Mandated Transparency and the Production of Research: Evidence from a Data- and Code-Sharing Policy*

John M. Barrios† Roman Chychyla‡

Preliminary draft — July 22, 2026

Abstract

We examine what happens to research when a journal requires authors to document their data and share the code that builds their samples. The *Journal of Accounting Research* (*JAR*) imposed such a rule on all empirical submissions in January 2015 while its closest competitors did not. We compare roughly 4,600 articles across six journals before and after the mandate. Consistent with disclosure disciplining authors, post-mandate *JAR* papers are written more plainly, stake fewer hedged claims, and report more completely. The hedging decline is concentrated in papers with the weakest headline evidence. Consistent with the mandate re-sorting who submits, *JAR* shifts toward more novel work by a more elite author pool, and its papers are cited less but by more prominent outlets. The writing effects are absent from exempt theory papers and reappear when a second journal adopts its own mandate. Overall, opening the research pipeline changes not just what can be checked and replicated, but what gets written and who publishes where.

Keywords: research transparency, open science, data-sharing policy, difference-in-differences, synthetic control, textual analysis, economics of science.
JEL: A14, C18, C52, L82, O33.

*We thank *Menaka Hampole*, *Miguel Minutti-Meza*, and *Thomas Wollmann* for helpful comments and discussions. Replication code and data are described in the Appendix and will be posted publicly upon publication. While we have reviewed for each of the six journals, neither of us have held an editorial or advisory role over the period studied. All measures are built from publicly available article text, metadata, and citation records; no journal provided data or cooperation. We study *JAR* because its January 2015 policy affords a clean start date, not to evaluate the journal, and we report costs and benefits together. John Barrios and Roman Chychyla have been rejected at least once in each of the 6 journals. Comments welcome.

†Yale University and NBER. Email: `john.m.barrios@yale.edu`.

‡University of Miami, Miami Herbert Business School. Email: `rchychyla@bus.miami.edu`

1 Introduction

Academic research trades in a marketplace of ideas and, like any asset, it is priced. Findings that readers trust are used and built on, while findings that they doubt are discounted (**?**). Over the past decade, readers have had more reason to doubt. Replications have failed at uncomfortable rates in psychology, economics, and finance (**?????**), and a parallel body of work has shown how flexibility in analysis choices produces statistical significance (**????**). When researchers can choose among many defensible sample cuts, controls, and specifications and report only those that clear a threshold, the published record drifts away from the true underlying evidence. This is not necessarily driven by fraud, but rather by the ordinary discretion that empirical work requires, exercised by people who face strong incentives to publish.

The response to this problem, across fields, has converged on one idea, transparency. Funders and journals now ask for pre-analysis plans, registered reports, and, most concretely, the data and code behind the results of a paper (**???**). The logic behind this idea is the same as for any disclosure regime. If others can inspect the materials, authors anticipate the scrutiny and write differently in the first place. This idea is familiar to economists, who have long argued that the institutions and incentives of science shape what scientists produce, from the priority rule and the reward structure of open science (**??**) to the way teams, stars, and access to data redirect the questions researchers pursue (**????**). Openness itself acts as a treatment, when biological research materials were placed in the public domain, both the rate and the direction of follow-on work changed (**?**). Disclosure mandates are now among the most widely adopted of these reforms.

Yet, almost all we know about these mandates concerns the replicability of the papers, far less is understood about how the anticipation of disclosure changes the research that gets written and the population of researchers who write it. A requirement to document one's data and hand over the code that builds them is not only a check applied after the fact. It is an ex-ante cost and it falls hardest on the opaque, fragile, or hard-to-document work that the transparency movement means to discourage. An author deciding what to write, how plainly to write it, and where to send it weighs that cost before a referee sees the paper. Whether the mandate changes behavior, changes who submits, or neither is a supply-side question, not a replication question, and it is this question that we examine in this paper.

To study these questions, we examine the *Journal of Accounting Research*'s (*JAR*) 2015 Data Policy. Beginning on January 1, 2015, authors of empirical papers had to submit a datasheet on their data sources, download dates, and who handled the data, and before final acceptance the authors had to provide the code that constructs the analysis sample

from the raw files. These materials were then posted with the published article. *JAR* did not require the actual final dataset nor do in-house replications, rather, the compliance was self-certified. The other top journals in the field—*The Accounting Review*, the *Journal of Accounting and Economics*, *Contemporary Accounting Research*, and the *Review of Accounting Studies*—imposed no similar change through the sample period of 2012–2017. We use this change as our source of identifying variation.

The accounting literature serves as a useful setting to examine these questions. Empirical work in the field is dominated by archival studies built from large datasets, often assembled by hand from proprietary or licensed sources, where the path from raw data to a regression table is long and rarely visible to readers. *JAR* is one of the three premier journals in accounting, so the mandate applies to work at the top of the market, where the incentive to publish is strongest and the temptation to cut corners most consequential. The policy was adopted unilaterally, on a clean date, by a single journal whose closest competitors did not follow. That gives us both a sharp treatment and a set of near-substitute controls drawn from the same authors, topics, and methods. A second journal, *Management Science*, adopted a comparable rule four years later, which gives us a second treated unit and variation in the stringency of the mandate which we utilize to corroborate our inferences.[1]

The mandate imposes a compliance cost that does not fall evenly across projects. Authors can respond along two margins: they may change how they write and report, or they may redirect work to other outlets (or both). We call the first the *behavioral* margin and the second the *market* margin. Ex ante, the rule admits three responses, distinguished by where the compliance cost lands: in the conduct of the authors who keep submitting, in the composition of who submits at all, or nowhere. The first two can operate at once. Under a *discipline* view, authors who keep submitting to *JAR* anticipate that the data trail and sample-construction code behind a paper will be public, so they should write more plainly, stake fewer hedged claims—most of all where the headline evidence is weakest—and report more completely, leaving insignificant results on the page rather than trimming them at the significance threshold. Under *selection*, projects that are opaque or hard to document go to other outlets, well-documented work stays, and *JAR*'s author pool and portfolio tilt toward more original contributions. Under *skepticism*, self-certified disclosure with no data editor and no in-house verification binds no one, and neither margin moves. A journal-level comparison cannot tell discipline from selection, because both produce the same reduced-form change in what *JAR* publishes. Which specific combination is supported by the data is an empirical question.

[1]While Management Science is not strictly an "accounting" area journal and publishes articles in other fields (such as management, marketing and finance), it is considered a top outlet for accounting publications.

Testing these predictions requires different kinds of evidence. Our baseline is a difference-in-differences on papers submitted to *JAR* and five accounting journals from 2012–2017, with submission-year and topic fixed effects. Our variable of interest is the $JAR \times Post$ coefficient which measures the reduced-form change relative to the field. This comparison, however, mixes conduct with composition, so we add designs matched to the margin. Our behavioral tests lean on several designs like using author fixed effects to isolate the effect of the rule on the writing of individual authors, or using a second treated journal, *Management Science*'s 2019 mandate, as well as comparisons with exempt theory papers. Our market results rely on the baseline but are checked against finance journals as a never-treated benchmark and against randomization inference and synthetic-control estimators suited to one treated unit.

To examine "the research" we begin by measuring its text. Utilizing the text-as-data approach (**?**) lets us measure how a paper is written and what it claims, not only whether it was eventually cited. We read the full text of roughly 4,600 articles published in our sample journals around the policy date and build four families of outcomes. The first is how plainly a paper is written, measured by section-level Gunning–Fog readability and a hedging measure, following the disclosure-readability literature in accounting and finance (**????**). The second is how novel the paper is, measured by the distance of its abstract and introduction from the prior literature in embedding space, alongside the combinatorial measures of **?** and the text-distance measures of **?**. The third is how its results are reported, along the dimensions of completeness, in the number of supporting estimates standing behind each headline claim, as well as in the significance distribution, which lets us look for the bunching of test statistics just above conventional thresholds that would signal specification search (**??**). The fourth is how the paper is later received in terms of its citation count and the journal-status composition of its citations. Together, these measures let us examine how the supply of research responds to the mandate along the dimensions that many in the transparency movement care about.

The two margins turn out to move separately. We start with the behavioral margin, finding that authors write more plainly the papers they publish in *JAR*. Post-mandate abstracts are about one Gunning–Fog grade more readable, a drop of roughly a third of a standard deviation. The effect concentrates in the sections of the paper that frame and interpret the results (i.e., the abstract and introduction) where a genuine change in how authors explain themselves should be most visible, and is absent from the mechanically technical methods and results sections. Two caveats are needed when interpreting this result. The raw series of *JAR* papers was already drifting toward plainer abstracts in the two years before the rule, so the level shift is not clean and a change in editors or journal style around 2015 could also produce similar writing changes. The natural worry, in other words, is that this reflects a changed roster of authors or a coincident editorial change, rather than changed behavior,

we run several tests to address these concerns. The readability effect remains robust across three tests, none of which would be anticipated by a purely compositional explanation or by an editorial change confined to a single journal. It holds within author once we exploit a second journal's mandate: *Management Science*'s 2019 data and code policy, where the same authors write more plainly after the rule than before. It scales with the stringency of the policy across journals, ordered by how much each rule actually requires. And it is absent for theory papers in *JAR*, which carry no data or code and are therefore exempt from the mandate, the within-field placebo that a behavioral story predicts and a compositional story does not. Overall, we view our readability results as genuine behavioral responses to disclosure change.

Next, we turn to how authors present evidence, documenting that this also seems behavioral. Post-mandate *JAR* papers stake fewer headline results and surround each result with more supporting coefficients, consistent with anticipating that the analysis will be re-run. Rather than selectively emphasizing the strongest outcomes, authors report more completely. This shift toward thoroughness first, trimming marginal headline claims that point the same way, does not come with any rise in the strength of the test statistics authors report. We do not, in contrast, find much evidence that the distribution of reported test statistics pulls away from the just-significant region once the same inference standard is applied to each statistic. Put differently, the mandate changes how thoroughly the authors report their results, not the significance of what they report.

Next, we consider the market margin. We find that following the mandate, *JAR* publishes papers that are more measurably novel, across every distance- and breadth-based novelty metric we construct. The estimate is robust to randomization inference, synthetic controls and holds when compared to never-treated finance articles. Yet, when we hold the author identity fixed, the novelty change is essentially zero, suggesting that this result is driven by selection, rather than a behavior change. The mandate does not make a given author write more original papers, rather it sorts more original work into the journal, plausibly because authors with distinctive, well-documented projects are more willing and able to bear the disclosure cost and least worried about scrutiny. Along these lines, we also find that the author pool at *JAR* shifts in the same direction. Relative to controls, post-mandate *JAR* authors are about 0.3 standard deviations more likely to be at a top research department and carry a higher h-index; they are directionally more junior and write somewhat shorter papers, and they draw on a different mix of proprietary and public data.

Our paper citation analysis also supports a market-side (selection) channel. Post-mandate, *JAR* papers are cited less in raw counts, but the citations that remain come from more prominent journals. Thus, even as the total audience shrinks, the marginal readers are drawn

from more prestigious venues. These more prominent citations come both from the field's leading accounting outlets and from outside the field: premier business journals account for a larger share of the smaller citation pool. Consistent with this re-sorting, hedging in the papers falls as well, but the decline is not uniform, it is concentrated in the contribution claims of papers whose headline evidence is weakest. Papers with the most to gain from dressing up thin results stop overselling once disclosure becomes mandatory. The overall drop in hedging also appears in the exempt theory papers, so it is partly due to changes in journal style. However, its concentration in empirically weaker studies cannot be explained by style alone, since a stylistic shift cannot selectively target papers with the weakest main test statistics.

Overall, our results point to the mandate changing both how a given author writes and which authors and which work the journal attracts; interpreting our reduced-form effects correctly therefore requires separating behavior from selection. The checks we report above align with this split: novelty is robust to cross-journal randomization inference and the synthetic control at the floor those designs allow, while the writing and citation effects are large and consistently signed (though less precisely estimated with one treated journal). We report the full inference battery behind each table in the appendix. Finally, this study has several important caveats, which we flag here and discuss in detail below. Treatment is assigned to a single journal, so power is limited, and several of our nulls are better read as imprecise than as evidence of no effect. Our measures of writing, novelty, and reported inference are built from text and inherit the limits of the underlying classifiers and embeddings, we treat them as informative proxies, not ground truth. And because the policy binds at submission, treatment timing is partly imputed from publication dates for the control journals, though our main results are unchanged on the observed-date-only subsample.

Our study contributes to three literatures. First, we speak to the economics of science and to research transparency (????). The economics-of-science literature historically studies how institutions, incentives, and access to ideas shape the production of knowledge (??????). A related strand shows that openness can redirect what scientists choose to work on (?). While the transparency strand of this literature asks whether existing results replicate, we show that anticipating disclosure changes the supply of research, and that this response runs through the market for submissions as much as through any individual author.

Second, we contribute to the economics of disclosure and to the textual analysis of disclosure literature (????????). This literature asks whether mandatory reporting changes real behavior or only what is disclosed, who bears the cost when underlying information is proprietary or costly to release, and what spillovers fall on parties that the disclosure was not meant to cover (??). A journal's data-and-code mandate poses the same three questions for

authors and papers, it changes how the submitted work is written and reported, it reallocates which work ends up in which journal, and it can shift the market-level composition of citations and audience attention. We treat the research paper as a disclosure document and measure those changes with text-based proxies for prose and reported inference, alongside measures of novelty and the composition of submissions and citations.

Third, we offer a template for separating behavior from composition when a policy is adopted by a single actor. The core idea is that authors can change how they write conditional on submitting, while the market can also re-sort which papers and authors arrive in the treated outlet. We implement this separation with author fixed effects, a second treated journal, within-field placebo checks, and a never-treated field benchmark, with design-based inference throughout (**??????**). The template is portable to any setting where one unit changes disclosure rules and the control units are close substitutes in the same market. The general-equilibrium intuition behind this is straightforward, if all competing outlets adopt the mandate, authors can no longer swap submission targets to avoid disclosure, so the re-sorting (composition) channel should flatten, while within-outlet changes in how papers report evidence can remain.

The remainder of the paper proceeds as follows. Section **??** describes the policy and describes our conceptual framework. Section **??** describes the corpus, the text-based measures, and the sample. Section **??** lays out the empirical strategy, our approach to inference, and a map of which design supports which claim. Section **??** presents the behavioral and market findings together with the parallel-trends, treatment-timing, and editorial-change checks that identification rests on. Section **??** discusses mechanisms and limitations, and Section **??** concludes.

2 Institutional Background and Conceptual Framework

2.1 The *JAR* data- and code-sharing policy

The *Journal of Accounting Research* is one of the three leading journals in accounting, along with *The Accounting Review* (*TAR*) and the *Journal of Accounting and Economics* (*JAE*). Its Data Policy took effect for all submissions received on or after January 1, 2015. The policy required, at the paper's initial submission, a data-description sheet disclosing how the raw data were obtained or generated, the sources, the dates on which the data were downloaded, the instrument for surveys and experiments, and which authors handled the data and ran the analyses. Authors using data obtained on a proprietary basis had to also give the editors, privately, a contact at the providing organization and the terms of the

data-sharing agreement. Prior to final acceptance, authors had to submit the computer code that converts the raw data into the dataset used in the analysis (or, where parts of that process were proprietary, a step-by-step description precise enough for another researcher to reconstruct the sample), and the datasheet and code were posted on the journal's website after publication. The requirement covered archival, experimental, field, survey, and simulation work. Purely analytical papers were not subject to the requirement.

Two aspects of the 2015 policy are as important as what it actually required. First, it established a disclosure obligation rather than a mandate to replicate the results. The journal did not require the final dataset itself or the code performing the statistical analyses, data sharing was encouraged, and authors were obligated only to retain their data and programs for six years. Moreover, *JAR* had no data editor and did not run in-house replication during our sample period. Compliance was self-certified, and the data sheet and sample-construction code were open to review by the referees during review and by others after publication. *JAR* did not extend the requirement to full analysis and table-generating code, log files, and mandatory identifiers until an April 2024 revision, well outside our window, a November 2016 update had renamed the policy the "Data and Code Sharing Policy" and added structural papers, leaving the substantive requirements unchanged. In contrast, *Management Science* introduced an active data editor in 2020. This makes clear that *JAR*'s was a disclosure rule, not a verification regime. (Appendix **??** documents the policy's full version history and the requirements each version imposed.) Second, the other leading accounting journals did not impose a comparable mandate over the 2012–2017 period. This asymmetry makes *JAR* the treated unit and the remaining accounting journals a candidate control group.

2.2 Descriptive framework

This disclosure mandate is best understood as raising the cost of publishing opaque or fragile empirical work in *JAR*. Once the provenance of the data must be disclosed, sources named, download dates fixed, a provider contact supplied, and the code that builds the sample must be handed over and posted, a result that survives only under a particular sample cut, an unusual winsorization, or an undocumented data step becomes far riskier to submit. A referee can now ask for the offending line of code, and any reader can find it after publication. The mandate works through exposure rather than through the journal re-running the analysis, and its incidence is heaviest on the practices the transparency movement targets (i.e., specification search and selective reporting) (**???**).

We use a simple framework to organize these ideas before we take them to the data. Consider an author with a project deciding whether and where to submit. Each journal

offers a prestige payoff, but imposes a cost of compliance with its standards. *JAR*'s mandate raises that compliance cost, and raises it the most for the most *opaque* projects, those whose headline result leans on choices the author would rather not disclose. Authors compare the prestige-net-of-cost of *JAR* to that of the alternatives and submit where the net payoff is highest. Two classes of responses follow.

