 2604.21965, arXiv, 2026.

Anton Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

Anton Korinek. Generative ai for economic research: Llms learn to collaborate and reason. NBER Working Paper 33198, National Bureau of Economic Research, 2024.

Juhani T. Linnainmaa and Michael R. Roberts. The history of the cross-section of stock returns. *Review of Financial Studies*, 31(7):2606–2649, 2018.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework. *Annual Review of Economics*, 18:283–316, 2026.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.

Albert J. Menkveld, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüß, Michael Razen, Utz Weitzel, et al. Nonstandard errors. *Journal of Finance*, 79(3):2339–2390, 2024.

Brian A. Nosek, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, Julia M. Rohrer, Felipe Romero, Anne M. Scheel, Laura D. Scherer, Felix D. Schönbrodt, and Simine Vazire. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73:719–748, 2022.

Robert Novy-Marx and Mihail Velikov. Artificial intelligence–powered (finance) scholarship. *Journal of Economic Literature*, 64(1):5–37, 2026.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Christophe Pérignon, Olivier Akmansoy, Christophe Hurlin, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Albert J. Menkveld, Michael Razen, and Utz Weitzel. Computational reproducibility in finance: Evidence from 1,000 tests. *Review of Financial Studies*, 37(11):3558–3593, 2024.

G. William Schwert. Anomalies and market efficiency. In George M. Constantinides, Milton Harris, and René M. Stulz, editors, *Handbook of the Economics of Finance*, volume 1, pages 939–974. Elsevier, 2003.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.

Table 1. From the published universe to the replication samples. The table starts with every research article the three journals published in 2000–2020, taken from the publishers' own journal metadata (the 5,455 articles behind Figure A1 less 25 non-research items: biographical essays, presidential addresses, review articles, errata, and comments), and classifies each article by the outcome of its replication attempt. Papers are counted once (by DOI). *Theory* covers papers with no replicable empirical table, including simulation-only and survey papers. *Confidential, hand-collected, or author-owned data* covers samples the authors assembled or obtained under restricted access (hand-collected records, author surveys, confidential firm or supervisory data, restricted government microdata). *Commercial data not available to us* covers vendor products outside our subscriptions or no longer sold. *Other failures* pools corrupt source PDFs, estimators outside the pipeline's scope, coverage gaps in data we do hold, public data behind access barriers, and attempts that ran but did not reproduce the paper's sample or estimates. The three *Replicated, but* rows are replications that ran end to end and were excluded by the sample rules of the analysis: the rebuilt sample covers less than half of the paper's period or is outside 0.5–1.5× the paper's size; the design has no in-sample replication or the paper-to-replication comparison fails the integrity audit; or the original claim is a null result, which the closeness measures cannot score. Replications whose blind re-audit judged the required data unavailable (154 papers) are classified by that audit's stated reason into the data rows above. The 52 research articles that were never attempted (coverage gaps in the bulk downloads) are allocated across the four failure rows in proportion to their size; 29 first-pass replications that never entered the validated pipeline are counted under other failures. The last row counts final-sample papers with a paired out-of-sample estimate, the sample behind every out-of-sample statistic in the paper.

	Papers	Share (%)
Research articles published in JF, JFE, and RFS, 2000–2020	5,430	100.0
Theory, or no empirical tables	999	18.4
Requires confidential, hand-collected, or author-owned data	859	15.8
Requires commercial data not available to us	1,002	18.5
Other failures	603	11.1
Replicated, but rebuilt sample does not match the paper	260	4.8
Replicated, but excluded by design or audit checks	98	1.8
Replicated, but the paper's central claim is a null result	25	0.5
Final in-sample replication sample	1,584	29.2
With an out-of-sample test	1,005	18.5

Table 2. Summary statistics for the main sample. The table describes the closeness measures (Panel A) and the paper and replication characteristics (Panel B) across main-sample papers. Each observation is a paper; coefficient-level quantities are first averaged within paper. Key estimates are the paper results designated during the closeness-extraction pass as carrying the target table's central empirical claim. IS denotes the in-sample replication, which rebuilds the paper's own sample period. OOS denotes the out-of-sample replication, which re-runs the design on a machine-readable period that starts strictly after the paper's sample and lasts at least five years. A category places an estimate into negative and significant, insignificant, or positive and significant at the 5% level. Original–IS and Original–OOS matches are the shares of key estimates whose replication category equals the published category. The window-adjusted OOS match classifies each OOS coefficient at the in-sample implied precision $|\hat{\beta}_{IS}/t_{IS}|$, removing the mechanical significance loss from the shorter post-sample window. The IS/original coeff. ratio is the within-paper geometric mean for same-sign non-null estimates. OOS/IS and OOS/original coeff. ratios are oriented to make the published sign positive; one means unchanged magnitude and a negative value means sign reversal. Panel B reports publication year; the fraction of the paper's estimation window covered; replication and OOS sample sizes, each divided by the paper's reported size; the number of key estimates; OpenAlex citations per completed calendar year through 2025, with missing rates replaced by their journal-year medians; and the OpenAlex author count. In Panel A, bootstrap standard errors of the mean and the median (100,000 resamples of papers with replacement) are in parentheses; a standard error of zero for a median reflects a mass point at the median value (for example, more than half of the papers have an Original–IS category match of exactly one), not a lack of resampling.

Panel A: closeness measures

Variable	Num. obs.	Mean	Median	Std. dev.	P10	P90
Original–IS category match	1,343	0.75	1.00	0.35	0.00	1.00
		(0.01)	(0.00)			
IS/original coeff. ratio	1,541	0.95	0.95	0.77	0.37	1.28
		(0.02)	(0.01)			
Original–OOS category match	1,005	0.42	0.33	0.43	0.00	1.00
		(0.01)	(0.04)			
Original–OOS match, adjusted	1,005	0.48	0.50	0.43	0.00	1.00
		(0.01)	(0.01)			
OOS/IS coeff. ratio	980	0.80	0.59	1.79	-0.44	1.98
		(0.06)	(0.03)			
OOS/original coeff. ratio	1,005	0.63	0.45	1.21	-0.32	1.65
		(0.04)	(0.02)			

Panel B: paper and replication characteristics

Variable	Num. obs.	Mean	Median	Std. dev.	P10	P90
Publication year	1,584	2011	2011	5.58	2003	2019
Paper-period coverage	1,524	0.98	1.00	0.07	0.98	1.00
Replication/paper sample size	1,512	0.99	1.00	0.14	0.83	1.15
OOS/paper sample size	474	0.79	0.58	1.52	0.21	1.28
Key estimates per paper	1,580	3.35	3.00	2.56	1.00	5.00
Annual citations	1,584	25.41	14.77	34.47	3.16	56.26
Number of authors	1,583	2.38	2.00	0.90	1.00	3.00

Table 3. Original-result strength and subsequent replication evidence. The table shows how in-sample recovery (Panel A) and out-of-sample retention (Panel B) vary with the strength of the original published result. A paper's synthetic t is the median t-statistic across its originally significant directional key estimates. The original measure uses absolute published t-statistics, while IS and OOS values are oriented to the published sign so that reversals enter negatively. Quartiles are formed ex ante from the original synthetic t and retained across panels. Panel B uses window-adjusted OOS t-statistics and category matches, evaluating each OOS coefficient at the in-sample implied precision $|\hat{\beta}_{IS}/t_{IS}|$ to remove the mechanical loss of significance from shorter post-sample windows. Coefficient ratios do not depend on window length. Category-match columns report the mean paper-level share of key estimates with the same sign-and-5%-significance category as published. Coefficient-ratio columns report the median across papers of paper-mean replication/original ratios, oriented so the published sign is positive. Coefficient-level t-statistics and ratios are winsorized at the pooled 1st and 99th percentiles before aggregation. Because stronger published results have more distance from the significance threshold, the coefficient-ratio columns show whether magnitude retention also varies with original strength.