First, we consider how an author who keeps targeting *JAR* produces and presents a paper once its data trail and sample-construction code will be public, we label this the *behavioral* margin—the *discipline* response. Our framework produces three predictions. The first concerns the clarity of the prose. A large literature in accounting and finance treats textual complexity as an economic choice rather than a stylistic accident, firms write more opaquely when they have reason to obscure weak performance, plain-language measures of readability predict how information is subsequently impounded, and clearer disclosure carries real consequences for valuation and attention (**??????**). Authors tailor the language strategically to the audience they expect to read it (**?**). The same logic applies to research writing, when the pipeline behind a paper is public and any reader can retrace it, opaque exposition no longer shields fragile analysis, and clear documentation becomes complementary to the disclosed materials. Post-mandate *JAR* papers should therefore read more plainly in the sections that describe and interpret the analysis (prediction P1). A closely related prediction concerns the language of certainty, disclosure rewards committing to a result and documenting how it was obtained, and penalizes the vague, defensively qualified claims that let a fragile finding be quietly walked back, so we expect a lower rate of linguistic hedging, especially in the rhetorical moves where authors stake their contribution (prediction P2). The third behavioral prediction concerns how completely the authors report their evidence. When the full analysis is open to re-execution, reporting only the specifications that clear the significance threshold becomes riskier, so authors should report more thoroughly, leaving more of their insignificant results on the page, and backing each headline claim with more supporting estimates rather than a few strong ones (**???**). As a corollary, the distribution of reported test statistics should pull away from the just-significant thresholds toward the null (**???**) (prediction P3).

The remaining responses are not changes in how a fixed author works, but rather changes in *which* authors and projects *JAR* attracts and in how the resulting papers are received. This is the *market* margin a single-journal mandate sets in motion—the *selection* response—and the one a naive difference-in-differences cannot separate from behavior. Consider first the originality of the work. If marginal, threshold-driven results become harder to publish once the analysis is open to inspection, the competition for a *JAR* slot tilts toward conceptual novelty. A growing literature measures scientific novelty through atypical recombination of prior ideas

(**?**), through the disruptiveness of a work relative to its antecedents (**??**), and through the distance of a work's text from the existing literature (**?**). This literature documents that genuinely novel work is both higher-impact and systematically resisted by gatekeepers (**?**). A parallel literature shows that incentives and openness shape whether scientists pursue novel directions at all (**??**). We therefore expect measured novelty to rise (prediction P4). Whether it rises because a given author chooses more original projects or because the journal attracts more original work, behavior or composition, is precisely what a within-author design is meant to settle. The composition channel is general, because submission is a choice, authors whose projects are opaque, or whose data are costly or impossible to share, do not submit to *JAR*, while those with clean, well-documented projects are relatively more willing to pay the disclosure cost, the composition of *JAR*'s authors and projects therefore shifts at the cutoff, so part of any measured change in its output reflects *who* now publishes there rather than how a fixed population behaves. This is the classic concern that an estimated effect reflects selection on unobservably different units (**?**), and it is especially acute in a market where authors sort across a few high-stakes outlets and compete for priority and placement (**???**) (prediction G1).

Once published, the paper is then evaluated by the market through readership and citations. Work published under the disclosure regime may be read, trusted, and cited differently, plausibly more selectively but by higher-status outlets, so that raw citation counts could fall even as the prominence of the journals that cite them rises (prediction P5). Finally, the control journals are also part of this same market. Projects diverted from *JAR* could then be submitted to the other accounting journals, so the natural control is itself partially treated by the displaced flow. And field-wide forces (the broader open-science movement, secular trends in how research is written and cited) can move treated and control journals together (**??**). Either channel violates the stable-unit-treatment-value assumption and biases a within-accounting comparison (**??**) (prediction G2).

Finally, the framework admits a null. Compliance is self-certified and the journal verifies nothing, so the mandate may not bind at all—the *skepticism* response—in which case neither margin moves and none of the predictions above should hold. Taking the predictions to the data is therefore also a test of whether a disclosure rule without verification has teeth.

3 Data and Measurement

3.1 Corpus

Our corpus of articles consists of research articles published between 2008 and 2025 in six leading accounting journals: *JAR*, *TAR*, *JAE*, *Contemporary Accounting Research* (*CAR*), the *Review of Accounting Studies* (*RAS*), and accounting articles published in *Management Science* (*MS*). For the general-equilibrium analysis we additionally assemble the accounting-classified articles in the three leading finance journals (the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies*), which serve as a never-treated outer benchmark.[2] For each article, we obtain the full text, parsed into sections (abstract, introduction, theory/hypotheses, methodology, results, conclusions), together with bibliographic metadata, submission and publication dates where available, author identifiers, and the forward citation record.

3.2 Measures

We construct four families of outcome measures: writing (readability and hedging), novelty, reported inference, and citations.

Readability. For each section of a paper we compute the Gunning–Fog index, a standard readability measure that increases in average sentence length and the share of multisyllabic words; higher values denote more complex, less readable prose. The measure is widely adopted in the accounting and finance disclosure literature, where textual complexity is treated as a choice that conceals or reveals information (**????**).

Hedging. Our hedging measure is constructed in two stages. In the first stage we classify *sentences* as being hedged or not. We start from the corpus of **?**, 975 sentences drawn from academic journals and manually classified as hedged or not. Next, we use these data to fine-tune a large language model classifier, DistilRoBERTa, and label every sentence in every paper in our sample. The fine-tuned classifier recognizes epistemic uncertainty and qualification (constructions such as "may suggest," "appears to be consistent with," "we cannot rule out," "broadly in line with"), unlike a fixed keyword list, which misses the many ways authors qualify a claim and misclassifies in cases of negation and quotation. A sentence is classified as hedged when the model's predicted probability exceeds 0.5. The paper- or section-level

[2]We classify a *Management Science* or finance-journal article as accounting if it either studies an accounting topic according to the OpenAlex database or at least one of its authors has previously published in an accounting journal, per the BYU accounting-research rankings list.

hedging rate is then the share of hedged sentences, $h =$ (hedged sentences)/(total sentences). A higher rate means more qualified, less committal prose.

In the second stage we locate *where* in the paper the hedging occurs, because the framework predicts movement in the parts that frame and stake a claim, not in the mechanical reporting of results. We parse each introduction into paragraphs and classify each paragraph by its rhetorical function using the "create-a-research-space" (CARS) model of research-article introductions (**??**). CARS characterizes an introduction as a sequence of rhetorical moves in which the author establishes the territory, carves out a gap, and then occupies it; we classify each paragraph into the corresponding functions (motivation, gap, present research, results, and contribution), together with flags for whether a paragraph claims novelty, importance, or identification. With sentences scored for hedging and paragraphs scored for their function, we compute a hedging rate *within* each rhetorical area. Our headline writing outcome is the introduction hedging rate, as it is the most prominent area where the contribution is stated. A paragraph counts toward a function when that function is either its primary or its secondary label, which captures the gap or contribution language embedded in a paragraph whose main job is motivation. We validate the sentence classifier against held-out hand-labeled sentences and confirm that hedging coverage is high and balanced across treated and control journals to make sure that differential measurement error does not drive the results.

Novelty. We measure novelty in three ways, to ensure robustness. Our main measure is textual, as we embed each paper's abstract and introduction with a pretrained sentence-embedding model (Qwen3-Embedding-4B, 2560 dimensions) and take the mean cosine distance between the paper and its twenty nearest neighbors among articles submitted in the prior three years (k-nearest-neighbor distance, $k = 20$), pooling the prior-literature reference set across all journals so the measure is comparable across treated and control samples. A larger distance means the text is further from the recent literature. Because one embedding choice could drive our measure, we rebuild it across $k \in \{1, 5, 10, 50\}$, a five-year reference window, an abstract-only variant, a distance-to-field-centroid variant, and an "originality" variant, one minus the maximum similarity to any prior paper, in the spirit of **?**.[3]

Our second measure is combinatorial, built from the reference list of the papers rather than the text. Following the recombination logic of **?**, we score the atypicality of the journal pairs a paper cites against a degree-preserving null, and we summarize the diversity of the fields it draws on, the entropy and the non-accounting share of its cited literature, as well as the average age of its references. These citation-based measures use no text at all, so their agreement with the embedding measure is an independent check rather than a mechanical

[3]The effect is stable across all of them (Appendix **??**).

one. Third, we use the introduction classifier to flag whether a paper makes an explicit rhetorical claim of novelty, we keep this separate from the distance measures throughout, because a claim of novelty is a statement authors make, not a property of the text, and the two turn out to be close to uncorrelated.

Reported inference. We use the Qwen 3.6 27B large language model to extract the test statistics reported in each paper's tables and recover, for each reported coefficient, an approximate two-sided p-value from the reported t- or z-statistic or from the coefficient and its standard error. Where a table reports only significance stars, the p-value can be located only within a coarse band, so we exclude star-imputed results from the test-statistic outcomes. Because authors showcase some results and relegate others, we also classify each extracted result by salience: the full universe of extracted coefficients, the primary hypothesis tests, the results discussed in the introduction, and the paper's single most-important (headline) result. We then summarize, at the paper level and within each salience tier, the share of results falling in conventional p-value bands and, following **?**, the mass of z-statistics located in a caliper just above versus just below the 1%, 5%, and 10% thresholds, the standard signature of threshold-driven specification search.

Citations. We built our citation outcomes from the OpenAlex citation graph. We match each paper to OpenAlex by DOI (97.5% of the sample) and retrieve every indexed work that cites it, together with the citing work's journal and publication year. From this record we construct three types of outcome. The first is quantity: the log of one plus the paper's total accumulated citations (a count restricted to the four years after publication behaves similarly). The second is field-adjusted impact: the log of one plus the paper's field-weighted citation index, which scales the citations a paper receives by the count expected for works of the same field, type, and publication year. The third is the status composition of the citing journals, which asks where the citations come from rather than how many: the share of a paper's citations originating in the top six accounting outlets, the share originating in seven top finance and economics journals, and the mean prestige of the citing outlets, computed by matching each citing journal to its Scopus prestige scores, SJR (Scimago Journal Rank) and SNIP (source-normalized impact per paper), in the year the citation appears and averaging across the paper's citations.[4] Because SJR and SNIP exist only for Scopus-rated outlets, the

[4]The accounting set comprises the six journals of our sample universe: *JAR*, *TAR*, *JAE*, *CAR*, *RAS*, and *Management Science* (on the citing side, all *Management Science* articles count, not only its accounting section). The finance–economics set comprises the *Journal of Finance*, the *Journal of Financial Economics*, the *Review of Financial Studies*, the *Quarterly Journal of Economics*, the *American Economic Review*, the *Journal of Political Economy*, and *Econometrica*.

prestige means are computed over the Scopus-rated subset of a paper's citers. We also report variants that exclude the six home-field outlets, which isolate the prestige earned outside the accounting elite, in Table **??**.

3.3 Sample and treatment timing

For our main sample we restrict our attention to archival and experimental research papers with an abstract and a classifiable introduction, yielding roughly 4,600 papers in the six accounting journals. Treatment timing is assigned at the level of the submission date, since the policy binds at submission. To determine submission we use the manuscript's received date where observed, and otherwise impute it from the publication date net of a calibrated review lag, dropping papers whose imputed timing straddles the January 2015 cutoff ambiguously. A paper is classified as *Post* if its assignment date falls on or after January 1, 2015, and *JAR* if published in the journal. Our main sample window is the three years on either side of the policy (submission years 2012–2017), in robustness checks, we widen this window. The treated papers do not need to have their submission date imputed as *JAR* reports a received date for every article in the window, so its treatment timing is observed directly, and all imputation falls on the control journals (this is balanced across the pre and post periods).[5]

Table **??** reports descriptive statistics for the main outcomes and controls. The controls used (paper length, the number of authors, mean author h-index and career age, and an indicator for authors at top research departments) are the paper- and author-level characteristics available for both treated and control journals. Because authorship is key to the selection story, Tables **??** and **??** describe the author characteristics. Table **??** reports team and author characteristics by journal. Author teams average roughly two to three authors, and *JAR* sits at the top of the field on author seniority and top-20 affiliation, alongside *JAE*. Table **??** previews a key compositional fact, *JAR*'s author pool moves around the policy relative to controls. Specifically, post-mandate *JAR* tilts toward more top-20-department and more junior authors. The author panel underlying the within-author estimates contains 9,770 author-paper observations on 3,774 unique authors in the accounting sample, of whom 1,794 appear more than once and 603 publish in both *JAR* and a control journal. The within-author estimates in Section **??** identify off the subset of this panel observed inside the 2012–2017 estimation window (about 2,500 author-paper observations and 810 repeat authors), which is why the switcher counts there are smaller.

Figure **??** plots the raw data behind our estimates. For the main outcome in each text-based channel, it shows the yearly gap between the *JAR* mean and the pooled accounting-

[5]Section **??** confirms that re-estimating on the observed-date-only subsample leaves the headline results unchanged.

control mean, normalized so that the average 2012–2014 gap is zero. The pre-policy gap is noisy but trendless, and it shifts after 2015, toward lower hedging and complexity and higher novelty; the raw pre-to-post movement in each panel is close to the regression difference-in-differences reported alongside it (readability −1.30 raw against −1.25 estimated). The confidence bands also make plain how much year-to-year noise comes with a journal that publishes only a few dozen papers a year, which is why we treat inference as carefully as we do when we check parallel trends and design-based inference in the results below.

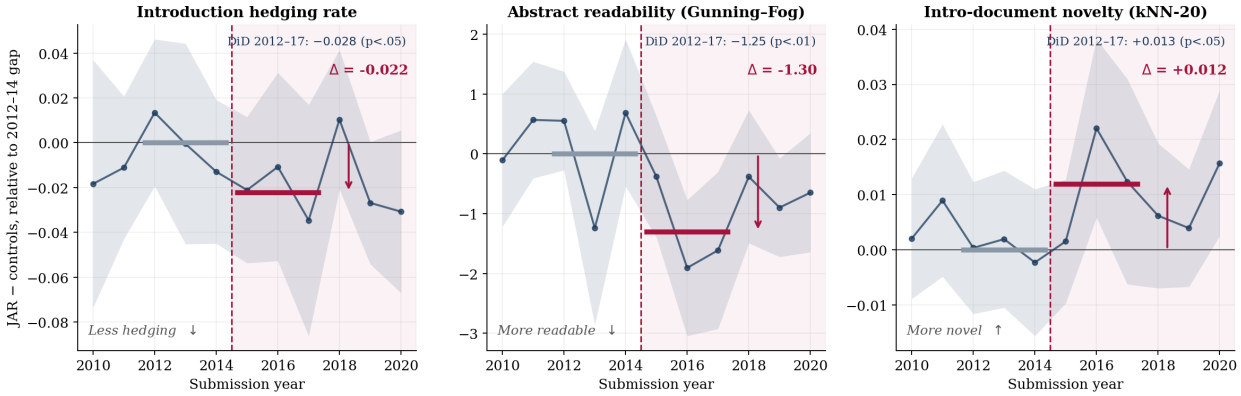

Figure 1: The raw *JAR*–control gap around the 2015 policy. Each panel plots the yearly difference between the *JAR* mean and the pooled accounting-control mean, demeaned so the 2012–2014 average gap is zero, with a 95% confidence band from the yearly difference in means. The grey and red bars mark the mean gap over the 2012–2014 pre-window and the 2015–2017 post-window (the estimation sample); the arrow is the raw level shift Δ, shown next to the corresponding regression difference-in-differences. Years after 2017 are displayed for context but fall outside the estimation window. Series are unadjusted; no controls or fixed effects enter this figure.

4 Empirical Strategy

4.1 Difference-in-differences

Our baseline compares papers in *JAR* to papers in the control journals, before and after the January 2015 policy. For paper i in journal j submitted in year t, we estimate

$$y_{ijt} = \beta\,(JAR_j \times Post_t) + \gamma\,JAR_j + \delta_t + \theta_{k(i)} + X'_{ijt}\lambda + \varepsilon_{ijt}, \tag{1}$$

where y_{ijt} is one of the writing, novelty, reported-inference, or citation measures from Section **??**; JAR_j marks the treated journal; $Post_t$ equals one for submission years 2015 and later; δ_t are submission-year fixed effects; and $\theta_{k(i)}$ are research-topic fixed effects. The coefficient β on the $JAR \times Post$ interaction is the difference-in-differences estimand: the

change in *JAR* papers from before to after the mandate, net of the contemporaneous change in the control journals.

The fixed effects and controls are chosen to absorb the obvious alternatives. Submission-year effects δ_t account for anything common to the field in a given year, including the broader open-science movement and secular trends in how accounting research is written and cited; identification comes only from *JAR*'s movement *relative* to the controls in the same year. Topic effects $\theta_{k(i)}$ ensure we are not simply comparing a different mix of subjects, whose writing and citation conventions differ. The control vector X_{ijt} holds fixed paper length, the number of authors, mean author h-index and career age, and an indicator for authors at top-20 departments, so that β does not absorb a shift in the size, seniority, or institutional composition of the work *JAR* attracts; Section **??** returns to that composition as an object of interest in its own right. When a citation or p-value measure is itself the outcome we drop it from the controls.

We report the estimates for two comparison sets. The "Top-3" design pits *JAR* against the other two leading accounting journals, *TAR* and *JAE*, the closest match in prestige and audience. The "All-Accounting" design adds *CAR*, *RAS*, and the accounting papers in *Management Science*. The two deliver a similar picture throughout. We estimate (**??**) by OLS and cluster standard errors by journal × submission-year.[6]

We trace the response over time to examine the assumption of parallel trends that underlies our specification (**??**). Specifically, we estimate an event-study analogue that replaces $Post_t$ with a full set of indicators for submission year relative to the 2015 cutoff, each interacted with *JAR*:

$$y_{ijt} = \sum_{s \neq -1} \beta_s \left(JAR_j \times \mathbf{1}[t - 2015 = s] \right) + \gamma\, JAR_j + \delta_t + \theta_{k(i)} + X'_{ijt}\lambda + \varepsilon_{ijt}. \tag{2}$$

The year before the policy ($s = -1$) is the omitted category, so each β_s measures the *JAR*–control gap in event-year s relative to that baseline. A flat, near-zero pre-period sequence supports the identifying assumption, the post-period coefficients trace the dynamic effect.

[6]The treatment is assigned to a single journal, which warrants a discussion on the standard errors. With one treated cluster, analytic cluster-robust standard errors can be anti-conservative (**????**). We therefore interpret the clustered t-statistics as descriptive and confirm every result with design-based procedures that do not lean on large-cluster asymptotics such as randomization inference that relabels each control journal as the treated unit, the restricted wild cluster bootstrap, an in-time placebo, and synthetic-control and synthetic difference-in-differences estimators built for a single treated unit (**???**). Appendix **??** reports these checks as well as multiple-testing-adjusted p-values across the family of outcomes (**??**).

5 Results

We structure our results along two dimensions. The first is *behavioral*: holding the author constant, we examine whether the rule changes how a given researcher writes and reports. The second is the *market* dimension: we assess whether the rule alters *which* authors and *which* projects the journal attracts, as well as how the resulting papers are subsequently cited.

5.1 The behavioral margin: what a given author changes

This subsection reports the changes in writing and reporting that appear in *JAR* papers once the disclosure rule is in place. We refer to these as the candidate behavioral responses. We examine writing first, then the reporting of statistical evidence, and then two dimensions of heterogeneity, the opacity of the underlying project and the strength of the paper's headline result that speak to potential mechanisms.

5.1.1 Writing: papers get plainer and less hedged

Table **??**, Panel A reports estimates of specification (**??**) for the writing outcomes. Post-mandate *JAR* papers are easier to read with abstract complexity falling by 1.25 Gunning–Fog points and introduction complexity by 0.96, each roughly a third of a standard deviation. In terms of economic magnitude, a 1.25-point Fog decline is slightly more than one U.S. grade level of reading difficulty. Consistent with the framework's first prediction, the largest changes appear in the sections a reader would use to evaluate the claims of the paper.[7]

We next decompose the readability gain into its two mechanical components, because the decomposition allows us to speak to whether the effect is simply cosmetic. The Gunning–Fog index is the sum of average sentence length and the share of complex, multisyllabic words, so the $JAR \times Post$ effect splits exactly into the part each component drives. Figure **??** reports the split by section. The complex-word channel accounts for roughly two-thirds of the introduction effect and nearly all of the whole-paper effect, and it accounts for essentially all of the statistical significance that survives the split. The sentence-length channel is small and never significant on its own. Post-mandate *JAR* authors are therefore not simply chopping long sentences into short ones, the easiest way to move a readability score without saying anything differently, but rather choosing plainer words. Moreover, an independent vocabulary-based index (SMOG) moves in lockstep with Fog (Appendix **??**), further strengthening our inferences.

[7] Section **??** examines a pre-period dip in abstract readability; as a result, we utilize the designs in Section **??** to further examine the behavioral interpretation for the drop.

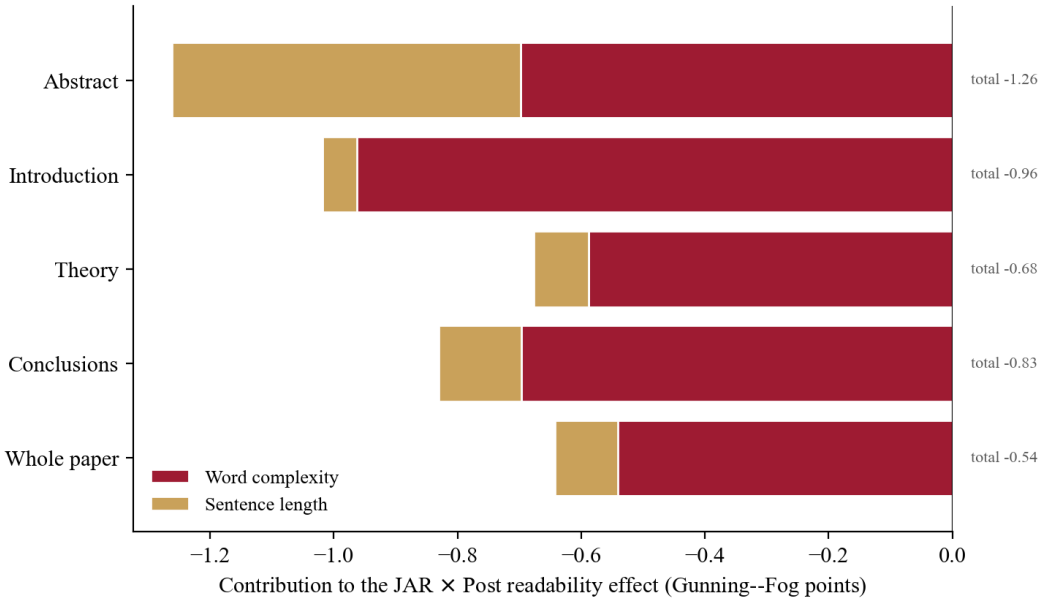

Figure 2: The readability gain is simpler words, not shorter sentences. The $JAR \times Post$ effect on Gunning–Fog, decomposed by section into the contribution of its two components: average sentence length (gold) and the share of complex, multisyllabic words (crimson). The two sum exactly to the total effect (labeled at right). The word-complexity component dominates in every section and accounts for the significance; the sentence-length component is small.

Next we examine hedging in authors' writing, finding that after the mandate the introduction hedging rate falls by 0.028, around 0.3 standard deviations. Read together, the writing results say that when the data construction of a paper is open to inspection, authors stop using vague, qualified prose to hedge a result and start saying plainly what they did and what they found. Two results ahead qualify this reading, and we flag them now. The average hedging decline also appears in the exempt theory papers, so part of it reflects journal style (Section **??**). Its concentration in the papers with the weakest headline evidence does not, since a style change cannot target the papers whose test statistics are thin (Section **??**).

The writing response varies across the different parts of the paper, and this variation serves as an internal check. Figure **??** plots the readability effect section by section and the hedging effect part by part. Both line up the same way. Readability improves by 0.7 to 1.3 Fog points in the abstract, introduction, theory, and conclusions, the sections that motivate and interpret the result, while the methodology and results sections, where notation and reported numbers fix the prose, do not move. Hedging falls almost entirely in the two rhetorical areas where authors stake a claim, the statement of the research gap (a drop of 0.087, about an eighth of its mean) and in claims of contribution (-0.091), while the motivation, present-research, and results claims are flat. A mandate that merely changed the typesetting, or that shifted JAR toward a different mix of topics, would not explain how

it was only the persuasive claims that moved in both panels. The underlying coefficients and t-statistics, section by section and claim by claim, are reported in Appendix Tables **??** and **??**.

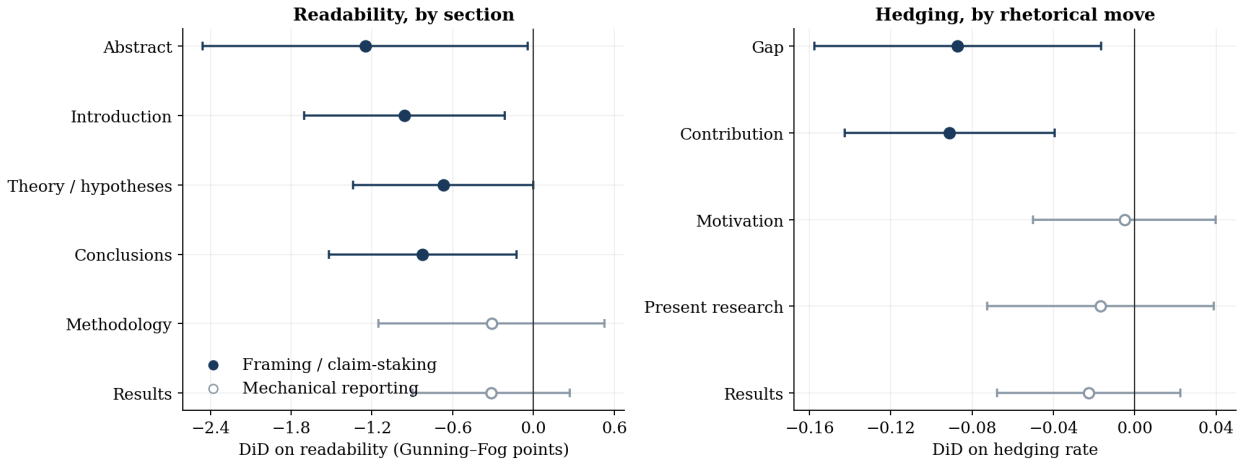

Figure 3: The mandate moves persuasive language, not mechanical reporting. $JAR \times Post$ estimates (All-Accounting; 95% confidence intervals). Left: readability (Gunning–Fog) by section. Right: hedging rate by introduction rhetorical move. Filled markers are framing/claim-staking components; hollow markers are mechanical reporting. The effect concentrates in the framing components and is absent in the mechanical ones.

The timing of the hedging decline also lines up with the policy. Figure **??** reports event-study estimates of (**??**) for the gap and contribution claims. In the gap move the *JAR*–control gap is near zero through the pre-period and turns negative once the mandate binds, the contribution claims show the same pattern more noisily, as one would expect from a sub-paragraph measure in a single selective journal. Overall, the dynamics are consistent with the framework's prediction that authors become less guarded about their gap and contribution claims once the paper's data trail and code will be public.

5.1.2 Reported inference: more nulls on the page, no robust retreat from the thresholds

We next examine how authors report the statistical evidence beneath the prose. If the mandate discouraged tuning a specification until it just cleared a significance cutoff, the test statistics authors report should pull away from the conventional thresholds and toward the null. We find that half of this prediction holds: post-mandate *JAR* papers leave more of their insignificant results on the page, but the apparent retreat from the one-percent threshold is a migration to the five-percent threshold rather than a move into insignificance.

Table **??** establishes the baseline. Reported significance is the norm in this literature:

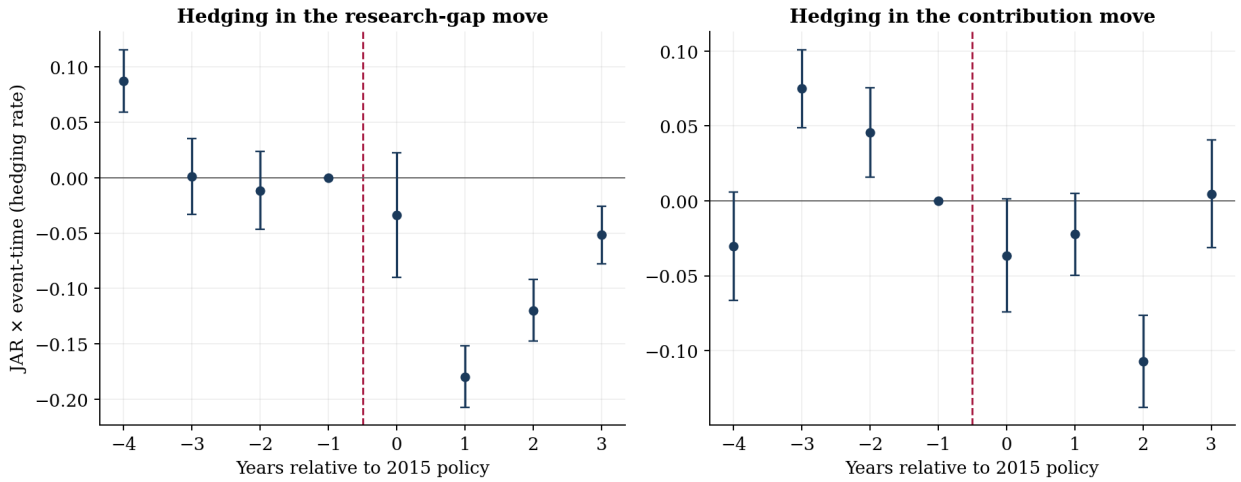

Figure 4: Hedging in the claim-staking moves, by event time. *JAR* × event-time coefficients from (**??**) for the research-gap and contribution rhetorical moves (reference year −1; 95% confidence intervals clustered by journal×submission year).

across every extracted coefficient the median reported p-value is 0.13 and 62.6% clear the five-percent cutoff, and among the single most-important result per paper the median p-value is 0.002 and 92.5% clear five percent. The introduction- and abstract-discussed results fall between these two extremes. Star-imputed rows move these levels trivially—the with- and without-imputation columns never differ by more than 0.001—so the threshold aggregates that follow are built on genuine test statistics only.

Table **??** estimates the specification (**??**) on the share of the reported results of a paper falling in each $|z|$ band, for four nested sets of results: every coefficient extracted (Panel A), the results discussed in the introduction (Panel B), the primary hypothesis tests (Panel C) and the single headline result (Panel D). Across the universe of extracted coefficients, the share of statistically insignificant results ($|z| < 1.65$) rises by 4.3 percentage points against *TAR*+*JAE* ($t = 2.53$) and by 4.4 points against all accounting controls ($t = 2.23$), from a pre-mandate *JAR* mean of 27.7%. To put this magnitude in context, this is a relative increase of about one-sixth in the share of reported nulls. No single significant band offsets the gain, and mean $|z|$ is essentially unchanged, so the rise reflects added mass at the null end of the distribution rather than a reshuffling across thresholds. Section **??** identifies the source as authors staking fewer headline numbers and surrounding each with more supporting estimates, disproportionately weaker ones, so the null share rises because reporting is fuller.

The evidence at the thresholds themselves is weaker. Reading down the salience hierarchy, the share of showcased results bunched just above the one-percent cutoff falls, and it falls most where the incentive to push a statistic across a line is largest: by 0.077 for the primary results that carry each paper, against 0.015 across the universe of incidental coefficients (**??**).

Taken alone, this pattern would suggest deterred specification search. Three additional pieces of evidence argue against that reading. First, the mass that leaves the just-above-one-percent caliper does not move into the null region; it reappears just above the *five*-percent cutoff, where the showcased share rises by about 0.10. Second, the overall share of significant showcased results, if anything, increases, and the null share among showcased results does not rise. Figure **??** shows the corresponding raw shift. The pre-policy spike in the density of absolute z-statistics in the $[2.0, 2.6)$ region, just above the five- and one-percent cutoffs, thins after the policy, but the mass does not collapse into the body below $z = 2$; much of it settles just above the five-percent cutoff, consistent with statistics relocating across thresholds rather than leaving significance.

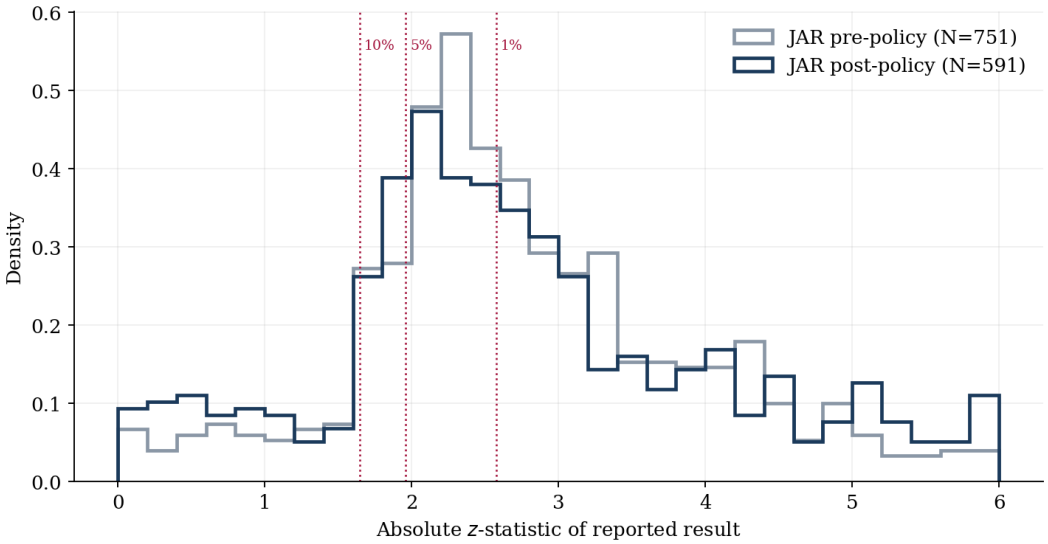

Figure 5: Reported test statistics in *JAR*, before vs. after the policy. Density of the absolute z-statistic of reported results in *JAR* papers, pre- and post-mandate. Dotted lines mark the ten-, five-, and one-percent two-sided thresholds. Post-mandate, the spike just above the one-percent threshold thins while mass appears just above the five-percent threshold, the pattern of a relocation across cutoffs rather than a retreat from significance (Section **??**).

Third, with a single treated journal, the one-percent decrease does not survive randomization inference as *JAR* is at the extreme of the placebo distribution, but with a one-sided p of only 0.17. We therefore treat the test-statistic channel as an unresolved null, consistent with reported statistics sorting one threshold down rather than with less specification search. Appendix Figure **??** reports the complete caliper picture for the various thresholds.

5.1.3 Reporting discipline: fewer headline claims, more supporting evidence

A natural follow-up question is how do the reported-inference results of Section **??** come about, do post-mandate *JAR* papers test more thoroughly or simply report fewer, stronger headline

numbers? Following the test-multiplicity logic of **?**, we decompose each paper's reported results into its headline ("primary") results and the full census of reported coefficients, and estimate equation (**??**) on the counts.

Table **??** reports the results. First, *JAR* papers report *fewer* headline results after the mandate with the log count of primary results falling by 0.11, which translates to about 10% fewer headline claims in proportional terms, with the raw mean dropping from 2.6 to 2.0 per paper. Second, the total number of reported coefficients does not fall. If anything it rises, from roughly 49 to 65 per paper, with the number of tables unchanged, so the ratio of total to headline coefficients climbs sharply (a difference-in-differences of about +14). Put differently, each headline claim in a post-mandate *JAR* paper is surrounded by many more supporting estimates. This is the pattern one would expect when the pipeline behind a result is open to inspection, authors stake fewer headline claims and back up each with more robustness. This pattern also has a selection flavor, since reporting fewer headline results is consistent with trimming marginal findings. On this margin the discipline and selection channels seem to go in the same direction.

To disentangle the behavioral from the compositional margin in this case we next focus on within-author variation. Absorbing author identity with author fixed effects (the design Section **??** formalizes), we find that the same authors report about 15 more coefficients in their post-mandate *JAR* papers than in their other work. That is roughly five-sixths of the cross-sectional gap of 18. The rise is not driven by the more junior or elite entrants as triple-interactions are insignificant. To isolate what these additional coefficients are, we turn to their significance levels and magnitudes. We find that the share of reported results that have clear conventional significance falls by seven to eight percentage points, while the mean magnitude of the test statistics is unchanged. This suggests that the coefficients that authors are adding are disproportionately weaker, non-significant ones. The added mass is in the supporting coefficients, not the main ones, which is how the falling overall significant share coexists with Section **??**'s finding that the significant share among showcased results, if anything, rises. In other words, post-mandate *JAR* papers seem to show more of the supporting evidence and lean less on a few headline numbers, with no inflation in the strength of the statistics they report (see Section **??** and Appendix Figure **??**).[8]

Next, if the response is driven by disclosure cost, it should be larger where that cost is larger, such as more opaque projects whose data and pipeline are harder to document and share. We test this prediction directly. From the disclosed data descriptions in papers,

[8]We note that this is a *JAR*-specific response: *Management Science*'s 2019 mandate (the second treated journal of Section **??**) does not reproduce it, so we read it as discipline particular to the *JAR* setting rather than a general property of disclosure mandates.

extracted with the same Qwen 3.6 27B model used for reported inference, we construct a continuous project-opacity index (private versus public ownership, restricted versus open access, proprietary-private status, with a Compustat-like adjustment that lowers the opacity score for canned public panels that can simply be acquired for nominal fees) and add a triple interaction $JAR \times Post \times Opacity$ to the baseline.