Panel A: original paper to in-sample replication

Original-strength quartile	Papers	Original synthetic t-stat	IS synthetic t-stat	Original–IS category match	IS/original coeff. ratio
Overall	1,326	3.59	3.25	74.04%	0.93
				(1.04)	(0.01)
Q1 (weakest)	348	2.49	2.10	56.20%	0.89
				(2.33)	(0.02)
Q2	315	3.22	2.87	70.97%	0.91
				(2.13)	(0.03)
Q3	333	4.27	3.82	80.76%	0.93
				(1.83)	(0.02)
Q4 (strongest)	330	9.06	7.99	88.99%	0.98
				(1.47)	(0.02)

Panel B: original paper to out-of-sample replication (window-adjusted)

Original-strength quartile	Papers	Original synthetic t-stat	Adj. OOS synthetic t-stat	Adj. Original–OOS category match	OOS/original coeff. ratio
Overall	1,005	3.72	1.84	47.94%	0.45
				(1.36)	(0.02)
Q1 (weakest)	239	2.52	0.95	31.09%	0.36
				(2.66)	(0.06)
Q2	231	3.22	1.08	35.55%	0.34
				(2.65)	(0.06)
Q3	262	4.33	1.94	49.57%	0.42
				(2.61)	(0.05)
Q4 (strongest)	273	9.05	5.39	71.62%	0.62
				(2.23)	(0.05)

Table 4. Replication outcomes by field, citations, institution rank, and team size. Each panel sorts main-sample papers on one characteristic and reports the same three column groups: synthetic t-stats, category-match rates, and coefficient ratios. The synthetic t-stat columns are within-group medians of the paper-level median t-statistic over originally significant directional key estimates: the absolute published value (Original), the in-sample replication value oriented to the published sign (IS), and the window-adjusted out-of-sample value (Adj. OOS), which classifies each OOS coefficient at the in-sample implied precision $|\hat{\beta}_{IS}/t_{IS}|$ so the shorter post-sample window does not mechanically lower it. Category-match columns are group means of the paper-level share of key estimates whose replication sign-and-5%-significance category equals the published category; the adjusted OOS match applies the same in-sample-precision classification as the adjusted t-stat column. Coeff.-ratio columns are group medians of paper-mean replication/original ratios, oriented so the published sign is positive. Research fields come from an LLM classification of each paper's abstract; the household-finance and other categories are combined into one residual group, so the panel covers the full sample. Citation percentiles rank OpenAlex citations per year within each publication decade; the top-10% and top-5% rows are subsets of the 75–100% range. Top 5 schools are Chicago, Wharton, Harvard, Stanford, and MIT (any author affiliation); remaining rows use the best-ranked matched affiliation across the author team, and papers with no ranked affiliation are grouped as unranked/industry/other. Author counts come from OpenAlex.

Panel A: research field

		Synthetic t-stat			Category match			Coeff. ratio	
	Papers	Original	IS	Adj. OOS	IS	OOS	Adj. OOS	IS	OOS
Asset pricing	764	3.62	3.30	1.68	81.0%	35.8%	46.2%	0.97	0.42
Corporate finance	479	3.48	2.95	2.26	69.2%	48.5%	51.0%	0.91	0.49
Intermediation	94	3.19	2.84	2.02	68.2%	41.3%	49.7%	0.85	0.48
Investments	93	3.48	3.38	1.38	71.5%	42.8%	38.1%	0.97	0.45
Microstructure	67	6.00	4.96	2.36	69.9%	58.1%	52.7%	0.90	0.46
Household fin. / other	87	3.71	3.58	4.75	76.1%	56.2%	61.1%	1.00	0.85

Panel B: annual citations (percentile within publication decade)

		Synthetic t-stat			Category match			Coeff. ratio	
	Papers	Original	IS	Adj. OOS	IS	OOS	Adj. OOS	IS	OOS
0–25%	394	3.48	3.33	2.06	75.8%	40.8%	50.9%	0.95	0.51
25–50%	396	3.48	2.97	1.52	72.1%	43.4%	45.2%	0.93	0.44
50–75%	396	3.65	3.50	2.01	75.9%	41.7%	49.3%	0.95	0.45
75–90%	239	3.75	3.28	1.73	75.7%	39.4%	46.8%	0.97	0.46
Top 10%	159	3.60	3.05	2.07	77.4%	42.9%	46.3%	0.95	0.41
Top 5%	81	3.67	3.25	2.88	82.7%	51.5%	54.9%	0.98	0.60

Table 4 (continued). Replication outcomes by field, citations, institution rank, and team size.

Panel C: best institution rank across the author team

		Synthetic t-stat			Category match			Coeff. ratio	
	Papers	Original	IS	Adj. OOS	IS	OOS	Adj. OOS	IS	OOS
Top 5 schools	230	3.68	3.50	2.19	81.1%	40.8%	49.9%	0.98	0.45
Ranks 6–20	410	3.59	3.27	1.67	74.3%	41.1%	47.9%	0.97	0.48
Ranks 21–50	360	3.48	3.50	1.81	75.3%	44.0%	47.4%	0.94	0.43
Ranks 51–200	302	3.48	3.06	1.70	72.4%	36.5%	43.7%	0.93	0.41
Unranked / industry / other	282	3.61	3.04	2.05	74.0%	46.3%	52.1%	0.94	0.58

Panel D: number of authors

		Synthetic t-stat			Category match			Coeff. ratio	
	Papers	Original	IS	Adj. OOS	IS	OOS	Adj. OOS	IS	OOS
1 author	251	3.74	3.15	1.86	74.9%	41.1%	49.5%	0.95	0.51
2 authors	641	3.74	3.41	1.83	78.0%	40.7%	48.0%	0.96	0.45
3 authors	564	3.47	3.20	1.92	72.9%	45.0%	48.0%	0.94	0.46
4+ authors	128	3.48	2.89	1.31	70.8%	32.9%	44.5%	0.93	0.33

Table 5. Conditional predictors of replication outcomes. Each column is a paper-level OLS regression of one replication outcome on the displayed predictors jointly. Column (1) is the share of the paper's key estimates whose in-sample replication reproduces the published sign-and-5%-significance category, in percentage points; column (2) is the within-paper mean in-sample/original coeff. ratio (oriented so the published sign is positive; coefficient-level ratios winsorized at the pooled 1st/99th percentiles); column (3) is the window-adjusted OOS category match in percentage points, which classifies each OOS coefficient at the in-sample implied precision $|\hat{\beta}_{IS}/t_{IS}|$ so the shorter post-sample window does not mechanically lower it; column (4) is the within-paper mean OOS/original coeff. ratio. Predictors: log mean published $|t$-stat$|$ is the log of the mean absolute published t-statistic across the paper's originally significant directional key estimates; asset pricing and behavioral mechanism come from the field classification; publication year is centered at 2010 and scaled by ten years, so its coefficient is the change per decade of publication date; top-5 school equals one when at least one author is affiliated with Chicago, Wharton, Harvard, Stanford, or MIT; star author equals one when at least one author's OpenAlex h-index or career citations (excluding the paper itself) place them in the top 5% of main-sample authors. Match outcomes are in percentage points. Heteroskedasticity-robust (HC1) t-statistics are in parentheses. The estimates are conditional associations, not causal effects. *, **, ***: significant at 10%, 5%, 1%.

	(1) IS category match	(2) IS/orig. coeff. ratio	(3) Adj. OOS category match	(4) OOS/orig. coeff. ratio
log mean published $\|t$-stat$\|$	14.37***	0.07***	24.61***	0.18***
	(10.88)	(3.75)	(13.42)	(3.69)
Asset pricing	11.62***	0.08**	-0.14	-0.11
	(6.24)	(2.49)	(-0.05)	(-1.44)
Behavioral mechanism	3.27	-0.02	1.10	-0.14*
	(1.37)	(-0.57)	(0.33)	(-1.74)
Publication year, (year − 2010)/10	-4.29***	-0.08***	-2.00	-0.02
	(-2.67)	(-3.03)	(-0.86)	(-0.26)
Top-5 school	4.61*	0.10***	0.17	0.04
	(1.94)	(2.63)	(0.05)	(0.39)
Star author	7.09*	0.09	11.85**	0.13
	(1.85)	(1.36)	(1.97)	(0.77)
Observations	1,326	1,328	1,005	1,005
R^2	0.116	0.027	0.153	0.017

Table 6. Out-of-sample decay predicted by publication selection versus actual decay. The table compares the out-of-sample coefficient decay that publication selection alone would produce with the decay actually observed. Two selection models are fitted only to the published in-sample t-statistics of the key estimates and never see out-of-sample data. The empirical-Bayes model, adapted from Chen and Zimmermann (2020), infers the distribution of true effects from the shape of the published t-statistic distribution above the significance threshold. The selection model, adapted from Andrews and Kasy (2019), additionally estimates how much less likely an insignificant estimate is to appear in a published table. Because coefficients are in different units across papers, both models are fitted to t-statistics (each estimate standardized by its own standard error) restricted to $|t| \leq 20$, with at most four estimates per paper and papers weighted equally. Each model converts a published t-statistic into a predicted ratio of the true to the published coefficient; predicted and actual ratios are averaged within paper and the table reports medians across papers, overall and by quartile of published synthetic t-statistic. The share of decay explained divides the predicted median decline by the actual median decline. Paper-bootstrap 95% confidence intervals that re-estimate the models on each draw: actual overall ratio [0.42, 0.52]; predicted overall ratio [0.89, 0.91] (empirical Bayes) and [0.88, 0.91] (selection model); overall share explained [16%, 20%] and [16%, 22%]. Varying the fitting window, significance cutoff, capping, and weighting moves the explained share between 8% and 22% (Section 4).

Original-strength quartile	Papers	Actual OOS/original coeff. ratio	Predicted: empirical Bayes	Predicted: selection model	Decay explained: empirical Bayes	Decay explained: selection model
Overall	1,005	0.45	0.90	0.90	17.9%	18.5%
Q1 (weakest)	238	0.35	0.89	0.90	17.3%	14.7%
Q2	232	0.34	0.90	0.87	15.8%	19.3%
Q3	262	0.43	0.91	0.88	15.9%	20.2%
Q4 (strongest)	273	0.62	0.95	0.96	13.5%	11.8%

Table 7. Decay of replicated coefficients after the original sample ends and after publication. Each paper's key coefficients are re-estimated calendar year by calendar year with the specification unchanged; the yearly estimate is divided by the published coefficient (oriented so the published value is positive, so 1 reproduces the paper, 0 is no effect, and a negative value is a sign flip), winsorized at the pooled 1st/99th percentiles, and averaged across the paper's coefficients within each year. The unit is a paper-year; every paper-stage (in-sample, post-sample, post-publication) is weighted to sum to one so long windows do not dominate. The regression is $y_{it} = \alpha_i + \gamma_1 \,\text{PostSample}_{it} + \gamma_2 \,\text{PostPub}_{it} + \varepsilon_{it}$, where PostSample equals one in every year after the original sample ends and PostPub equals one from the journal publication year onward (nested), so γ_1 is the change once the sample ends, γ_2 the additional change associated with publication, and $\gamma_1 + \gamma_2$ the total post-publication change. Panel A is weighted OLS with paper fixed effects α_i (the main specification). Panel B is a pooled median regression of the same outcome on the same dummies without paper effects; its intercept is the in-sample median ratio. Asset pricing is the field classification; other fields are corporate finance, intermediation, investments, microstructure, household finance and other. Standard errors in parentheses: two-way clustered by paper and calendar year in Panel A; paper cluster bootstrap (1,000 draws) in Panel B. *, **, ***: significant at 10%, 5%, 1%.

	(1) All papers	(2) Asset pricing	(3) Other fields
Panel A: weighted OLS with paper fixed effects			
Post-sample, pre-publication (γ_1)	-0.281***	-0.406***	-0.119*
	(0.070)	(0.092)	(0.072)
Publication increment (γ_2)	0.025	0.062	-0.021
	(0.074)	(0.095)	(0.082)
Total post-publication ($\gamma_1 + \gamma_2$)	-0.255***	-0.345***	-0.140**
	(0.058)	(0.087)	(0.062)
In-sample mean level	0.895	0.944	0.834
Paper fixed effects	Yes	Yes	Yes
Calendar-year fixed effects	No	No	No
Papers	836	463	373
Paper-years	27,687	17,954	9,733
Panel B: pooled median regression			
Post-sample, pre-publication (γ_1)	-0.276***	-0.342***	-0.210***
	(0.034)	(0.038)	(0.041)
Publication increment (γ_2)	-0.020	0.000	-0.043
	(0.035)	(0.040)	(0.046)
Total post-publication ($\gamma_1 + \gamma_2$)	-0.296***	-0.342***	-0.254***
	(0.029)	(0.034)	(0.044)
In-sample median (intercept)	0.773	0.781	0.768
Papers	836	463	373
Paper-years	27,689	17,955	9,734

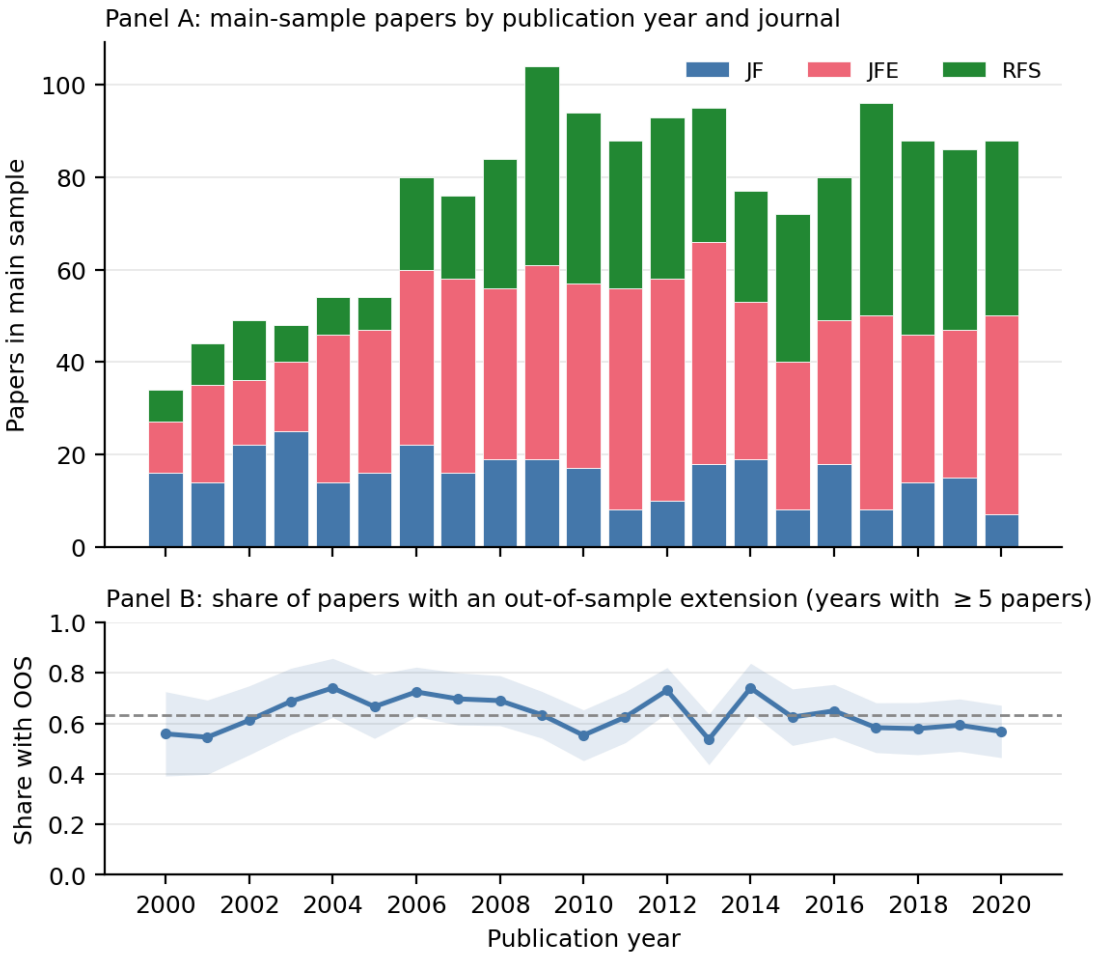


Figure 1. Sample coverage by publication year. The figure shows how the main sample and its out-of-sample availability are distributed over publication years. Panel A plots the number of main-sample papers by publication year, stacked by journal; the sample is restricted to the 2000–2020 project universe (later years are reserved for the out-of-sample test). Panel B plots the share of papers with a usable out-of-sample extension (years with at least five papers); the shaded region is a pointwise normal-approximation 95% interval and the dashed line is the full-sample availability rate. The publication year is the printed-issue year from the publisher's own journal metadata (OpenAlex or archive metadata for the <1% of papers without a publisher record). Recent years mechanically have shorter post-sample windows, which lowers OOS availability and motivates the window-length robustness checks.