Table **??** reports the estimates, and the evidence is mixed. The two most language-elastic outcomes move in the predicted direction: abstract readability improves more, and introduction hedging falls more, for higher-opacity projects (about -0.40 Fog and -0.021 hedging per standard deviation of opacity).[9] The same relation does not appear in main-text readability, reporting discipline, or novelty, and it does not survive when opacity is proxied by the single proprietary-private flag, so we regard the channel as suggestive. Whether the mandate also changed the opacity composition of the pool, exit rather than effort, is a sorting question, which we take up with the other composition results in Section **??**. Overall our results above indicate that the disclosure cost is paid through behavior, the same authors write more plainly on their harder-to-document projects.

5.1.4 Heterogeneity by the strength of the headline result

Next we focus on the incentive side. If the mandate deters authors from overstating thin evidence, the writing effects should be concentrated where overstatement pays most, in papers whose headline result is weak. To examine this, we split papers at the sample-median headline test statistic ($|t| = 3.12$, taken from each paper's self-identified most important result) and add $JAR \times Post \times Weak$ to the baseline specification.

Figure **??** graphs the net $JAR \times Post$ effect for weak (filled) and strong (hollow) papers, and Appendix Table **??** reports the underlying triple-interaction estimates. The decline in hedging is carried almost entirely by the weak papers. Introduction hedging falls by 0.053 among weak papers against an insignificant -0.011 among strong ones, and contribution hedging falls by 0.185 against an insignificant -0.029. We also see effects given the significant triple interactions (-0.042 and -0.156). In terms of magnitude, the weak-paper effects are five to six times larger than the strong-paper coefficients. In terms of pre-period standard-deviation units, these are declines of roughly two-thirds to a full standard deviation. These results suggest that the full-sample hedging declines documented above are not a broad shift in the journal style but rather they come from the papers whose headline evidence is thin.

Three additional checks discipline our reading of the result. First, the concentration is not mechanical. Pre-mandate, JAR authors already hedged their contribution claims more

[9] Both at the randomization-inference floor of 1/6 that five placebo journals allow, and both robust to a provenance-only index.

when the headline evidence was weaker (a slope of -0.058 per log point of headline strength), and the raw within-*JAR* correlation between contribution hedging and headline strength falls from -0.33 before the mandate to -0.12 after. The mandate does not shift all weak papers by a constant, rather it compresses precisely the strategic margin on which prose tracked evidence. Second, the concentration is specific to claim-staking language. Gap hedging is the exception in Figure **??**, falling for strong papers (-0.152, $t = -4.24$) rather than weak ones (a net of -0.065), and the readability improvements are shared by both groups, the weak-paper abstract-Fog net difference of -1.34 is at most marginal, and neither readability triple interaction is distinguishable from zero. The mandate changes how strongly authors claim a contribution, not the complexity of their sentences.

Finally, the concentration does not just reflect the selection of papers. If the mandate pushed the weak papers with the most inflated claims to *JAR*'s neighbors, the weak papers still at *JAR* would be the ones that never overstated their evidence, and measured hedging would fall with no author changing a word. For that story to be true, the weak share of *JAR*'s pool must fall, and it does not. Relative to controls, the share of *JAR* papers whose headline statistic falls below the sample median rises by about eleven percentage points after the mandate (*JAR* ranks first among the six accounting journals under cross-journal relabeling). Within *JAR*, the raw share moves from 46 to 54 percent of the measurable pool, and the count of weak papers is flat at roughly eleven per year (33 pre, 32 post).[10] If anything, the mandate leaves *JAR* carrying more thin-evidence papers, not fewer, so the hedging decline among them reads as authors writing differently, not different authors writing. Put differently, once readers can check the evidence, the writing response is concentrated exactly where the incentive to overstate thin evidence is strongest.

5.2 Behavior or composition? Three tests that separate the margins

Any of the behavioral findings so far could, in principle, be produced by composition alone, the same reduced-form change would appear if the mandate merely re-sorted who publishes in *JAR*. We use three designs to separate these two margins. First, we exploit within-author variation, which holds the researcher fixed, so any remaining effect is the same person writing differently. Second, we reproduce each test on an independent mandate at a second treated

[10]This split requires a numeric headline statistic, and the share of *JAR* papers carrying one falls by about 15 percentage points relative to controls as designs whose headline is not a t-statistic enter alongside the rare-topic broadening of Section **??**, so we read the weak-share rise as a statement about the measurable pool. Selection could still operate through entry (i.e., weak papers arriving after the mandate may be written by authors who never hedged), but that margin is what the within-author evidence of Section **??** tackles.

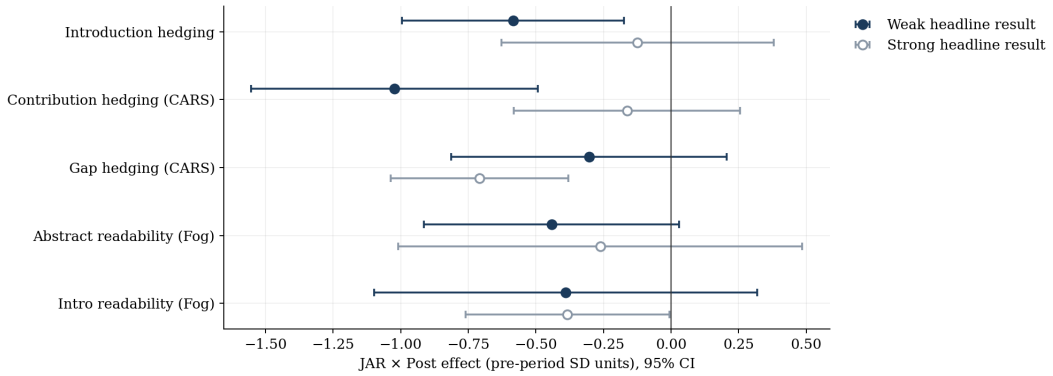

Figure 6: Writing effects by headline-result strength. Net $JAR \times Post$ effects for papers whose headline test statistic is below (filled) versus above (hollow) the sample median, in pre-period standard-deviation units with 95% confidence intervals (analytic, clustered journal×year). Estimates from the triple-interaction specification of Appendix Table **??**; weak-paper net effects are delta-method sums of $JAR \times Post$ and the triple interaction.

journal. Finally, for the text outcomes, we use theory papers, which carry no data or code and are therefore exempt from the mandate, as a within-field placebo. Findings that survive all three designs we read as behavioral, while findings that do not, we attribute to the mandate re-sorting the pool of papers.

5.2.1 Within-author estimates and robustness to selection on observables

The first and most demanding test holds the researcher fixed. Asking, how much of the measured change in *JAR*'s output is a change in behavior, and how much is it a change in the types of authors that publish in *JAR*? Two complementary exercises bound the answer, and together they localize the mechanism: first we reweight the controls to look like *JAR* on observables, then we utilize author fixed effects.

We first ask whether the reduced-form effects survive by making the control group look like *JAR* on observables. Using entropy balancing (**?**), we reweight the control journals to have the same average author h-index, seniority, team size, paper length, and top-department affiliation with those of *JAR*, and re-estimate specification (**??**) on the reweighted sample. The estimates are essentially unchanged (Table **??**, Panel A), with introduction hedging falling by 0.029, abstract complexity by roughly 1.19 Fog points, and novelty rises by 0.013, each with the same sign and magnitude as the baseline. This suggests that the behavioral results are robust to selection on the observable characteristics of authors that we can measure.

Next, we go on to absorb the author identity directly. Expanding the data to author–paper observations and adding author fixed effects, we identify β only from within-author variation, which means the same researcher publishing in *JAR* after the mandate relative to their other

work. The within-author estimates collapse to zero (Table **??**, Panel B). For example, the introduction-hedging coefficient is −0.002, abstract complexity +0.08, and novelty +0.003, and the same holds in the subsample of "switcher" authors who publish in both treated and control journals.

Together, the two tests localize the margin of selection. The cross-sectional effects survive balancing on observables but vanish once author identity is held fixed, which means that the relevant selection is not on the handful of covariates a referee would think to control for, but on the identity, and the largely unobservable type, of the authors who choose *JAR* after the mandate. This is the selection channel (prediction G1) and is the most important qualification for a purely behavioral reading of the reduced form estimates. One caveat cuts the other way, the within-author contrast nets out any response an author carries to all their outlets, so if the mandate changed writing habits portfolio-wide, this design would miss it, the null bounds the *JAR*-specific behavioral component, not the total. Additionally, we caution that the within-author design is demanding in a single small treated journal, only about 810 authors contribute repeat observations and 227 are switchers. The minimum effect the within-author estimator can detect at conventional (80 percent) power ranges from 27 to 37 percent of an outcome's standard deviation across the four columns. For abstract readability and novelty, that detection threshold lies below the cross-sectional estimate, so the within-author null is informative in the sense that the behavioral component is genuinely smaller than the cross-section. For hedging and citations, the threshold is at or above the cross-sectional estimate, so there the within-author null reflects limited power as much as composition.

Finally, the cross-sectional effects are broadly shared, not concentrated. Interacting $JAR \times Post$ with an indicator for experimental (rather than archival) work, and separately with an indicator for junior authors, yields no statistically significant heterogeneity in any channel (Table **??**, Panel C), consistent with a broad shift not confined to any one methodology or career stage.

5.2.2 A second treated journal and a dose-response design

The biggest limitation of the design is that we only have one journal treated. To deal with this we examine *Management Science*'s adoption of a stronger data- and code-sharing policy in 2019 (they required the full analysis and table code, log files, and, from 2020, an active data editor that re-executes submissions), giving us a second treated journal. Treating its accounting papers as a second treated unit in a two-way fixed-effects design (journal and submission-year effects), readability improves by 0.63 Fog points after 2019 as reported in Table **??**. This translates into about three-fifths of a U.S. grade level, the same sign and about

half of *JAR*'s magnitude. As we show below, the right comparison is to *JAR*'s behavioral component rather than its full cross-section, and there the match is close.

That an independent journal with a much stronger rule moves the same outcome the same way is a strong external check to our readability inferences. The second journal also supplies the within-author test that *JAR*'s thin panel cannot power. Absorbing author fixed effects alongside the journal's own style, the same authors write their post-2019 *Management Science* abstracts 0.64 Fog points more plainly than their other work, with the main text moving by −0.46, these within-author magnitudes match the journal-level estimate. However, their hedging, as at *JAR*, does not fall.[11] Our behavioral interpretation of readability therefore does not rest on cross-author variation at either treated journal. The hedging, novelty, and citation channels are *JAR*-specific. Next, we use the full cross-journal variation in policy stringency. We code every journal-period on an eight-item index that counts what each policy actually requires—sample-construction code, analysis code, table-generation code, a comprehensive log file, a prohibition on manual data steps, mandatory sample identifiers, active data-editor verification, and a disclosure document at submission—read directly from the policy texts (Appendix **??** defines the index and documents the coding). Within our window the index runs from 0 at the never-mandating journals, to 2 at *JAR* from 2015 (and at *CAR*, which adopted a comparable disclosure rule in mid-2020), to 5 and then 6 at *Management Science*, whose 2019 policy requires the full code chain and log files and whose 2020 revision adds the data editor. Regressing each outcome on the index with journal and submission-year fixed effects, readability scales with the strength of the rule: a one-unit increase in the index lowers Fog by 0.15 points, statistically significant whether we cluster on journal×year or on the six journals, and similar under an ordinal recoding of the same record. Because only two journals mandate within the window, the index's variation is concentrated: the estimate keeps its sign when any single journal is dropped but loses precision when either treated journal is excluded, so we read the dose-response as corroboration for the second-journal design rather than stand-alone evidence. No other outcome shows a similar dose-response.

Across both extensions, readability is the robust, generalizable result, while the remaining channels are particular to *JAR*. Additionally, *JAR*'s informative within-author null caps the

[11] Author–paper panel across the six accounting journals, 2008–2021 submissions: 7,747 observations on 1,791 repeat authors, with author, submission-year, and topic fixed effects, the *Management Science* journal effect, and the controls of equation (**??**); standard errors clustered by author. The abstract estimate is −0.75 restricting to the 2016–2021 window around the policy, and is unchanged controlling for $JAR \times Post$. The journal-level hedging coefficient in Table **??** is positive (+0.026) and significant but does not read as a policy response: an event study shows *MS*–control deviations of the same size well before the rule (2012, 2013, and 2017), the coefficient falls to +0.015 over the full 2008–2021 window and to +0.010 without the 2020–2021 submission years, no weak-paper gradient accompanies it (the triple interaction is insignificant), and relabeling a control journal with the same 2019 date produces a drift of comparable size. We therefore read the *MS* evidence as the absence of a hedging response, not a perverse one.

behavioral component below the -1.25 cross-section, and *MS*'s within-author estimates put that component near -0.6. We interpret this as the remainder reflecting the sorting that Section **??** documents directly, which suggests that the headline readability effect is roughly equal parts behavior and sorting.

5.2.3 A within-journal placebo: theory papers

Given that the mandate requires authors to file datasheets and post the code that builds their samples, it can only bind on empirical work. Theory papers in *JAR* are therefore exposed to the same editors, journal style, and time trends as their empirical counterparts but are not treated by the disclosure rule. If the readability effect is driven by the mandate rather than a journal-wide drift, it should appear for empirical papers and be absent for theory papers.

To test this, we reintroduce the analytical papers we otherwise exclude and estimate the difference-in-differences separately by paper type, and a triple difference that interacts $JAR \times Post$ with an indicator for empirical work. The separation is clean, as shown in Figure **??**. For empirical papers, abstract readability improves by 1.25 Fog points, about one and a quarter U.S. grade levels, the introduction by 0.96, and the main text by 0.53. For theory papers, the same coefficients are near zero and statistically indistinguishable from it. The triple-difference coefficients carry the right sign and a magnitude close to the empirical-only effect, and *JAR* is the most extreme of the nine journals when each is relabeled as treated.

Despite the results, we are cautious in our interpretation here since *JAR* publishes only about two dozen theory papers per period, so the triple difference is estimated imprecisely, its standard error is several times that of the empirical-only estimate, and the design could not reject even an empirical-sized effect most of the time. We therefore read the theory papers as corroborating the disclosure interpretation of the readability result, alongside the within-author and second-journal evidence, rather than as an independent test. The contrast does not extend to hedging: empirical and theory papers hedge less by nearly the same amount, which points to a journal-wide change in style, not a response to the mandate, and reinforces our reading of readability, not hedging, as the behavioral disclosure channel.

5.3 The market margin: compositional shift in authors and works

We next turn to the market margin asking whether the mandate changed which authors and which projects *JAR* attracts, and how the resulting papers are cited. These are selection effects rather than changes in any one author's conduct, given our results above. We specifically examine the novelty and topic composition of the work first, then the composition of the

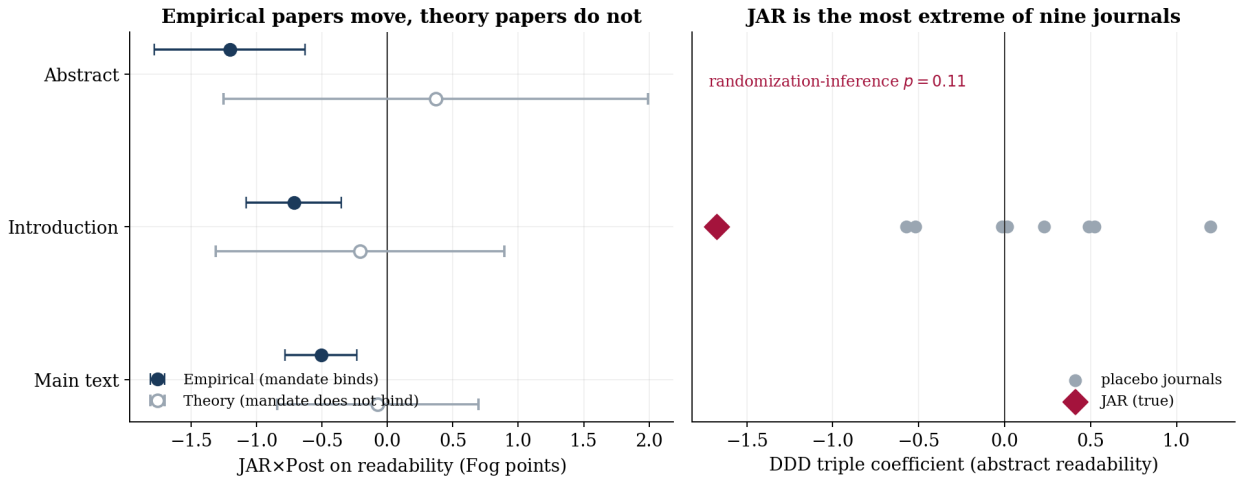

Figure 7: The mandate moves empirical papers, not theory papers. Left: $JAR \times Post$ effect on readability (Gunning–Fog) estimated separately for empirical papers (the mandate binds) and analytical/theory papers (it does not); 95% confidence intervals. Right: the abstract-readability triple-difference coefficient with each of the nine journals relabeled as treated; *JAR* (diamond) is the most extreme.

author pool, then citations, and we close by testing the assumption that the control journals are themselves untreated, which matters for our inferences.