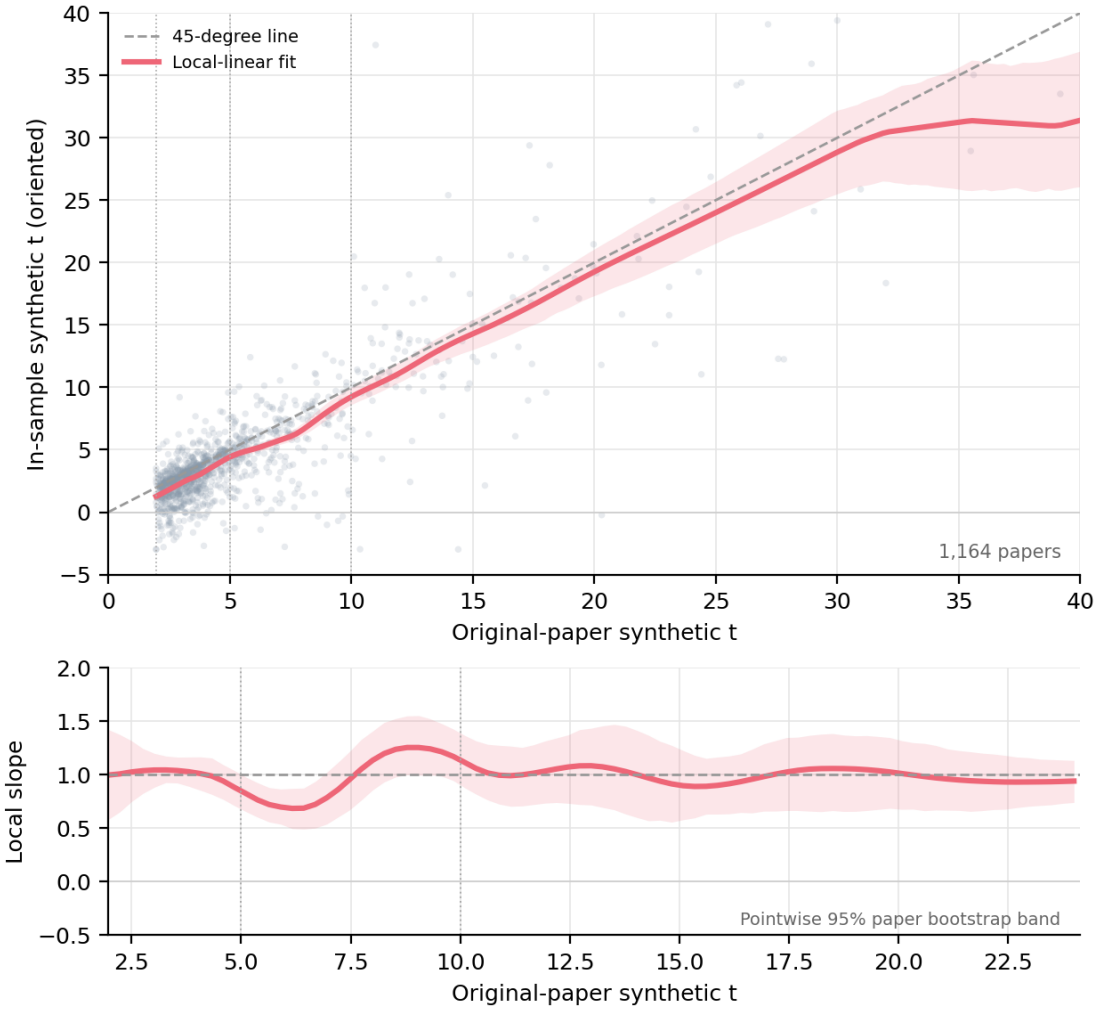


Figure 2. In-sample recovery as a function of published statistical strength. The figure shows how the statistical strength of the in-sample replication varies with the strength of the original published result. Each dot in the upper panel represents one paper. For each paper, the original synthetic t is the median absolute published t-statistic across originally significant directional key estimates (the paper's designated headline coefficients), and the IS synthetic t is the median replication t over the same estimates after orienting them to the published sign. Coefficient-level t-statistics are winsorized at the pooled 1st and 99th percentiles before taking medians, so every paper receives one observation regardless of its number of key estimates. A negative IS value indicates that at least half of the paper's eligible evidence is reversed. The red curve is local-linear LOWESS using 40% of observations per neighborhood, with a pointwise 95% interval from 400 paper bootstrap draws. The dashed 45-degree line represents equal statistical strength. The lower panel is the derivative of the LOWESS curve through the 97.5th percentile of the horizontal variable; a local slope of one means one-for-one transmission. The synthetic t is a descriptive paper summary, not a test statistic with a t reference distribution. Axis limits are for display only; all winsorized observations enter the fit.