5.3.1 Novelty: papers move further from the prior literature

We first ask whether the work *JAR* publishes moves away from the prior literature. The framework predicts it should, if a marginal empirical result is harder to defend once the data trail and code are open, the authors who continue to choose *JAR* compete on a different margin, and the natural one is the novelty of the idea. Table **??**, Panel B reports the estimates. The embedding distance of the introduction from its twenty nearest predecessors rises by 0.013, roughly a third of a standard deviation, and abstract novelty by 0.016. The share of papers that explicitly assert novelty rises by about five percentage points but is not statistically significant. We interpret that rhetorical margin apart from the distance measures, since the two turn out to be close to uncorrelated.

The distance result is the sharpest in the paper.[12] It is robust to design-based inference, randomization and synthetic control Appendix **??**), and it is the one outcome on which *JAR* separates cleanly from the placebo journals. Holding the author fixed, however, the novelty change is essentially zero (Section **??**), the mandate does not seem to make a given author

[12]Because we examine many outcomes across the four channels, some individually statistically significant coefficients are expected by chance, adjusting for multiple testing within each channel (**??**) leaves novelty as the channel that survives.

write more original papers, it seems to sort more original work into the journal. Moreover, the novelty result does not hinge on how we measure novelty. It holds across every embedding variant we construct, nearest-neighbor distances at k from 1 to 50, a five-year reference window, and an originality measure, each landing between one-third and one-half of a standard deviation, and it reappears in measures built entirely from the citation network, with no text input. Post-mandate *JAR* papers draw on a wider range of fields, and the non-accounting share of their references rises by about half a standard deviation ($t \approx 7$), robust to how many references they carry. The fact that these two different measures, constructed from different data, move together (Figure **??**; Appendix **??**), makes it less likely that our inferences are driven by some data issue.

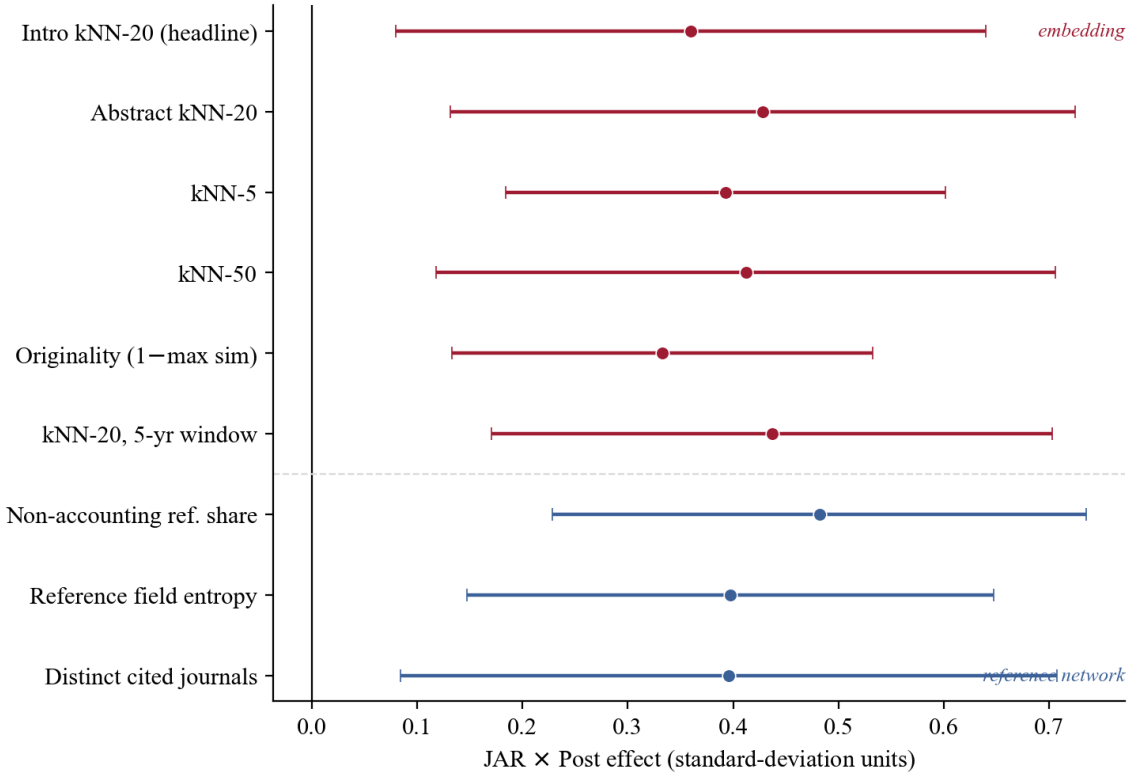

Figure 8: Novelty is robust across measures. *JAR* × *Post* difference-in-differences on each novelty measure, in standard-deviation units, with 95% confidence intervals (analytic clustering by journal×year; All-Accounting sample). Crimson markers are text-embedding measures; blue markers are built from the reference network and use no text. The effect is stable across every embedding choice and reappears in the citation-based measures. The figure shows the measures that move significantly toward greater novelty; the distance-to-centroid, reference-age, and Uzzi journal-pair measures appear in Table **??**.

Two features of the change in novelty help characterize the kind of novelty the mandate sorts in. First, it is local rather than global, the distance from a *JAR* paper to its nearest

predecessors rises sharply, but its distance to the field's center barely moves, so the mandate pushes *JAR* into the sparser corners of the field, not into a new part of it. Second, it is concentrated in the tail. Figure **??** reports quantile treatment effects showing how the effect roughly rises from about +0.007 at the median to +0.021 at the ninetieth percentile ($t = 2.4$), and the share of *JAR* papers as novel as the pre-mandate top decile rises by about ten percentage points, directionally consistent. The mandate thickens the upper tail of the novelty distribution instead of shifting the typical paper. Both patterns are sharpest in financial accounting and auditing, the archival areas where the disclosure cost binds the most, and the most novel post-mandate papers, on subjects such as analyst facial structure and the information in product-launch tweets, sit further from the field's staple questions than the papers they displace.

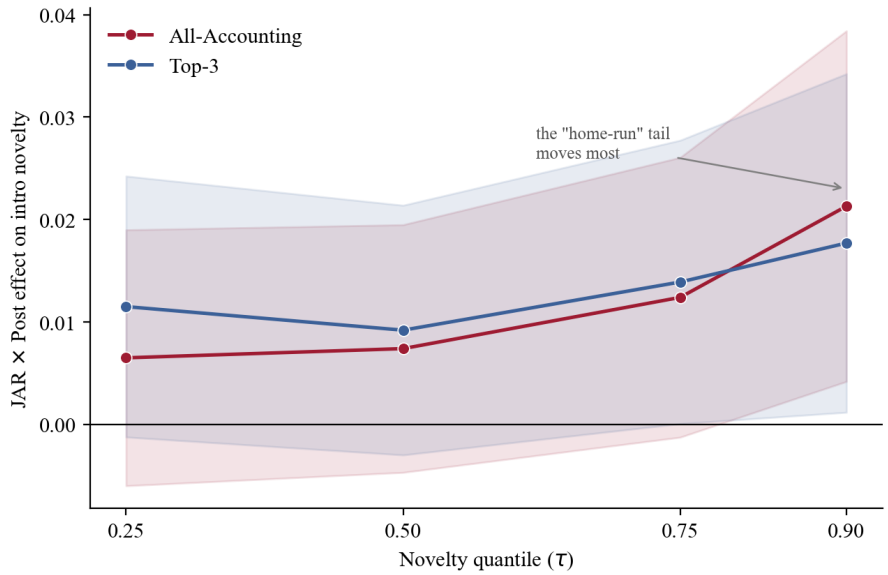

Figure 9: Quantile treatment effects on novelty. *JAR* × *Post* quantile treatment effects on introduction novelty at the 25th, 50th, 75th, and 90th percentiles (95% confidence intervals), for both comparison sets. The effect roughly doubles from the median to the 90th percentile: the mandate produces more highly novel papers rather than shifting the typical one.

A natural alternative interpretation is that *JAR* simply began publishing different topics, with the embeddings picking up the relabeling. We test this directly and find that the core topic mix does not move. Across coarse topic classes, no share shifts survive a multiple-testing correction (smallest Benjamini–Hochberg $q = 0.20$). The overall distance between *JAR*'s topic distribution and the controls' (the Jensen–Shannon divergence) barely changes, and a multinomial test of the post-mandate mix against the controls' rejects nothing. Despite these results, we further remove the issue of topical changes by re-estimating the headline novelty regression inside financial accounting alone, the dominant area holding three-quarters

of the sample, with the coefficient continuing to be positive and significant on the mandate (+0.011). The novelty gain is therefore not a relabeled topic mix, but rather papers move further from their neighbors within the field's staple areas.[13]

What does change is *JAR*'s reach into small topics. We define a paper's topic as rare if its granular OpenAlex topic holds less than one percent of the control-journal corpus (the threshold is computed omitting *JAR*, so the definition cannot respond to *JAR*'s own behavior). Figure **??** reports the result. The share of *JAR* papers in rare topics jumps from 4.8 to 19.2 percent at the mandate while the controls drift from 8.3 to 10.9, a difference-in-differences of +12.9 percentage points. The change starts exactly at the 2015 submission cohort and persists through 2021. Under cross-journal relabeling, *JAR* ranks first among all six journals, the strongest statement the design permits, with a placebo coefficient 1.65 times the largest control's in absolute value.[14] On a nearly constant paper count, *JAR* goes from publishing in 11 distinct granular topics to 20. These papers read as genuine boundary-crossing rather than simply classifier noise, touching on topics such as machine-learning fraud detection, deception in CEO narratives, judicial ideology as litigation risk, integrity-hotline field work, and the politics of accounting regulation.[15] Read together with the stable core, the composition evidence suggests that the mandate did not change what *JAR* is about, rather it widened how far afield *JAR* reaches. This is the market-margin counterpart to the thicker novelty tail above (Appendix **??** reports the full topic-share and per-topic difference-in-differences evidence).

5.3.2 Selection: the composition of *JAR*'s author and project pool

Prediction G1 of our framework says that the mandate re-sorts those who publish in *JAR*. Table **??** and Figure **??** report estimates of difference differences with the author and the characteristics of the paper as outcomes. Relative to controls, post-mandate *JAR* authors are about 0.32 standard deviations more likely to be affiliated with a top-20 research department and carry a higher h-index (about 0.25 standard deviations). They are also directionally more junior and write somewhat shorter papers, though neither margin is statistically significant

[13] If we exclude the rare-topic papers described next, this roughly halves the novelty coefficient. Tail-opening into new topics and within-topic distance are two faces of the same phenomenon, embedding novelty is higher in thin neighborhoods precisely because entering them is novel, so we describe the effect as both-and, not as purely within-topic.

[14] The estimate survives rarity thresholds from 0.5 to 2 percent, leave-one-year-out estimation, and defining rarity from pre-period control shares alone.

[15] Two qualifiers are worth noting. *JAR*'s rare-topic share is volatile further back (15 percent in 2008, zero in 2009–10), so the 2012–14 base is on the low side of its history; the post-mandate level is, however, the first stretch of seven consecutive years above 16 percent. And the rare bucket is a residual of small topics, so we lean on the out-of-*JAR* definition, the threshold sweep, and a continuous rarity version (log inverse control share, +0.56), all concordant.

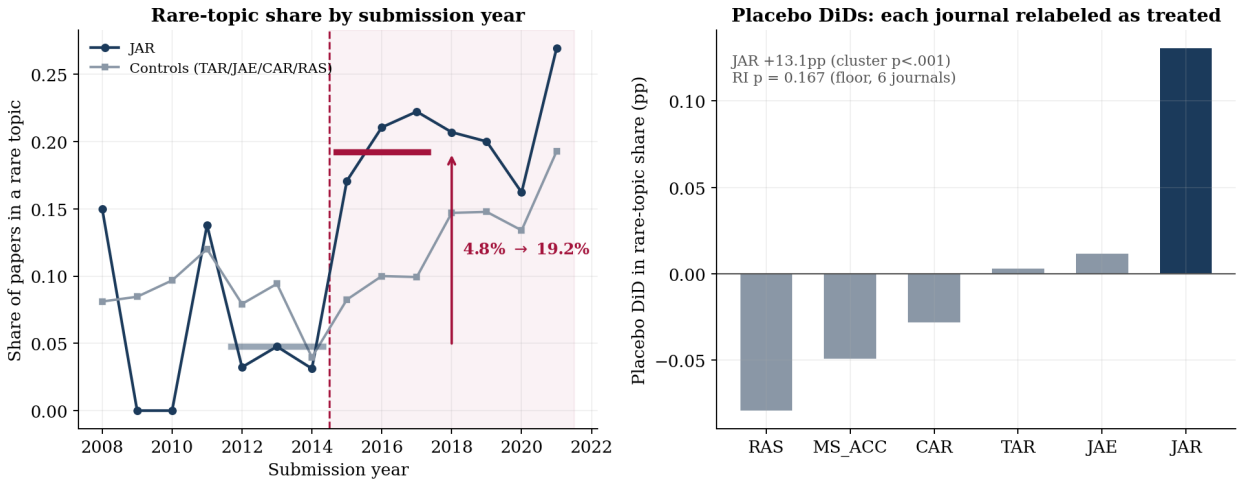

Figure 10: Rare-topic share around the 2015 mandate. A topic is rare if it holds less than one percent of the control-journal corpus (defined out of *JAR*). Left: yearly rare-topic shares, 2008–2021; bars mark the pooled 2012–14 and 2015–17 shares in the estimation sample (4.8% → 19.2%). Right: the difference-in-differences with each journal relabeled as treated; *JAR* ranks first among the six (RI at the 1/6 floor).

($p \approx 0.20$). The change in what *JAR* publishes therefore reflects, in part, a change in who publishes there. Oster-style bounds confirm what the balancing exercise in Section **??** showed, the coefficients are stable to observables, so the relevant selection is on author type (**?**).

The sorting extends to the data used in the projects. The share of *JAR* papers built on proprietary data rises by about eleven percentage points relative to controls after the mandate (a difference-in-differences of 0.11), consistent with proprietary data authors, who can substitute a documented data description for the data itself, facing a smaller compliance cost than authors of public-data work. The rise is a fact about provenance, not opacity, given that when measured with the continuous opacity index of Section **??**, the composition of the pool does not move (+0.05 standard deviations).[16] Read together, the two results indicate that the compliance cost fell on documentation, not on proprietariness, and we find no evidence of substitution between proprietary or opaque work over our window.

The composition shift has two distinct sources, and separating them characterizes how the market re-sorts the papers. We divided the post-mandate *JAR* authors into those returning to the journal and those publishing in it for the first time (Figure **??**), and the two groups pull the composition in different directions. The new juniors come from the entrants,

[16]The topic-composition evidence of Section **??** corroborates this null from the other direction. The one composition margin that does move, entry into rare topics, tilts toward more opaque, more proprietary subjects (pre-period opacity 0.46 versus 0.31, proprietary-data share 46 versus 21 percent), and the opacity shift it mechanically implies (+0.019 on the index) is almost exactly the insignificant sorting estimate here (+0.017, estimated on the wider ±7-year window). If the mandate were deflecting opaque work away from *JAR*, the entering topics would tilt the other way.

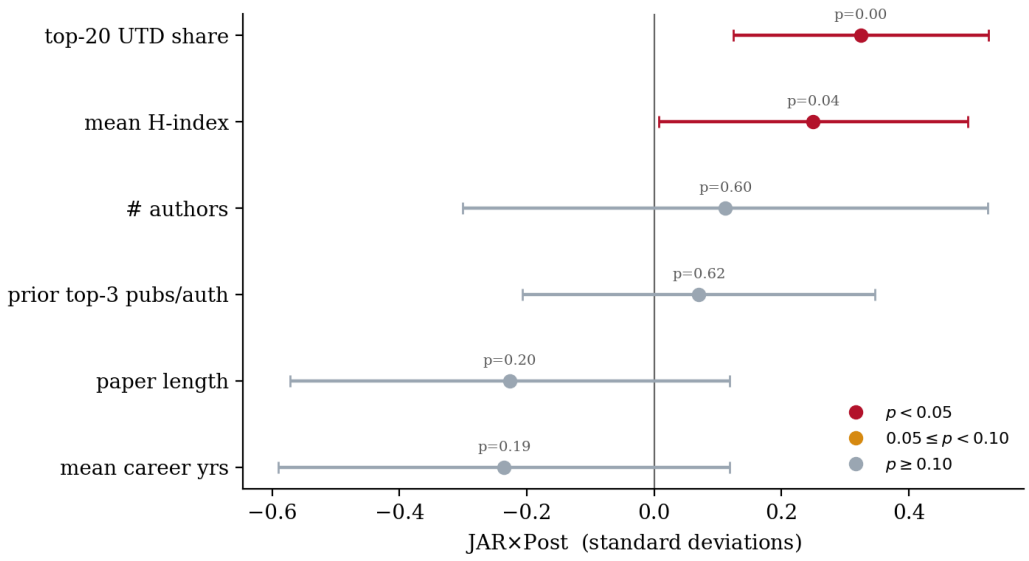

Figure 11: Selection: the composition of *JAR*'s author and project pool. $JAR \times Post$ difference-in-differences on each author/paper characteristic, expressed in standard-deviation units, with 95% confidence intervals (analytic clustering by journal×year). Markers are coloured by significance (red $p < 0.05$, orange $0.05 \leq p < 0.10$, grey $p \geq 0.10$) and annotated with the two-sided p-value. After the mandate *JAR* tilts toward more top-20-department, higher-h-index authors, with directionally (but no longer significantly) more junior authors and shorter papers.

authors new to *JAR* after the mandate are about 2.8 years more junior than the returning pool ($t = -6.4$), carry a lower h-index (14 versus 25), and are cited less than a third as often. The elite affiliation comes from the incumbents, with returning authors sitting at a top-20 department 66% of the time, against 47% for the new entrants. Established authors anchor the journal while a broad set of junior researchers enters alongside them, suggesting a widening of the author pool at the bottom of the seniority ladder rather than a wholesale turnover, drawn from many institutions, not a few (an institutional Herfindahl of 0.010, against 0.025 among returning authors). We looked for, but do not find, a *JAR*-specific change in team size or in the within-team spread of seniority, post-mandate *JAR* teams are no more mixed in seniority than the controls', and the raw rise in mixed-seniority teams is a field-wide shift, not a response to the mandate. The one team-level margin that does move is geographic, post-mandate *JAR* papers draw on fewer distinct institutions per team (a difference-in-differences of −0.19 standard deviations, which is robust to design-based inference), so the disclosed projects are somewhat more co-located, consistent with the added coordination that assembling and sharing a documented data package requires.