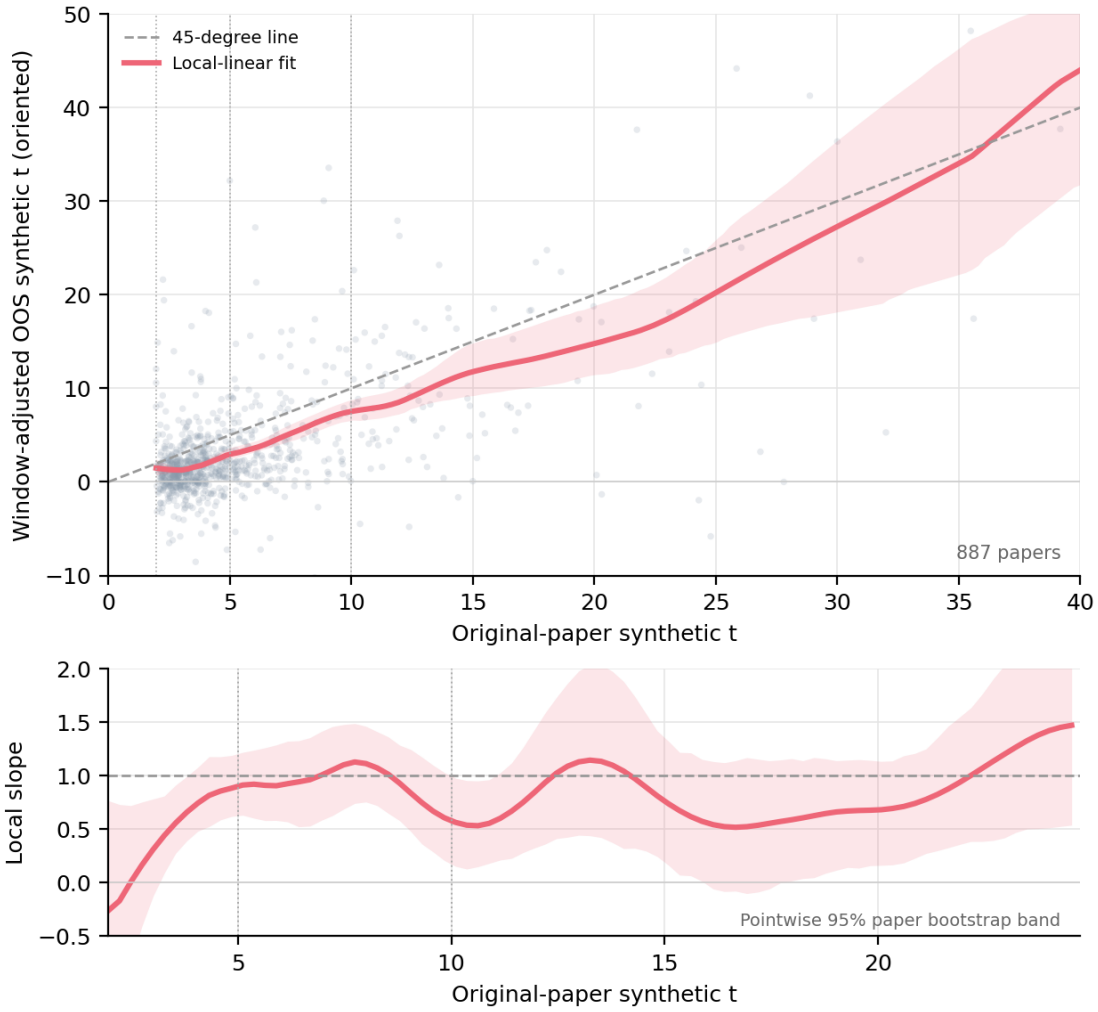


Figure 3. Window-adjusted out-of-sample statistical strength as a function of published strength. Each dot is one paper with valid OOS evidence: the horizontal axis is the paper's original synthetic t-stat (the median absolute published t-statistic over its originally significant key estimates) and the vertical axis is the window-adjusted OOS synthetic t-stat, the same median computed after dividing each OOS coefficient by the in-sample implied standard error $|\hat{\beta}_{IS}/t_{IS}|$ and orienting to the published sign. The adjustment removes the mechanical significance loss from the shorter post-sample window, so the adjusted t-stat reflects effect decay at the original window's precision (statistics with $|t| < 0.1$ on either side keep their unadjusted t; adjusted inputs are winsorized at their pooled 1st/99th percentiles before the within-paper median). Under the adjustment, an effect that is unchanged out of sample lies on the dashed 45-degree line, so vertical distance below the diagonal is genuine decay rather than mechanical precision loss. The red curve is local-linear LOWESS using 55% of papers per neighborhood, with a pointwise 95% interval from 400 paper bootstrap draws; the lower panel is its local slope. The synthetic t is a descriptive paper summary, not a test statistic.

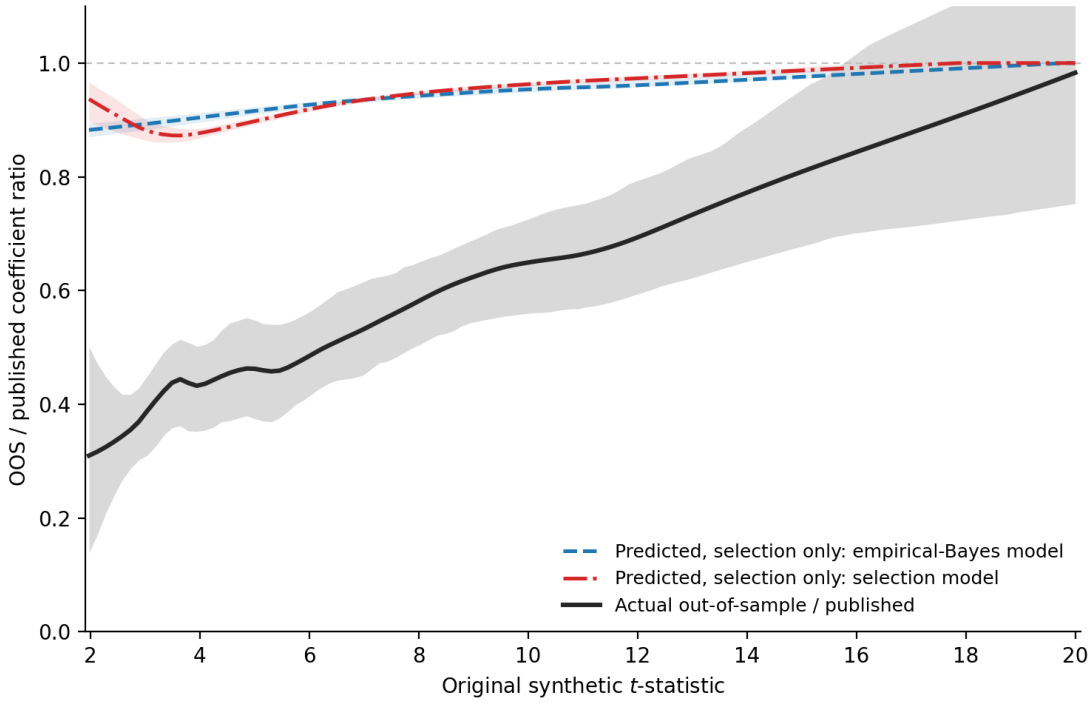


Figure 4. Publication selection predicts far less decay than observed. The black curve is a locally weighted regression (Cleveland, 1979) of the paper-level out-of-sample/published coefficient ratio on the paper's original synthetic t-stat (the median absolute published t-statistic over its originally significant key estimates); the gray region is a pointwise 95% interval from 200 paper bootstrap draws. The colored curves show the ratio each publication-selection model predicts if selection on reported significance were the only source of decay: the empirical-Bayes model adapted from Chen and Zimmermann (2020) and the selection model adapted from Andrews and Kasy (2019). Both models are fitted only to the published in-sample t-statistics and never see out-of-sample data; their bands re-estimate the models on each bootstrap draw. The vertical gap between the predicted and actual curves is decay that selection on reported significance cannot produce; Section 4 quantifies it at roughly four-fifths of the observed decline. The display spans the models' fitting window, synthetic t-stats up to 20.

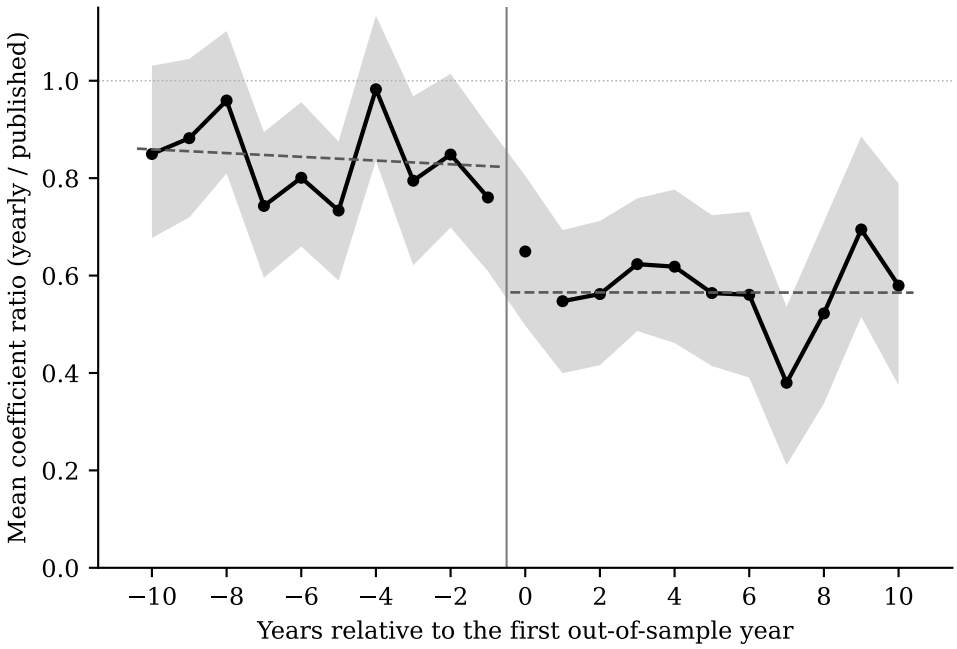


Figure 5. Coefficient ratios around the end of the original sample. Mean across papers of the yearly coefficient ratio by year relative to the first out-of-sample year. Each paper's key coefficients are re-estimated one calendar year at a time with the specification unchanged; the yearly estimate is divided by the published coefficient, oriented so the published value is positive (one reproduces the paper, zero is no effect), winsorized at the pooled 1st and 99th percentiles, and averaged across the paper's coefficients within a year; each paper is weighted equally within a year. The first out-of-sample year is drawn as an isolated point because for many papers it straddles the sample end. The shaded area is the 95% paper-bootstrap interval. Dashed lines are straight-line fits over years −10 to −2 and +1 to +10. Main sample, 836 papers. Figure A2 in the Appendix shows asset pricing and other fields separately.

Appendix A Replication Instructions

This appendix reproduces, lightly condensed, the instruction file the replication agent reads at the start of every paper (`tools/prompts/REPLICATION.md`). The agent works in a fresh session with the paper's text in `current-paper/` and follows the six parts below in order. Text in brackets summarizes housekeeping detail omitted for length.

Resources and environment. Read these reference files before starting: the WRDS schema reference (table schemas, join keys, filters); the variable-recipe file (standard Compustat/CRSP constructions); and `main_takeaways.md` (lessons from past replications: WRDS quirks, merge pitfalls, estimation tips). Tools: a WRDS download utility, Newey–West/Fama–MacBeth estimation utilities, and `pyfixest` for panel fixed-effects regressions.

Pre-cached WRDS core data: check `data/wrds-main/` first. It holds full-history parquet dumps of CRSP (daily and monthly stock files, names, delistings, indices, CCM link, mutual funds), Compustat (annual, quarterly, security daily and monthly, names, segments), and Fama–French factors, portfolios, and industry maps. Only query WRDS for tables not cached. Other pre-cached data (FRED, Datastream, Bankscope, SDC, credit ratings, and others) are indexed in `data/index_data.md`; check it before any external download.

Internet is open. Download non-WRDS data only when it is critical, not on WRDS or in `data/`, small (<1 GB), and easy to get (direct URL, no login or scraping). WRDS "sample" and "trial" libraries are tiny teaser tables: never use them as replication data; they serve only to inspect schemas.

Use Python for all code and kebab-case file names.

Part 1 — Feasibility check. Read the paper (and internet appendix if present) from its markdown conversion; convert the PDF first if needed. Follow the steps in order and stop at any gate that says STOP.

Step 1: Empirical screen. Rate 0–10 how empirical the paper is and classify its field (asset pricing, corporate finance, market microstructure, household finance, banking, other). *Gate:* if the score is ≤ 3, write `classification.json` and STOP (theory paper).

Step 2: Data screen. List every dataset the paper uses with a one-line description and a classification: WRDS (give `library.table`), PUBLIC (URL, size, login/scraping needed?), PURCHASABLE, or PROPRIETARY (contents; is there a public proxy?). Before classifying any dataset as unavailable, query WRDS for an equivalent table—many "proprietary" datasets from older papers now have WRDS equivalents—and, for government surveys and agency data, search the web for a public-use version (e.g., NSSBF, SCF, CPS, HMDA, Call Reports, Flow of Funds). Then assess table by table which datasets each main table requires; some tables may be replicable even if others need proprietary data. *Gate:* if no table can be replicated with available data, write `classification.json` and STOP. Otherwise flag the feasible tables.

Step 3: Target table. Pick the single most important feasible table. Describe the main finding in at

most 80 words. Identify the target table and the summary-statistics table used for sample-size comparison. Even when one coefficient is the headline result, plan to rerun the full target table once the core pipeline works; restrict the first pass to a narrower slice only when runtime or data force it.

Step 4: Methodology extraction. Write `methodology.md` with everything needed to replicate the target table: sample period and filters; variable definitions (exact construction, lags, scaling, winsorization); merge logic (keys, timing conventions); estimation method, fixed effects, and standard-error clustering; and any detail that appears only in the internet appendix. Tag every choice with its source: `[PAPER §II.A]`, `[PAPER Table 3 note]`, `[PAPER fn.14]`, or a short quote, when the paper states it; `[INFERRED]` with a brief rationale when the paper is silent and standard practice or judgment fills the gap. This tagging is critical: when a replication fails, inferred items are the first places to revisit. Use specific WRDS `library.table` names throughout. If the paper's sample predates WRDS coverage, note the earliest feasible start year (a shifted-period replication is acceptable and is scored as out-of-sample).

Step 5: Feasibility. Given `methodology.md`, can the table be executed with WRDS plus pre-cached public sources? *Gate:* if not, write `classification.json` with the reason and STOP. If yes, record status READY, in-sample versus out-of-sample, and proceed to Part 2.

Part 1 outputs: `classification.json` (empirical score, field, status, stop reason, replication type, dataset list with classifications, feasible tables, target table) and `methodology.md`.

Part 2 — Replication attempt. Re-read this file from here onward before continuing. Replicate the target table using `methodology.md` and the paper.

Setup. Code goes in `code/`, output in `results/`, constructed samples in `data/` of the paper folder. Use the pre-cached WRDS dumps as the primary source; query WRDS only for uncached tables and cache the result. If a needed dataset is unavailable, check the data index for a cached copy; if it is not there, log it in `missing-data.md` and continue, or STOP and write the final report if it is essential.

Runtime gate. Before committing to Part 2, estimate whether the sample build plus first-shot regression can finish within two hours after reasonable optimization (cached parquet, column pruning, date filters, aggregation before joins, a narrower feasible specification when appropriate). If the optimized runtime still exceeds two hours, do not attempt the replication: record the runtime barrier and STOP. [Rules for background jobs: log to files, record the process ID and exit code, poll until exit, verify that the expected output exists and is fresh; a running monitor is not completion; never compare results, write the evaluation, or archive while a job is running.]

Path and output integrity. [Scripts must run from both the repository root and the paper folder; canonical outputs only under the paper's `results/`, `code/`, and `data/`; the OOS script writes `oos-table.txt` and never overwrites `final-table.txt`.] `final-table.txt` must distinguish replication-generated estimates from paper-supplied comparison values. Never copy paper coefficients, t-statistics, or sample sizes into a field labeled as a replication result; when paper num-

bers are needed, label them as paper inputs with a source anchor to the paper's table, note, or equation, and spot-check them against the text before archiving.

Phase 1: Build sample. Re-read the takeaways and variable-recipe files. Write one script (`build-sample.py`) and run it once. Follow `methodology.md` literally; do not improvise. Download the full available period but filter to the paper's sample period, with the date filter as a parameter so Part 3 can reuse the code. Then follow the sample-validation protocol: match N at every level, compare summary statistics, diagnose mismatches by sub-period. If N is off by more than 30%, try up to three substantively different data constructions and document each. If still off after three attempts, STOP and write the final report (sample-construction failure).

Phase 2: First shot. Write one script (`run-regressions.py`) and run the target specification exactly as described. Save `first-shot-table.txt` and `first-shot-code.py`; these are never overwritten. Do not compare to the paper's published numbers until this phase completes.

Phase 3: Iterate. Compare the first shot to the published table (for out-of-sample replications, compare signs and significance only). If it already qualitatively replicates (signs and significance match), skip to Phase 4. Otherwise, at most three iterations. Start by revisiting the `[INFERRED]` items in `methodology.md`; each iteration must either fix an inferred assumption or cite a specific passage in the paper that motivates the change. No fishing. STOP after three attempts without convergence on signs and significance.

Phase 4: Report. Write `evaluation.json` with: result (EXACT | QUALITATIVE | PARTIAL | FAILURE), first-shot result, replication type (in-sample | out-of-sample), sample-size ratio (replication N / paper N), number of iterations, quality score, failure reason, and a long summary. Quality scores: 90–100 exact (estimates within 10%, significance identical); 70–89 qualitative (signs and significance match); 40–69 partial (some specifications only); 0–39 failure. Out-of-sample is scored on signs and significance only.

Required outputs: `evaluation.json`; `final-table.txt` (best replication output, covering the full target table when feasible; if only a subset, say so in the summary); `final-code.py` (one self-contained regression script); `journey.md` (what worked, what did not, ambiguities, documentation gaps); `takeaways.md` (generalizable lessons); `data.md` (data the paper uses for the main result and what was used or unavailable); `summary.md` (comparison of the target and replicated tables, the OOS results, how close the replication is, and likely reasons for differences). If no iteration was needed, the final output is the first shot.

Stop conditions (any one triggers Phase 4): main result replicated (exact or qualitative); essential data missing with no proxy; three sample-construction attempts with N still off by more than 30%; three specification attempts without convergence. At the end of Part 2 write a timing log with phase timings, WRDS query count, and iteration summary.

Part 3 — Out-of-sample test. Re-read this file from here onward. Assess whether an out-of-sample (OOS) test is feasible: the paper's sample must end before the available data end; the OOS period must be at least five years; the same variables and sources must be available in the

OOS period; and some tests are inherently not OOS-testable (e.g., event studies around a one-time regulatory change). If OOS is not feasible, write `oos-skip.md` explaining why and STOP.

Otherwise write `run-oos.py`, which reuses the cached data already downloaded for the full period, rebuilds the sample for the OOS period only by changing the date filter (the rest of the pipeline is identical to `build-sample.py`), runs the same specification as `final-code.py` on the OOS sample, and saves `oos-table.txt`. Do not modify the specification to fit the OOS data: the point is to test whether the original finding holds, not to optimize for a new period.

Write `oos-evaluation.json` with the OOS period, OOS N, whether signs match, whether significance matches, the OOS result—CONFIRMED (signs and significance hold), WEAKENED (signs hold, significance drops), REVERSED (sign flips), or INCONCLUSIVE (too few observations or data issues)—and notes. Add an "Out-of-Sample Test" section to `journey.md` comparing the in-sample and OOS results, and update the timing log.