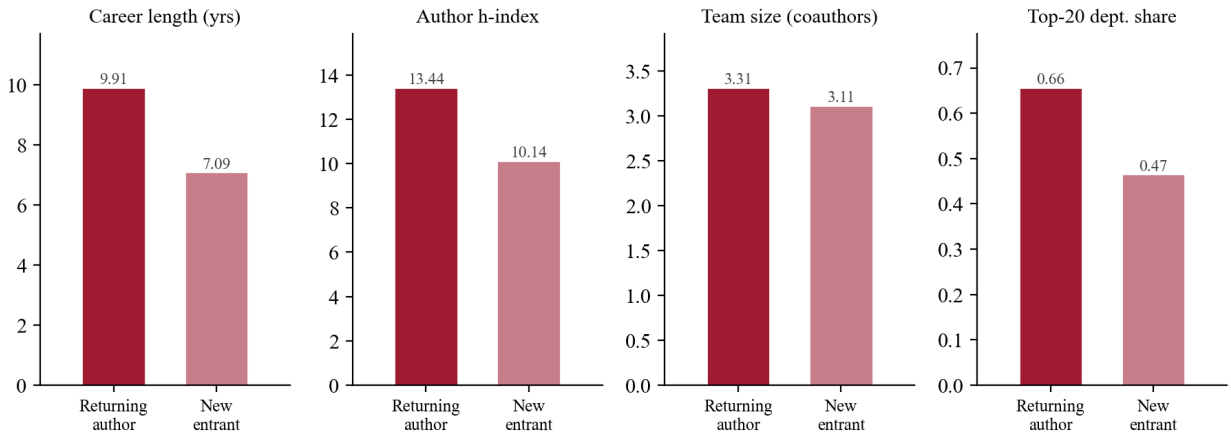

Figure 12: Characteristics of post-mandate entrants and returning authors. Mean characteristics of post-mandate *JAR* authors, split into those returning to the journal and those appearing for the first time. New entrants are more junior and less often at a top-20 department, on slightly smaller teams; the elite affiliation comes from the returning authors.

5.3.3 Citations: fewer, but from better journals

We next examine how the market receives the post-mandate papers. Table **??** reports the estimates. Raw citations fall (Panel A), with log total citations dropping by 0.29 against the top-3 controls, roughly a quarter fewer citations, and the field-weighted index moves with it.[17] The decline is not just reflecting the mechanical price of the journal's rare-topic reach, excluding the rare-topic entrants leaves it slightly larger, and the rare-topic papers themselves are, if anything, cited more than the rest of the post-mandate pool (Section **??** reports this decomposition).

However, the citations that remain come from higher-status outlets. Panel B traces the citing journals across nested tiers, the share of a paper's citations from *TAR* and *JAE* rises by 1.8–1.9 percentage points, the share from the top-six accounting journals by 2.5–2.8, the share from the top finance and economics journals by 0.2–0.3, and the share from the top-thirteen business journals by 2.8–3.0. The accounting and business tiers are each statistically significant at the one-percent level, and the small finance-and-economics tier is marginal. Panel C confirms the pattern in journal-prestige scores, the mean SJR of citing outlets rises by 0.38–0.43, which is roughly a sixth of the pre-mandate mean citing-journal SJR of 2.428, and the mean SNIP by 0.15–0.16. The estimates are consistent across the two comparison sets and sharper, if anything, in the broader All-Accounting design. Table **??** bounds how far the prominence extends: recomputing the mean SJR and SNIP of citing journals after dropping the top-6 accounting journals entirely as citers, the increases shrink to about 0.18

[17] Restricting the count to citations received within three years of publication gives the same sign and a similar share of the mean: −0.19 against the top-3 controls and −0.20 against all accounting controls.

SJR and 0.09 SNIP and are no longer statistically significant. The rising prestige of citing outlets is therefore concentrated among the top accounting and business journals.

Two further features describe the shift. First, the shift is of the whole distribution of citing-journal prestige, not of its tails, the $JAR \times Post$ effect on the mean citing SJR is essentially the same, between +0.28 and +0.39, at every decile from the tenth to the ninetieth, so the papers are cited up and down the prestige ladder by more prominent outlets rather than picking up a few marquee citations. Figure **??** shows the corresponding movement over time, the *JAR*–control gap in citing prestige is stable before the mandate and widens steadily after it. Second, the attention comes from a wider set of fields, the share of *JAR*'s citations originating inside business and accounting falls by about eight percentage points after the mandate, while the share from economics and finance, and from the social and decision sciences, rises correspondingly (Appendix Figure **??**). The shift does not come from *JAR* papers citing themselves, whose share does not rise, while the prestige gain is concentrated in the top accounting and business outlets (Appendix Table **??**). The complete set of prestige and tier measures appears in Appendix Figure **??**.

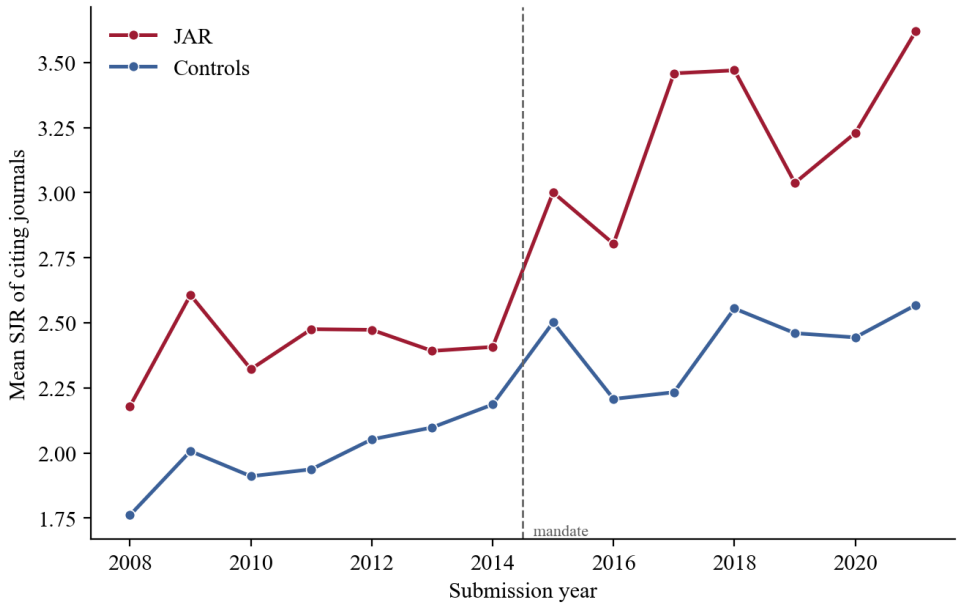

Figure 13: Post-mandate *JAR* papers are cited by more prominent journals. Mean SJR of citing journals by submission year, for *JAR* and the pooled accounting controls; the dashed line marks January 2015. The *JAR*–control gap is stable before the policy and widens after it.

5.3.4 General equilibrium: is the control group really untreated?

Every estimate so far compares *JAR* to the other accounting journals, and reads the gap between them as the effect of the mandate. That reading rests on an assumption the framework

warns against (prediction G2) that the control journals are an untreated counterfactual, unaffected by *JAR*'s policy and by whatever else is moving the field. In a market for research, neither half of that assumption is safe. Submission is a choice, so any projects diverted from *JAR* would land in *TAR*, *JAE*, and the rest, partially "treating" the controls with the displaced flow of papers. And accounting is a small, tightly connected field, so secular forces (the broader open-science movement, changing journal styles, shifting citation norms) move *JAR* and its controls together. Either force breaks the stable-unit-treatment-value assumption, but the two threats bias different comparisons. Spillovers onto the controls, like displaced submissions pushing *TAR* and *JAE* in *JAR*'s direction, would shrink the treated–control gap, so the within-accounting difference-in-differences would understate the true effect. A field-wide trend that lifts all accounting journals together is differenced out of that within-accounting contrast, what it contaminates is any estimate benchmarked outside the field, which would read the common accounting trend as a *JAR* effect and *overstate* it. Neither bias can be signed *a priori* so we work to measure them.

The first test exploits a market that *JAR*'s rule plausibly does not affect. A data- and code-sharing requirement at an accounting journal does not divert finance submissions, and the three leading finance journals (*Journal of Finance*, *Journal of Financial Economics*, *Review of Financial Studies*) adopted no comparable mandate over our window. We therefore use their accounting-related papers as a *never-treated outer benchmark* and run two contrasts for each writing outcome (Table **??**; Figure **??**). We restrict the benchmark to the writing outcomes because they are the ones an outside market evaluates on equal terms, readability and hedging are computed identically for finance and accounting text, whereas citation levels and norms differ mechanically across fields, and our novelty measure is defined relative to the accounting literature's own neighborhood. A cross-field contrast on those outcomes would not compare like with like. The first contrast asks whether the controls themselves are contaminated: a difference-in-differences comparing the accounting controls (excluding *JAR*) to finance around 2015, where a nonzero coefficient means the supposedly untreated controls are in fact moving, whether from displaced submissions or from a common field trend. The second re-estimates the headline effect with finance, rather than the accounting controls, as *JAR*'s comparison group, a cleaner control immune to within-field spillovers. Reading the two together is a triple difference: the finance-benchmarked *JAR* effect, net of whatever the accounting controls were doing anyway.

The two contrasts provide a clear picture. For readability, the finance-benchmarked effect is of the same order as the within-accounting one (-1.19 Fog points, a bit more than one U.S. grade level—against -1.25 relative to the accounting controls), and the contamination contrast is small and insignificant, so the readability result is robust to whether the comparison

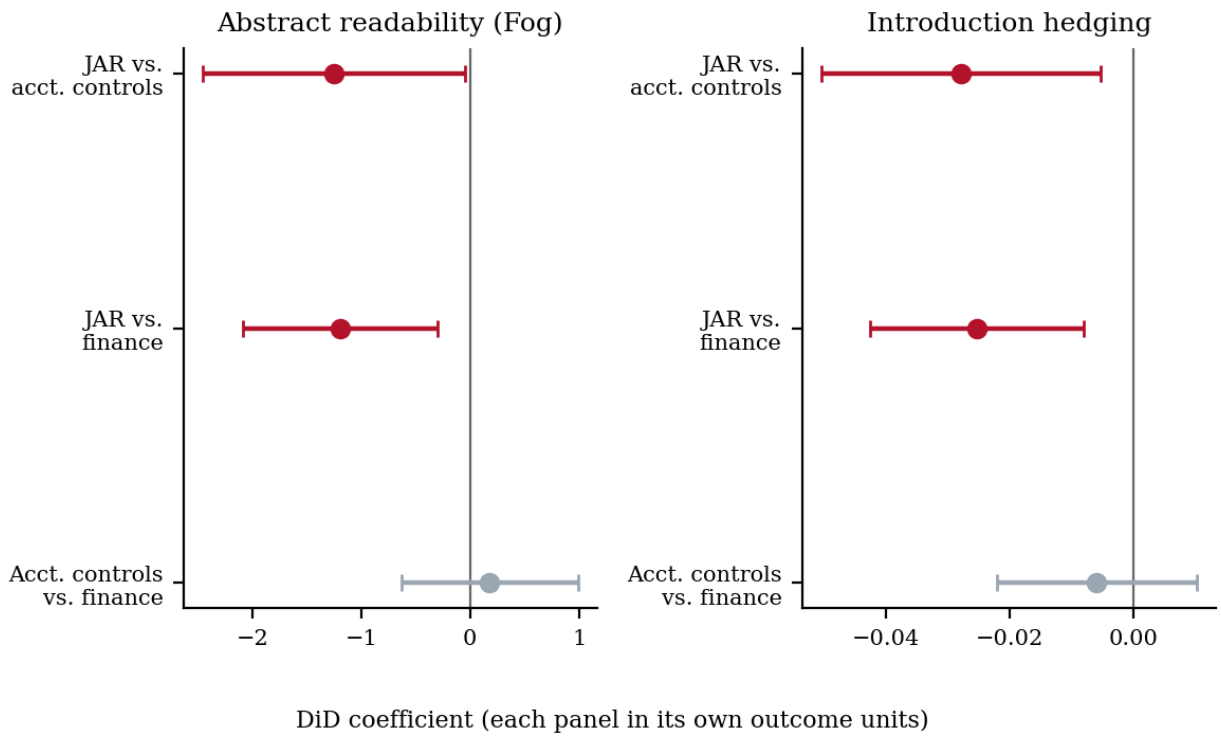

Figure 14: Benchmarking against the never-treated finance journals. One panel per writing outcome (each in its own units). Within each panel: the headline $JAR \times Post$ effect against the accounting controls (top), the same effect against the never-treated finance benchmark (middle), and the contamination contrast—the accounting controls themselves versus finance (bottom). Bars are 95% confidence intervals, coloured by significance (red $p < 0.05$, orange $0.05 \leq p < 0.10$, grey $p \geq 0.10$). For readability the finance-benchmarked effect is of the same order as the within-accounting one and the contamination contrast is small and insignificant, so the result is robust to the choice of control group; for hedging the contamination contrast sits on zero, so the controls are clean.

group is other accounting journals or an outside benchmark. For hedging, the contamination contrast is essentially zero (coefficient ≈ 0), so whatever hedging difference exists is not a control-group artifact. The lesson generalizes beyond this policy: when the treated and control units trade in the same market, a two-group difference-in-differences can confuse a field-wide trend with a treatment effect, and an outside, never-treated benchmark is what tells the two apart. Here it confirms the writing result rather than overturning it.

5.4 Identification and inference

Three design questions remain once the two margins are separated, whether *JAR* and the controls were on parallel paths before the policy, whether the estimates are an artifact of imputed treatment timing, and whether an editorial change that happened to coincide with the mandate could produce the same results. We take them in turn.[18]

5.4.1 Parallel trends

The headline difference-in-differences assume *JAR* and its controls were moving together before the policy. Figure **??** reports event-study versions of Equation (**??**), replacing *Post* with submission-year-relative-to-2015 indicators interacted with *JAR* (reference year -1). For embedding novelty, the pre-policy coefficients are flat and statistically indistinguishable from zero, and the effect appears only after 2015. For abstract readability the pre-trend is less clean, a negative pre-period coefficient two years before the policy indicates that *JAR* abstracts were already trending toward lower complexity. Total citations display a pre-existing downward drift that is partly mechanical, since older papers have had longer to accumulate citations. We therefore read the novelty result as cleanly identified. For readability, the support that does not lean on the pre-period comparison is the within-author and placebo evidence of Section **??**. For citations, the compositional (share-based) facts of Section **??** and the timing check below.

5.4.2 Robustness to imputed treatment timing

Because we assign treatment at submission and submission dates are imputed for some control papers, a natural concern is that the imputation drives the results. Table **??** shows that it does not. First, the treated unit needs no imputation as *JAR* reports a received date for every article in the window, so all imputation falls on the controls, where only around nine percent need to be imputed for *TAR+JAE* and they are balanced across the pre and post periods

[18] Design-based inference for the single-treated-cluster problem is done in Appendix **??**.

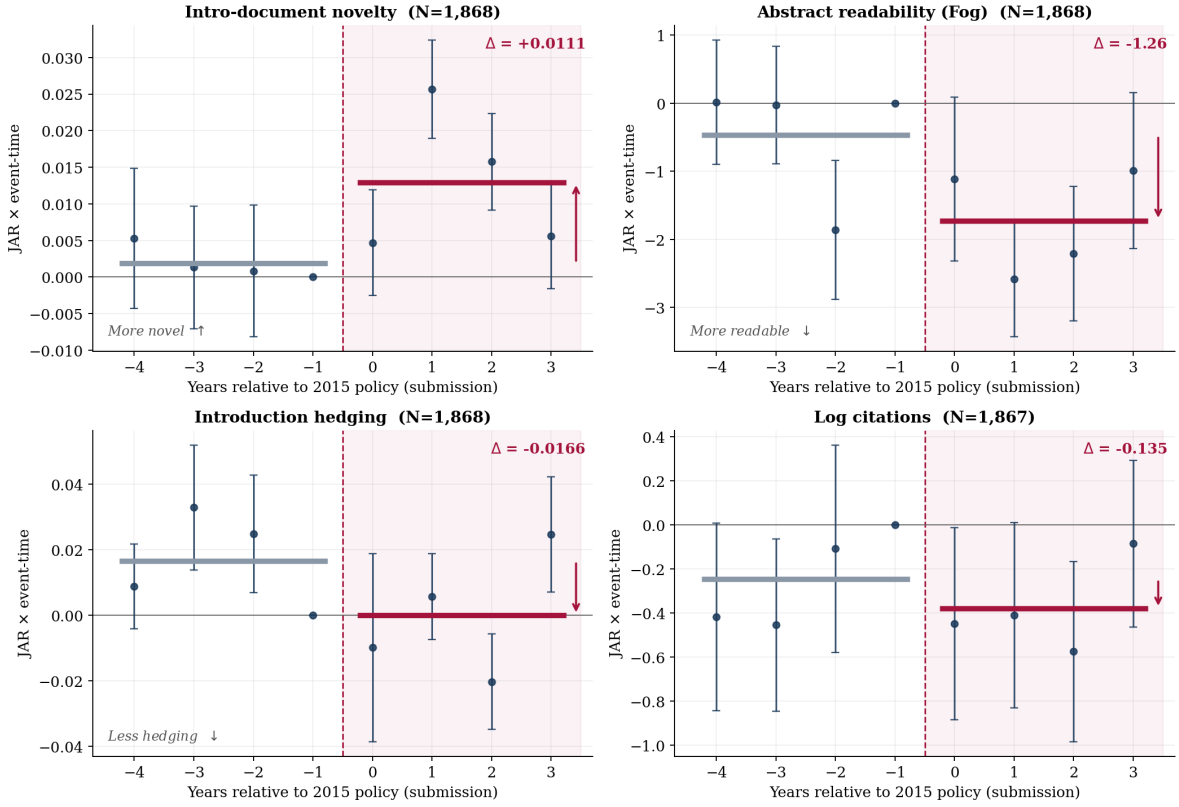

Figure 15: Event-study (dynamic DiD) estimates. Each panel plots the $JAR \times$ event-time coefficients (95% confidence intervals, analytic clustering by journal×year), reference year -1 (2014). The grey and red bars mark the mean of the pre-period and post-period coefficients, and Δ is their difference—the level shift at the mandate. Flat, near-zero pre-period coefficients support parallel trends; the readability pre-period dip at -2 and the citation drift discussed in the text are visible as departures of the pre-period points from the grey bar.

(33.1 versus 33.4 percent in the broad set). *JAR* loses exactly one paper to the ambiguous-window screen. Second, re-estimating the baseline on the observed-date-only subsample, dropping every imputed observation, leaves the abstract-readability, introduction-novelty, and introduction-hedging effects unchanged in sign and significance and, if anything, slightly larger (abstract readability moves from -1.26 to -1.32, novelty from 0.014 to 0.015), while the citation-count effect, the least precisely estimated channel, keeps its sign. This evidence suggests a mild non-differential attenuation in the full sample, not an effect manufactured by imputation, which would shrink when the imputed observations are removed. Third, the realized review lag lengthened modestly after 2015, but the *JAR*-versus-control difference in that lengthening is statistically insignificant (a difference-in-differences of roughly +72 days) and, by construction, moves only imputed control dates, never *JAR*'s. Since non-differential misclassification of the post indicator attenuates the estimate toward zero, our full-sample coefficients are, if anything, conservative.

5.4.3 Editorial change?

A final concern is that our writing and selection results may simply reflect an editor or house-style shock that happened to land at the same time as the mandate rather than the mandate itself. With a single treated journal, any *JAR*-specific shock that coincided with January 2015, an editor transition, a change in desk-reject norms, a house-style revision, is collinear with the mandate, and the field-wide year fixed effects cannot absorb it. We address the concern in two ways. First we examine changes in who edits the journal, which is directly testable. Next we examine changes in the stylistic regime imposed from the masthead, which cannot be tested head-on.

If *JAR*'s post-2015 shifts merely track a change in editors, they should disappear once the editor is held fixed. Table **??** and Figure **??** show that they do not. Three editors—Berger, Leuz, and Skinner—each handled at least five papers on both sides of the mandate and together account for the majority of the journal's decisions. Within this subsample, with editor fixed effects, the same before/after pattern reappears: abstracts get plainer by 1.1 Fog points, introductions hedge 1.7 percentage points less, novelty rises by 0.007, and review durations lengthen by roughly 20%. The readability and hedging estimates essentially match the whole-journal figures, and all four keep their sign and significance.[19] The same people, editing at the same desk, changed what they accepted once the rule bound.

[19]The one outcome that does not move within editor is data opacity, which is consistent with the paper's null result on project sorting (Section **??**). The lone exception is Skinner, whose auditing share rises from 6% to 18% over the window, his opacity change is more than twice the journal-wide figure and survives topic fixed effects on his desk alone, so it reflects a shift in the mix of papers routed to his desk rather than a change in what a given editor accepts.

A falsification test sharpens the point, *JAR*'s actual masthead turnover came in 2019, not 2015, and nothing breaks there, among post-mandate papers the 2019 break on readability, hedging, and novelty is small and insignificant (readability -0.35, $p = 0.22$), and the editors who joined in 2019 look like the continuing ones on every outcome. We are careful about the reach of this test, while it rules out that the 2015 shifts are an artifact of the editors, it does not separate the demand margin from supply, since the same editor still accepts from a changed submission pool.

The diffuse version of a change in the stylistic regime imposed from the masthead leaves no editor to hold fixed, so here we rely on the separating designs. The theory-paper placebo holds the editorial regime fixed: theory papers pass through the same editors and the same desk in the same years, yet, carrying no data or code to disclose, they do not move (Section **??**), which a stylistic change imposed by an editor would not respect. The second treated journal (Section **??**) identifies the readability response from variation that an idiosyncratic *JAR* editorial change cannot generate, a different journal, in a different field, under different editors, moves the same way after its own mandate and does so within author, the same researchers writing more plainly once a rule binds. The weak-paper relationship is a second pattern no editor-imposed style regime should produce, for the reason Section **??** gives. A coincident editorial shift could affect the two-group reduced form, but it cannot produce the convergence of these independent designs, which is why we rely on their inference for the behavioral claim rather than on the headline difference-in-differences alone.

6 Discussion

Our results document how a transparency mandate adopted by a single journal changes writing and papers, when only that journal, and not its competitors, adopts. Field-wide adoption is a different experiment, and our framework distinguishes these two. The market-margin results are relative-price effects in the sense that novelty rises, elite authors arrive, and attention concentrates at *JAR* because authors can still circumvent the disclosure cost by submitting elsewhere. If every journal adopted the rule, that escape valve closes, which would flatten the re-sorting across journals, and the composition gains we document, which came partly at the other journals' expense, would not aggregate to the whole field.

The behavioral results are the opposite case, a given author's plainer writing and fuller reporting do not depend on where the marginal paper goes, they reappear under an independent mandate at *Management Science*, and they scale with policy stringency, so these are the effects we would expect to survive field-wide adoption. What other things would universal adoption change becomes the more important question, with no untreated journal of similar

quality to absorb high-documentation-cost projects, the adjustment must run through what gets researched and how it is documented rather than where it is sent. All this to say that our reduced form is not a forecast of a field-wide mandate, it is the market's response while an outside option still exists.[20]

6.1 Is the research better?

Next comes the question of whether research quality improved. A mandate can change research without improving it, so we close by asking whether the shifts we document leave the field better off, and we try to answer it to the extent the evidence allows. The question of research quality has both a private and a social side, which may not agree. Privately, *JAR* appears to come out ahead, its post-mandate papers are plainer, more novel, and cited by more prominent outlets, the portfolio an academic journal competes to assemble. Socially, the results are more mixed, and part of the private gain is a transfer rather than a creation of value.

Start with the specification fishing, the aspect that the transparency movement cares about, where the case is strongest. Post-mandate *JAR* papers leave more of their insignificant results on the page, the share of reported nulls rises by about a sixth, and they back each headline claim with more supporting coefficients rather than a few strong ones. Whatever else the mandate did, it made the evidentiary record behind a *JAR* paper more complete and more scrutinizable, which is the stated objective of the policy and of the broader movement (**??**). Plainer prose points the same way, as a reader pays less to extract what a paper did and found.

However, these benefits come with costs, which we do not net out. The first is reach. Post-mandate *JAR* papers are cited about a quarter less often, although the citations that remain come from more prominent journals. To the extent that citations proxy for use, some use was lost, and lost on comparable work, not only on the papers that reached into sparser topics. Whether the field prefers narrower, more elite attention to broader attention is a value judgment we do not take a stand on. The second margin is the work the rule prices, and here the evidence tempers the standard worry rather than confirming it. The proprietary cost of disclosure theory that the accounting literature has studied for four decades (**??**) predicts that a sharing mandate deflects proprietary-data work to the journal's neighbors. Over our

[20]The general-equilibrium results provide a methodological lesson that extends beyond this setting. The natural control group for a single-journal mandate, the other journals in the field, is endogenous to the very behavior under study, both because submissions are reallocated and because the field moves together. The robustness of our estimates to a never-treated outside benchmark had to be verified, not assumed (Section **??**). Studies that evaluate journal- or institution-level policies should benchmark against a plausibly untreated outside market wherever one exists.

window it did not, because the 2015 rule asked for a datasheet and sample-construction code rather than the data themselves, proprietary-data authors could comply at low marginal cost, as proxied by the rise in the share of proprietary-data papers in *JAR* while its measured opacity does not move (Sections **??** and **??**). The redistribution we can document runs through authors instead: the elite researchers who sort into *JAR* would otherwise have published elsewhere in the field, so part of *JAR*'s gain is a within-field transfer, not new value. The deflection cost remains the right worry for stronger policies, here *Management Science*'s verification regime and the full analysis code required by *JAR*'s own 2024 revision impose the cost on the work the 2015 rule spared. Moreover, a single-journal design cannot measure the net effect in any case, because we never observe the projects of the authors who chose not to come.

Two forces skew even the reporting gains as pure welfare. The cleanest behavioral gain, more complete reporting, is *JAR*-specific and does not reappear at *Management Science*, so we cannot claim it as a result of the disclosure mandate. And a rule strong enough to push proprietary-data work elsewhere would trade one kind of credibility for another. For example, results that anyone can re-run, in exchange for questions only proprietary data can answer. The 2015 mandate largely avoided that trade-off, the datasheet substituted for the data, but its stricter successors do not, and whether the trade is worth making is itself the welfare question, which we are not able to answer with this data.

7 Conclusion

We studied how a data- and code-sharing mandate requiring authors to disclose the provenance of their data and hand over the code that builds their samples changes the research a journal publishes, using *JAR*'s January 2015 policy and a corpus of roughly 4,600 accounting articles. The reduced-form patterns are economically meaningful. Post-mandate *JAR* papers are written about one Gunning–Fog grade more plainly, hedge less in the parts of the paper that state the contribution, especially where the headline evidence is weakest, score higher on every distance- and breadth-based novelty measure, leave more of their insignificant results on the page, and draw a two- to three-percentage-point larger share of their citations from leading journals. Because treatment is assigned to a single journal, we held each result to design-based inference and to explicit tests for the general-equilibrium responses a single-journal mandate provokes. The novelty effect clears cross-journal randomization inference and a synthetic control and is corroborated by reference-network measures that use no text. Our readability effect is corroborated by a second treated journal and a stringency dose-response, while the hedging and citation effects are large and consistently signed. We also document direct

evidence that the mandate re-sorted *JAR*'s author pool, toward more elite and slightly junior authors.

Our work shows how transparency requirements do not merely make existing research checkable; they change what research is written, by whom, and where it is sent. For editors and scholarly associations, the takeaway is that a disclosure rule is a market intervention whose effects run through the composition of submissions as much as the conduct of any one author, and that field-wide adoption can manufacture trends that any single journal's policy will appear to cause. We hope our work encourages future research to ask whether the same forces operate under registries and pre-analysis plans, and in settings with enough treated journals to identify the market response directly.

Tables

Table 1: Descriptive statistics

	N	Mean	SD	Median
Outcomes				
Abstract readability (Fog)	1,370	19.29	2.98	19.21
Introduction readability (Fog)	1,370	17.82	1.73	17.76
Introduction hedging rate	1,370	0.506	0.097	0.511
Intro-document novelty (kNN-20)	1,370	0.532	0.035	0.531
Log total citations, $\log(1 + c)$	1,370	4.16	1.13	4.20
Cite share from top-6 acct.	1,356	0.118	0.083	0.106
Controls				
Paper length (words, 000s)	1,370	10.7	2.0	10.5
Number of authors	1,369	2.75	0.99	3.00
Mean author h-index	1,369	9.5	5.2	8.8
Mean career length (years)	1,356	7.8	5.2	7.0
Top-20 department share	1,370	0.384	0.487	0.000

Notes. Sample of archival and experimental research articles in the six accounting journals, submission years 2012–2017 (the headline DiD window). Readability is the Gunning Fog index (higher = more complex); novelty is mean cosine distance to the 20 nearest prior-three-year documents; hedging is the share of uncertainty-marked sentences.

Table 2: Main difference-in-differences estimates, by channel

Outcome ($JAR \times Post$)	Top-3 coef.	All-Acct coef.	Pred.
Panel A. Writing (Gunning–Fog; hedging rate)			
Abstract readability (Fog)	-1.352^{***}	-1.250^{**}	P1
Introduction readability (Fog)	-1.079^{***}	-0.959^{**}	P1
Conclusions readability (Fog)	-1.006^{***}	-0.823^{**}	P1
Introduction hedging rate	-0.027^{**}	-0.028^{**}	P2
Contribution hedging rate	-0.094^{***}	-0.091^{***}	P2
Panel B. Novelty			
Intro-document novelty (kNN-20)	$+0.014^{***}$	$+0.013^{**}$	P4
Abstract novelty (kNN-20)	$+0.020^{***}$	$+0.016^{***}$	P4
Novelty claim (indicator)	$+0.045$	$+0.049$	P4
Panel C. Reported inference			
Share insignificant ($\lvert z\rvert < 1.65$), all results	$+0.039^{**}$	$+0.039^{*}$	P3
Panel D. Citations			
Log total citations	-0.287^{**}	-0.277^{*}	P5
Log field-weighted citations	-0.213^{*}	-0.221	P5
Cite share from top-6 accounting	$+0.025^{***}$	$+0.028^{***}$	P5
Mean prestige of citing outlets	$+0.425^{***}$	$+0.382^{*}$	P5

Notes. Each cell is the $JAR \times Post$ coefficient from specification (**??**) with submission-year and topic fixed effects and paper/author controls. "Top-3" compares JAR to $TAR{+}JAE$; "All-Acct" compares JAR to all five accounting control journals. Stars denote analytic clustered significance and are descriptive only: Appendix **??** reports design-based inference appropriate for a single treated cluster. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Mechanism: test multiplicity (discipline vs. selection)

Outcome ($JAR \times Post$)	Top-3	All-Acct
Log # headline (primary) results	-0.098^{*}	-0.104^{**}
Log # reported coefficients (total)	$+0.153^{**}$	$+0.101^{*}$
Log # tables	$+0.006$	-0.019
Multiplicity ratio (total / primary)	$+15.54^{***}$	$+14.87^{***}$

Notes. Each cell is the $JAR \times Post$ coefficient from specification (**??**) with the listed count as the outcome. Headline ("primary") results are coefficients flagged as a paper's main findings; the total is the full census of reported coefficients across all tables. For JAR, the mean number of primary results falls from 2.6 (pre) to 2.0 (post) while the mean total rises from ∼49 to ∼65, so each headline claim is backed by more supporting estimates after the mandate—the discipline signature. p-values are analytic and subject to the single-treated-cluster caveat of Appendix **??**. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Author fixed effects, entropy balancing, and heterogeneity

	Hedging (intro)	Readability (abstract)	Novelty (intro-doc)	Log citations
Panel A. Entropy-balanced DiD (controls reweighted to JAR)				
$JAR \times Post$	-0.029^{**}	-1.187^{*}	$+0.013^{***}$	-0.324^{*}
p-value	0.016	0.058	0.004	0.058
Panel B. Author fixed effects (within-author / switchers)				
$JAR \times Post$, all repeat authors	-0.002	$+0.080$	$+0.003$	-0.005
p-value	0.889	0.825	0.535	0.960
$JAR \times Post$, switchers only	$+0.007$	$+0.214$	$+0.003$	$+0.073$
p-value	0.619	0.564	0.449	0.530
Panel C. Heterogeneity (triple-interaction coefficient)				
$JAR \times Post \times$ experimental	-0.021	-0.467	$+0.016$	$+0.016$
p-value	0.508	0.582	0.276	0.963
$JAR \times Post \times$ junior author	-0.025	$+1.165^{**}$	$+0.006$	$+0.246$
p-value	0.236	0.041	0.352	0.374

Notes. All columns use the All-Accounting sample. *Panel A* reweights control papers by entropy-balancing weights that equate the means of author h-index, career length, team size, paper length, and top-department share with *JAR*'s (**?**); post-balancing standardized covariate imbalance is zero. *Panel B* expands to author–paper observations and absorbs author fixed effects, so the coefficient is identified from within-author variation; standard errors are clustered by author ($N \approx 2{,}518$ author–paper observations, 810 repeat authors, 227 switchers). *Panel C* reports the triple-interaction of $JAR \times Post$ with the listed indicator. p-values are analytic and, for Panels A and C, remain subject to the single-treated-cluster caveat of Appendix **??**. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: A second treated journal and a dose-response design

	Abstract readability	Introduction hedging	Novelty (intro-doc)	Log citations
Management Science × Post (2019)	-0.625^{**}	$+0.026^{***}$	-0.001	$+0.337^{**}$
Policy index (eight-item count), acct. journals	-0.151^{**}	$+0.002$	-0.001	$+0.013$

Notes. Row 1 treats *Management Science*'s accounting papers as a second treated unit (post-2019), in a two-way fixed-effects design with journal, submission-year, and topic effects plus controls ($N = 1{,}295$; control journals *TAR*, *JAE*, *RAS*). Row 2 regresses each outcome on the journal×period policy index—an eight-item count of what each policy requires, coded from the journals' policy documents (defined in Appendix **??**, Table **??**); in our window the index runs from 0 (never-mandating journals) through 2 (*JAR* from 2015; *CAR* from mid-2020) to 5–6 (*MS* from 2019, adding a data editor in 2020)—with journal, submission-year, and topic effects across the six accounting journals ($N = 3{,}480$). Readability is the only channel that generalizes across journals in the predicted direction and scales with policy stringency. *Management Science*'s positive hedging coefficient does not read as a policy response—deviations of the same size predate the 2019 rule and the estimate is fragile to the estimation window (see text)—and novelty and citations do not track the policy index, so the behavioral margins beyond readability are journal-specific. Standard errors clustered on journal×year. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: Selection: change in *JAR*'s author and paper composition

Characteristic ($JAR \times Post$)	DiD (SD units)	p-value
Top-20 department share	+0.32	0.002
Mean author h-index	+0.25	0.044
Mean author career length	−0.24	0.193
Paper length	−0.23	0.197
Number of authors	+0.11	0.595
Prior top-3 pubs. per author	+0.07	0.619

Notes. Each row is a DiD with the listed *characteristic* as the outcome, expressed in standard-deviation units of that characteristic. Post-mandate *JAR* authors are more likely to be at top-20 departments and carry a higher h-index; the shifts toward more junior authors and shorter papers are directionally similar but not statistically significant—direct evidence of the selection channel (G1).