Part 4 — Replication report. Read the report-format file and follow it to produce a LaTeX replication report (`replication-report.tex` → PDF) for the paper.

Part 5 — Update shared knowledge. Append the most useful generalizable lessons to the shared `main_takeaways.md` in under 100 words: only things that help replicate other papers (WRDS quirks, variable gotchas, estimation tricks, merge pitfalls), no paper-specific facts. If the replication downloaded or constructed a public dataset likely useful for other papers (e.g., CFTC positions, DGTW benchmark returns, macro announcement calendars, country-level institutional variables), save a clean parquet copy under `data/from-replications/` with a short index (source URL, download date, variables, coverage, triggering paper) and register it in the data index. Do not save WRDS data there.

Part 6 — Archive and clean. Copy the paper folder to the completed-papers archive, verify that the key files exist at the destination, then clear the working folder. Never delete before verifying.

Appendix B Additional Tables and Figures

This appendix contains additional tables and figures.

Table A1. Definitions of key analysis variables. Key estimates are the paper results designated, before any replication was run, as carrying the target table's central claim. Unless noted otherwise, each quantity is first built per estimate and then averaged so that every paper counts once. IS denotes the in-sample replication, which rebuilds the paper's own period; OOS denotes the out-of-sample extension, which re-runs the same design on at least five post-sample years.

Variable	Construction	Interpretation
Original synthetic t-stat	The median absolute t-statistic across the paper's key published estimates: the one to four coefficients that carry its central claim and are significant in the paper. Extreme t-statistics are trimmed (winsorized) at the pooled 1st/99th percentiles before taking the median.	How strong the published evidence was, measured before any replication is run. A summary index, not a test statistic.
IS or OOS synthetic t-stat	The same median, computed from the replication's t-statistics. Signs are first set so the published effect counts as positive; the OOS version also has a window-adjusted variant (below).	How strong the evidence is when the analysis is redone. A negative value means the typical estimate went the opposite way.
Category match (broad, in-sample)	Every estimate is placed in one of three classes: negative and significant, insignificant, or positive and significant at the 5% level. The match is the share of a paper's key estimates whose replication lands in the published class.	Did the replication reach the paper's qualitative conclusion on the paper's own period (Table 2)?
Original–OOS category match	The same three-class comparison for the post-sample estimate, restricted to key estimates that are significant in the paper. Passing does not require the in-sample replication to have succeeded.	Does the published conclusion still hold in the years after the paper's sample ends?
Window-adjusted OOS match	Same comparison, except each OOS coefficient is judged at the original window's precision: it is divided by the in-sample standard error, recovered as $\lvert\hat{\beta}_{IS}/t_{IS}\rvert$, which carries whatever clustering the design uses. Estimates with $\lvert t\text{-stat}\rvert < 0.1$ on either side are left as they are.	Removes the significance a result loses purely because the post-sample window is shorter. What remains is a real change in the effect.
No-decay benchmark match	The paper's own in-sample coefficient judged at the OOS window's precision (divided by the OOS standard error $\lvert\hat{\beta}_{OOS}/t_{OOS}\rvert$); same rows and the same $\lvert t\text{-stat}\rvert < 0.1$ rule.	The OOS match rate we would see if every effect were exactly as strong as published and only the shorter window mattered.

Table A1 (continued). Definitions of key analysis variables.

Variable	Construction	Interpretation
Coeff. ratio	The replication coefficient divided by the published coefficient, with signs set so the published effect is positive. Extreme ratios are trimmed at the pooled 1st/99th percentiles; each paper's key estimates are then averaged.	One means the effect is exactly as large as published; zero means it disappeared; a negative value means it flipped sign. Does not depend on the window length.
IS/original coeff. ratio (geometric mean)	The geometric mean of the in-sample ratios over key estimates that keep the published sign.	A size check for the in-sample replication that treats halving and doubling symmetrically; sign flips are excluded.
Paper-period coverage	The share of the paper's estimation window that the rebuilt replication sample covers.	A same-sample check; the main sample requires at least 50%.
Sample-size discrepancy	$\lvert \log_2(N_{IS}/N_{paper}) \rvert$ using the archived primary sample size.	Zero means the replication and paper samples are equally large; one means one is twice the other.
Annual citations	OpenAlex citations through 2025 divided by the number of complete calendar years since publication.	Attention adjusted for a paper's age; missing values are filled with the journal-year median. Percentile groups are formed within each publication decade.
Institution rank and top-5 school	The best-ranked matched affiliation across the author team. Top-5 school means at least one author at Chicago, Wharton, Harvard, Stanford, or MIT.	Descriptive measures of author affiliation, not measures of research quality.

Table A2. Robustness sorts: replication outcomes across secondary splits. Appendix analog of the strength-sort table for three splits that are informative but not part of the main analysis. Panel A splits by the model family that performed the blind re-audit; assignment of papers to audit models was not random, so differences mix paper composition with any auditor effect. Panel B splits at the median of the headline methodology measure — the share of core method choices the paper states explicitly (wave-1 coverage; unmeasured papers excluded from this panel). Panel C splits by core-data fidelity, the conditional-effects battery definition: a paper is in the substitute/proxy group when at least one core data source was replaced by a substitute or proxy series; papers whose source fidelity was never recorded are excluded. Outcomes are paper-level: category-match shares are means, coefficient ratios are medians of the within-paper mean oriented winsorized ratio. The adjusted OOS match classifies each OOS coefficient at the in-sample implied precision $|\hat{\beta}_{IS}/t_{IS}|$, removing the mechanical significance loss from the shorter post-sample window. The difference row reports the between-group gap with a Welch t-statistic in parentheses for the mean-based columns.

Panel A: by blind-audit model

	Papers	OOS papers	Original–IS category match	IS/original coeff. ratio (med.)	Original–OOS category match	Adj. Original–OOS category match	OOS/original coeff. ratio (med.)
GPT-audited	1,086	821	75.5%	0.95	41.8%	47.6%	0.45
Claude-audited	254	183	73.7%	0.94	40.9%	49.1%	0.47
Difference (t)			+1.7 (0.7)	+0.02	+0.9 (0.3)	-1.5 (-0.4)	-0.01

Panel B: by core documentation explicitness (headline methodology measure)

	Papers	OOS papers	Original–IS category match	IS/original coeff. ratio (med.)	Original–OOS category match	Adj. Original–OOS category match	OOS/original coeff. ratio (med.)
Above median	865	643	73.8%	0.94	42.7%	49.2%	0.47
Below median	478	362	77.4%	0.97	39.9%	45.7%	0.42
Difference (t)			-3.6 (-1.9)	-0.03	+2.8 (1.0)	+3.6 (1.3)	+0.05

Panel C: by core data fidelity

	Papers	OOS papers	Original–IS category match	IS/original coeff. ratio (med.)	Original–OOS category match	Adj. Original–OOS category match	OOS/original coeff. ratio (med.)
Original core sources	827	628	77.6%	0.96	43.3%	50.0%	0.45
Substitute/proxy core source	482	352	69.7%	0.90	38.8%	44.2%	0.44
Difference (t)			+7.9 (3.8)	+0.06	+4.4 (1.6)	+5.8 (2.0)	+0.01

Table A3. Robustness of Table 7: calendar-year fixed effects and fixed-effects median regression. Same panel, outcome, weights and stage dummies as Table 7. Panel A adds calendar-year fixed effects to the main OLS specification; because event time equals calendar year minus each paper's cohort, stage and calendar effects are only partly separable (post-sample years exist only in 1995–2019, post-publication years only from 2000), so this is a reference specification. Panel B is the Canay (2011) two-step fixed-effects median regression: paper effects are estimated by the OLS of Panel A of Table 7, subtracted from y, and the stage coefficients are estimated by median regression on the demeaned outcome; it assumes the paper effect is a pure location shift, and gives smaller magnitudes than OLS because each paper's OLS level is pulled up by its largest years. Standard errors in parentheses: two-way clustered by paper and calendar year in Panel A; paper cluster bootstrap re-estimating both steps (1,000 draws) in Panel B. *, **, ***: significant at 10%, 5%, 1%.

	(1) All papers	(2) Asset pricing	(3) Other fields
Panel A: weighted OLS with paper and calendar-year fixed effects			
Post-sample, pre-publication (γ_1)	-0.231***	-0.344***	-0.146*
	(0.086)	(0.127)	(0.087)
Publication increment (γ_2)	0.034	0.016	0.008
	(0.066)	(0.073)	(0.093)
Total post-publication ($\gamma_1 + \gamma_2$)	-0.198***	-0.327***	-0.139
	(0.071)	(0.113)	(0.102)
In-sample mean level	0.895	0.944	0.834
Paper fixed effects	Yes	Yes	Yes
Calendar-year fixed effects	Yes	Yes	Yes
Papers	836	463	373
Paper-years	27,666	17,933	9,731
Panel B: two-step fixed-effects median regression (Canay 2011)			
Post-sample, pre-publication (γ_1)	-0.160***	-0.254***	-0.107***
	(0.028)	(0.049)	(0.028)
Publication increment (γ_2)	-0.033	-0.015	-0.032
	(0.026)	(0.045)	(0.033)
Total post-publication ($\gamma_1 + \gamma_2$)	-0.193***	-0.270***	-0.139***
	(0.024)	(0.034)	(0.036)
Papers	836	463	373
Paper-years	27,689	17,955	9,734

Table A4. Is there a discontinuity at the end of the original sample? For each paper, a straight line is fitted to the yearly coefficient ratios (yearly estimate divided by the published coefficient, as in Table 7) over the five years before a cutoff and, separately, over the five years starting at the cutoff; the jump is the difference between the two fitted values at the cutoff. Rows report the median across papers of the jump, of the slope on each side (change in the ratio per year), and of the plain difference in median ratios between the two five-year windows. The first row uses the true cutoff, the first out-of-sample year; placebo rows move the cutoff inside the original sample (using in-sample years only) or five years into the out-of-sample window (out-of-sample years only). Data mining that inflates the whole in-sample estimate predicts a negative jump at the true cutoff and none at the placebo cutoffs; gradual market change predicts negative slopes on both sides and no jump. Bootstrap standard errors clustered by paper (1,000 draws) in parentheses. *, **, ***: significant at 10%, 5%, 1%.

Cutoff	Jump at cutoff	Slope before	Slope after	Median ratio, after − before	Papers
Sample end (true boundary)	-0.070	-0.004	-0.007	-0.110***	757
	(0.045)	(0.009)	(0.010)	(0.027)	
Placebo: 5 years before sample end	0.010	-0.004	-0.004	-0.029	710
	(0.039)	(0.011)	(0.009)	(0.022)	
Placebo: 8 years before sample end	0.013	0.003	0.000	-0.011	660
	(0.039)	(0.009)	(0.013)	(0.036)	
Placebo: 12 years before sample end	-0.005	0.002	-0.011	0.010	571
	(0.053)	(0.016)	(0.011)	(0.030)	
Placebo: 5 years after sample end	-0.056	-0.004	0.007	-0.029	638
	(0.041)	(0.011)	(0.012)	(0.027)	
Sample end, asset pricing	-0.123	-0.001	-0.008	-0.119***	434
	(0.092)	(0.018)	(0.017)	(0.040)	
Sample end, other fields	-0.036	-0.009	-0.007	-0.097***	323
	(0.048)	(0.010)	(0.012)	(0.036)	

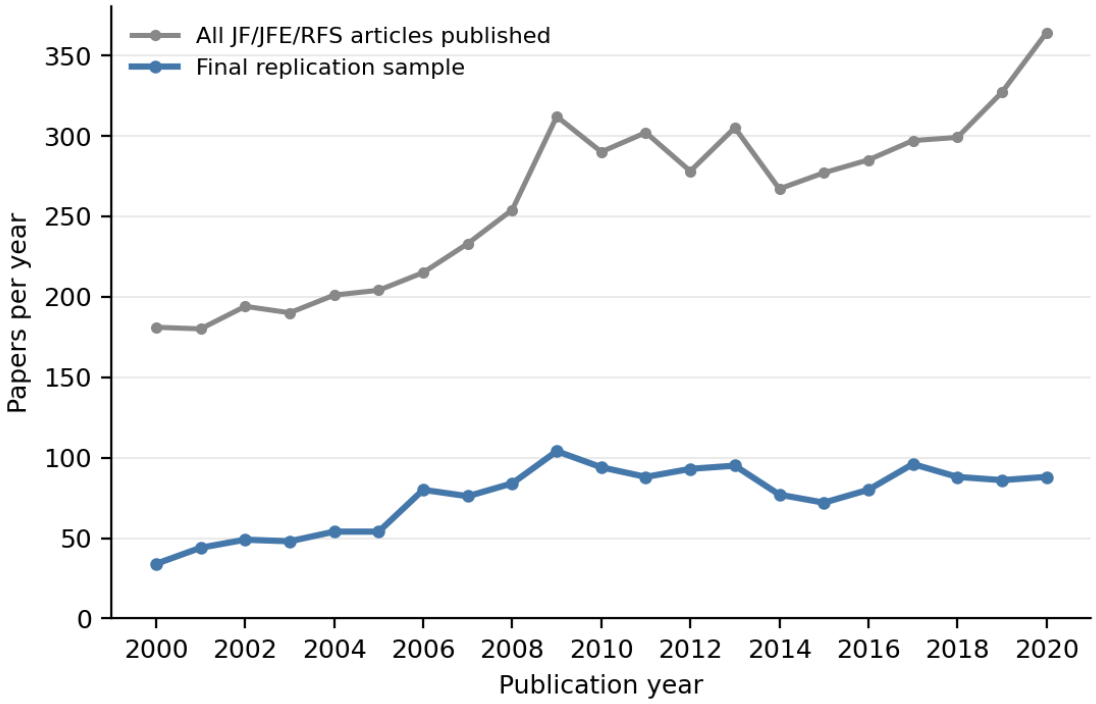


Figure A1. Replication coverage of the published universe. The figure shows what share of everything the three journals published ends up in the replication sample. For each publication year 2000–2020, the gray line counts every article the three journals published, taken from the publisher metadata of the full-journal bulk downloads; the blue line counts the papers in the final replication sample. The published count comprises full-length research, review, and survey articles; editorial material (reports, book reviews, errata, discussions) is not part of the metadata. Over the full window the final sample covers 1,584 of the 5,430 research articles (29%; 5,455 published items including non-research material); the gap consists of theory papers, empirical papers whose data are proprietary or otherwise unavailable, and replications removed by the sample filters of Table 1. Publication years are printed-issue years from the publisher metadata.

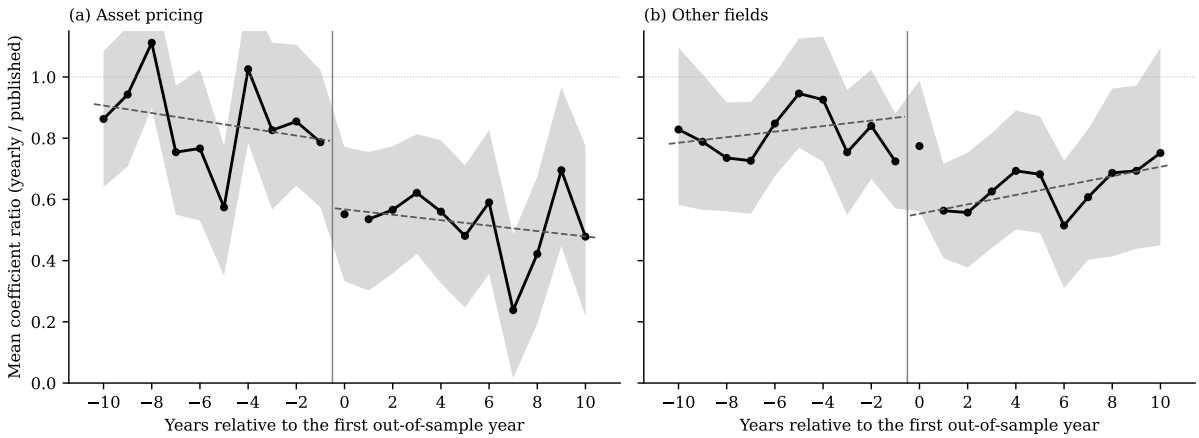


Figure A2. Coefficient ratios around the end of the original sample, asset pricing versus other fields. As Figure 5, separately for asset-pricing papers (463) and papers in all other fields (373).