Table 7: Citations: count, citing-journal status, and prestige

Outcome ($JAR \times Post$)	Top-3	All-Acct
Panel A. Citation count		
Log total citations	-0.287^{**}	-0.277^{*}
Log field-weighted citations	-0.213^{*}	−0.221
Panel B. Share of citations from...		
TAR + *JAE*	$+0.018^{***}$	$+0.019^{***}$
top-6 accounting journals	$+0.025^{***}$	$+0.028^{***}$
top-7 finance/economics journals	$+0.003^{*}$	+0.002
top-13 business journals	$+0.028^{***}$	$+0.030^{***}$
Panel C. Prestige of citing journals		
Mean SJR of citing outlets	$+0.425^{***}$	$+0.382^{*}$
Mean SNIP of citing outlets	$+0.160^{***}$	$+0.146^{**}$

Notes. Each cell is the $JAR \times Post$ coefficient from (**??**). Panel A uses logged total and field-weighted citation counts. Panel B reports the share of a paper's citations originating in nested journal tiers (SJR is the Scimago Journal Rank; SNIP, source-normalized impact per paper). Raw citations fall while the status of citing journals rises across every tier. Stars are analytic clustered significance and descriptive only. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8: General equilibrium: accounting controls vs. a never-treated finance benchmark

	Contamination (acct. controls vs. finance)		*JAR* vs. finance (clean control)	
Outcome	coef.	p	coef.	p
Abstract readability	+0.179	0.67	−1.192	0.009
Introduction hedging	−0.006	0.47	−0.025	0.004

Notes. "Contamination" is a DiD comparing the accounting *control* journals (excluding *JAR*) to the finance benchmark around the policy: a nonzero value means the controls are themselves moved by field-wide forces. "*JAR* vs. finance" uses finance as the (cleaner) control for *JAR*. The readability effect against the clean control (−1.19) is of the same order as the within-accounting estimate (−1.25), and the contamination contrast is small and insignificant. The result is robust to the choice of control group and hedging shows no contamination.

Online Appendix

Mandated Transparency and the Production of Research: Evidence from a Data- and Code-Sharing Policy

A The History of *JAR*'s code and data policy

This appendix documents the full history of the *JAR* data policy. Table **??** summarizes the four versions of the policy, the coding is taken directly from the policy documents posted by the journal and the accompanying FAQ.

The original "Data Policy for the *Journal of Accounting Research*," announced in December 2014, took effect for all submissions received on or after January 1, 2015. It imposed two obligations. At initial submission, authors had to file a data-description sheet disclosing how the raw data were obtained or generated—sources, download dates, the instrument for surveys and experiments—and which authors handled the data and ran the analyses; authors using proprietary data had also to give the editors, privately, a contact at the providing organization and the terms of the data-sharing agreement. Prior to final acceptance, authors had to submit the computer code that converts the raw data into the analysis dataset (or, for genuinely proprietary portions, a step-by-step description precise enough for another researcher to reconstruct the sample), and the datasheet and code were posted on the journal's website after publication. The policy covered archival, experimental, field, survey, and simulation work; purely analytical papers were exempt. It did not require the final dataset, the statistical analysis code, or the table-generation code, and compliance was self-certified—the journal appointed no data editor and ran no in-house replication.

The two subsequent updates inside and just beyond our sample window left these requirements unchanged. The November 2016 revision renamed the policy the "Data and Code Sharing Policy," added an FAQ, and extended coverage to structural papers, without altering what authors had to submit. The November 2022 revision expanded the proprietary-data disclosure: authors must now indicate whether the data provider restricts publication, requires pre-approval of results, or retains control of the data. On the feature-count index in Table **??**, all three versions score 2 of 8.

The substantive escalation came only in April 2024, well outside our window. Version 4 requires the full replication package short of the data themselves: statistical analysis code, table-generation code, a comprehensive log file showing the execution of the entire codebase, observation identifiers for the final sample (previously "whenever feasible," now mandatory), and an explicit prohibition on manual data manipulation outside the submitted

code. Even under Version 4, compliance remains self-certified: unlike *Management Science*, which appointed a data editor in 2020 to verify replication packages before publication, *JAR* has no active verification. Two implications follow for our design. First, within our sample the policy in force is the 2015 disclosure rule described in Section **??**, a single, sharp adoption date with no confounding mid-sample tightening. Second, the 2024 increase implies that our estimates speak to a comparatively light mandate, the behavioral responses we document arise from disclosure and exposure alone, not from verification.

B Additional descriptive and robustness tables

This appendix collects the descriptive and robustness tables referenced in the main text: the citing-journal prestige result excluding the top-six accounting journals (Table **??**), author-pool descriptives (Tables **??** and **??**), the reported-p-value descriptives (Table **??**), the significance-distribution difference-in-differences (Table **??**), the opacity-heterogeneity estimates (Table **??**), the writing effects split by headline-result strength (Table **??**), and the treatment-timing robustness (Table **??**).

Table OA1: The *JAR* data- and code-sharing policy across its four versions

Effective for submissions from	Version 1 Jan. 2015	Version 2 Nov. 2016	Version 3 Nov. 2022	Version 4 Apr. 2024
Panel A: Required at initial submission				
Data-description sheet (sources, download dates, who handled the data)	✓	✓	✓	✓
Provider contact and terms for proprietary data (privately, to the editors)	✓	✓	✓[a]	✓
Panel B: Required prior to final acceptance				
Sample-construction code (raw data → analysis dataset)[b]	✓	✓	✓	✓
Statistical/econometric analysis code	–	–	–	✓
Table-generation code	–	–	–	✓
Comprehensive log file of the full code execution	–	–	–	✓
Prohibition on manual (out-of-code) data manipulation	–	–	–	✓
Observation identifiers for the final sample	Encouraged	Encouraged	Encouraged	✓
Final analysis dataset	–	–	–	–
Panel C: Enforcement				
Compliance self-certified by authors	✓	✓	✓	✓
Data editor / in-house verification of packages	–	–	–	–
Feature-count stringency index (0–8)[c]	2	2	2	7

Notes. Requirements are coded from the four policy documents posted by the journal ("Data Policy for the *Journal of Accounting Research*," December 2014, and the "Data and Code Sharing Policy" as updated November 2016, November 2022, and April 2024) and the journal's FAQ dated April 1, 2024. Version 1 took effect for all submissions received on or after January 1, 2015; the November 2016 and November 2022 documents state only the month of the update, so those boundaries are approximate. A checkmark denotes a mandatory requirement; "–" denotes absence; "Encouraged" denotes language of the "whenever feasible" form. [a]Version 3's substantive change is expanded proprietary-data disclosure: authors must indicate whether the provider restricts publication, requires pre-approval of results, or retains control of the data. [b]All versions allow a detailed step-by-step description in place of code for genuinely proprietary portions of the sample-construction process. [c]Sum of eight binary indicators: the six code/data requirements in Panel B (counting identifiers only when mandatory), the submission-stage disclosure sheet, and active editorial verification. Version 2 renamed the policy the "Data and Code Sharing Policy" and extended coverage to structural papers; Version 3 left the code requirements unchanged; neither moves the index.

Table OA2: Citing-journal prestige: robustness to excluding the top-6 accounting journals

	JAR mean		*JAR* × *Post* difference-in-differences		
Outcome	Pre	Post	Within-*JAR*	vs. *TAR*+*JAE*	vs. All-Acct
Mean SJR of citing journals	2.428	3.059	+0.704***	+0.399***	+0.383*
			(+4.40)	(+4.03)	(+1.78)
excl. top-6 accounting	1.485	1.873	+0.388***	+0.203	+0.184
			(+3.04)	(+1.63)	(+0.83)
Mean SNIP of citing journals	1.864	2.124	+0.283***	+0.149***	+0.145**
			(+5.26)	(+4.78)	(+2.04)
excl. top-6 accounting	1.501	1.693	+0.196***	+0.093**	+0.088
			(+5.34)	(+2.25)	(+1.13)

Notes. Each cell is a difference-in-differences estimate of the *JAR* × *Post* effect on the mean Scopus prestige (SJR or SNIP) of the journals citing a paper in its first three years, with t-statistics in parentheses. *Within-JAR* regresses the outcome on Post in a *JAR*-only sample (topic FE, SE clustered by year); the two DiD columns add JAR × Post with topic and submission-year FE and SEs clustered by journal×year, over the 2012–2017 window. The *excl. top-6* rows recompute the mean citing prestige after dropping citers that are themselves top-6 accounting journals (*JAR*, *JAE*, *TAR*, *CAR*, *RAS*, accounting *Management Science*). Excluding them, the DiD increases attenuate and, against the accounting controls, are no longer significant, indicating the citing-prestige gain is concentrated among the top-6 accounting and adjacent business journals whose citation share rises. Controls X as in equation (**??**). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table OA3: Author and team characteristics by journal

	JAR	TAR	JAE	CAR	RAS	MS
Number of authors	2.62	2.70	2.66	2.94	2.86	2.76
Mean author h-index	8.8	9.0	9.4	9.8	8.7	11.6
Mean career length (years)	7.4	7.4	7.0	9.0	7.5	8.6
Top-20 department share	0.50	0.37	0.51	0.29	0.36	0.33
Prior top-3 pubs. per author	2.98	3.19	3.09	2.26	2.41	1.45
Papers	162	436	195	269	135	173

Notes. Means of paper- and author-level characteristics for the empirical sample (archival and experimental research articles, submission years 2012–2017). h-index, career length, and top-3 publication counts are averaged across a paper's authors; the top-20 share is the fraction of papers with at least one author at a top-20 research department.

Table OA4: Author characteristics: *JAR* versus controls, before and after the policy

	JAR		Controls		
	Pre	Post	Pre	Post	DiD
Number of authors	2.48	2.77	2.66	2.86	0.10
Mean author h-index	7.5	10.3	8.7	10.3	1.3
Mean career length (years)	7.2	7.6	7.2	8.4	-0.9
Top-20 department share	0.42	0.59	0.38	0.36	0.19
Prior top-3 pubs. per author	2.60	3.40	2.52	2.72	0.60

Notes. Group means by treatment status and period (submission before vs. on/after January 2015). The final column is the difference-in-differences (*JAR* post−pre minus Controls post−pre). Post-mandate *JAR* shifts toward more top-20-department and more junior authors; see Section **??** for inference.

Table OA5: Statistical results extracted from papers in the empirical sample

	N	Mean	SD	Q1	Median	Q3
All results						
Mean p-value (reported)	4,347	0.145	0.100	0.068	0.133	0.204
Mean p-value (with imputation)	4,448	0.146	0.100	0.068	0.133	0.205
% significant at $p \leq 0.10$	4,448	72.5	18.6	60.0	74.3	87.0
% significant at $p \leq 0.05$	4,448	62.6	21.3	48.9	63.3	78.6
% significant at $p \leq 0.01$	4,448	42.6	24.2	24.4	40.6	59.3
Intro-discussed results						
Mean p-value (reported)	4,372	0.052	0.083	0.007	0.020	0.052
Mean p-value (with imputation)	4,445	0.053	0.082	0.007	0.020	0.054
% significant at $p \leq 0.10$	4,445	92.4	15.1	88.9	100.0	100.0
% significant at $p \leq 0.05$	4,445	84.1	20.7	75.0	91.7	100.0
% significant at $p \leq 0.01$	4,445	57.4	30.6	33.3	57.1	83.3
Abstract-discussed results						
Mean p-value (reported)	4,330	0.046	0.089	0.003	0.015	0.037
Mean p-value (with imputation)	4,414	0.046	0.088	0.003	0.015	0.038
% significant at $p \leq 0.10$	4,414	93.4	16.2	100.0	100.0	100.0
% significant at $p \leq 0.05$	4,414	86.1	22.6	75.0	100.0	100.0
% significant at $p \leq 0.01$	4,414	60.1	34.7	33.3	62.5	100.0
Primary results						
Mean p-value (reported)	4,354	0.028	0.077	0.000	0.007	0.023
Mean p-value (with imputation)	4,448	0.029	0.079	0.000	0.007	0.023
% significant at $p \leq 0.10$	4,448	96.5	15.0	100.0	100.0	100.0
% significant at $p \leq 0.05$	4,448	91.1	23.3	100.0	100.0	100.0
% significant at $p \leq 0.01$	4,448	66.0	40.2	33.3	100.0	100.0
Most important result						
Mean p-value (reported)	4,333	0.023	0.084	0.000	0.002	0.016
Mean p-value (with imputation)	4,448	0.024	0.086	0.000	0.003	0.017
% significant at $p \leq 0.10$	4,448	97.1	16.7	100.0	100.0	100.0
% significant at $p \leq 0.05$	4,448	92.5	26.4	100.0	100.0	100.0
% significant at $p \leq 0.01$	4,448	68.5	46.5	0.0	100.0	100.0
Mean sample size	4,398	223,241	2,843,096	921	7,504	33,776

Notes. Sample is the Table **??** empirical sample of archival and experimental research articles, limited to the 4,448 papers whose *most-important* (headline) result carries a p-value (reported, $|t|$-derived, or imputed from significance stars). Statistical results are coefficient estimates extracted from each paper's tables at the regression-row level; the unit in each row is the paper, and each subset row is conditional on the paper having a usable value for that statistic. *All results* uses the *comprehensive* extraction — every coefficient of every column of every table/panel — restricted to empirical papers also present in the main extraction; the *intro-discussed*, *abstract-discussed*, *primary*, and *most-important* subsets use the main extraction (the main result per table/panel, which carries those flags). *Mean p-value (reported)* uses an explicit p-value when reported, otherwise a p-value derived from a numeric test statistic via the 2-sided normal approximation. *Mean p-value (with imputation)* additionally fills a midpoint p-value for rows that report only stars (3–4→0.005, 2→0.030, 1→0.075, 0→0.500). *Intro-/abstract-discussed* are results whose linked discussion paragraphs appear in the introduction/abstract. *Primary* results are flagged as the paper's main hypothesis-test coefficients; *most important* is the single headline result (one per paper). Because each row counts only papers with a usable value, subsets built on a single result (e.g. *most important*) cover marginally fewer papers than subsets that pool several results per paper.

Table OA6: Inference: difference-in-differences on the statistical-significance distribution of reported empirical results around the *JAR* January 2015 data- and code-sharing policy

Panel A: Aggregates across all extracted results

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 83$)	Post-Policy ($N = 73$)	Within-JAR ($N = 156$)		JAR vs. TAR + JAE ($N = 760$)		JAR vs. All Top ($N = 1,291$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$\|z\| \geq 3.29$ ($p < 0.001$)	0.246	0.232	-0.026	-1.15	-0.010	-0.39	-0.015	-0.60
$\|z\| \in [2.58, 3.29)$ ($p \in [0.001, 0.01)$)	0.156	0.142	-0.014	-0.58	-0.014	-0.70	-0.009	-0.45
$\|z\| \in [1.96, 2.58)$ ($p \in [0.01, 0.05)$)	0.221	0.202	-0.023	-1.19	-0.026	-1.34	-0.016	-0.73
$\|z\| \in [1.65, 1.96)$ ($p \in [0.05, 0.10)$)	0.100	0.096	+0.001	+0.11	+0.006	+0.58	-0.003	-0.32
$\|z\| < 1.65$ ($p \geq 0.10$)	0.277	0.328	+0.061**	+2.43	+0.043**	+2.53	+0.044**	+2.23
Mean $\|z\|$	2.62	2.54	-0.160*	-1.73	-0.019	-0.36	-0.057	-0.77

Panel B: Results discussed in the introduction

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 81$)	Post-Policy ($N = 74$)	Within-JAR ($N = 155$)		JAR vs. TAR + JAE ($N = 759$)		JAR vs. All Top ($N = 1,296$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$\|z\| \geq 3.29$ ($p < 0.001$)	0.320	0.334	-0.023	-0.81	+0.021	+1.05	+0.014	+0.52
$\|z\| \in [2.58, 3.29)$ ($p \in [0.001, 0.01)$)	0.226	0.214	-0.024	-0.74	+0.000	+0.01	-0.003	-0.12
$\|z\| \in [1.96, 2.58)$ ($p \in [0.01, 0.05)$)	0.294	0.267	-0.007	-0.25	-0.069**	-2.23	-0.051	-1.48
$\|z\| \in [1.65, 1.96)$ ($p \in [0.05, 0.10)$)	0.079	0.102	+0.036**	+2.33	+0.022	+1.21	+0.012	+0.61
$\|z\| < 1.65$ ($p \geq 0.10$)	0.081	0.084	+0.018	+0.53	+0.025	+0.75	+0.028	+0.83
Mean $\|z\|$	3.16	3.26	-0.084	-0.78	+0.176**	+2.07	+0.092	+0.90

Table ?? (continued). Primary results and the paper's most-important (headline) result.

Panel C: Aggregates across primary results

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 82$)	Post-Policy ($N = 73$)	Within-JAR ($N = 155$)		JAR vs. TAR + JAE ($N = 754$)		JAR vs. All Top ($N = 1,288$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$\|z\| \geq 3.29$ ($p < 0.001$)	0.387	0.388	-0.026	-0.79	+0.015	+0.33	+0.006	+0.12
$\|z\| \in [2.58, 3.29)$ ($p \in [0.001, 0.01)$)	0.259	0.216	-0.046	-1.08	-0.052	-1.19	-0.051	-1.08
$\|z\| \in [1.96, 2.58)$ ($p \in [0.01, 0.05)$)	0.256	0.323	+0.088***	+3.96	+0.060	+1.45	+0.068*	+1.71
$\|z\| \in [1.65, 1.96)$ ($p \in [0.05, 0.10)$)	0.044	0.035	-0.003	-0.31	-0.003	-0.17	-0.008	-0.51
$\|z\| < 1.65$ ($p \geq 0.10$)	0.054	0.038	-0.013	-0.36	-0.019	-0.58	-0.015	-0.45
Mean $\|z\|$	3.38	3.48	-0.023	-0.19	+0.281*	+1.76	+0.145	+0.90

Panel D: Paper's most-important result

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 82$)	Post-Policy ($N = 72$)	Within-JAR ($N = 154$)		JAR vs. TAR + JAE ($N = 752$)		JAR vs. All Top ($N = 1,283$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$\|z\| \geq 3.29$ ($p < 0.001$)	0.439	0.375	-0.089	-1.40	-0.080	-1.46	-0.095	-1.61
$\|z\| \in [2.58, 3.29)$ ($p \in [0.001, 0.01)$)	0.244	0.222	-0.015	-0.30	+0.014	+0.24	+0.015	+0.26
$\|z\| \in [1.96, 2.58)$ ($p \in [0.01, 0.05)$)	0.232	0.361	+0.142***	+2.77	+0.100	+1.63	+0.113*	+1.69
$\|z\| \in [1.65, 1.96)$ ($p \in [0.05, 0.10)$)	0.049	0.014	-0.026	-0.89	-0.014	-0.45	-0.016	-0.59
$\|z\| < 1.65$ ($p \geq 0.10$)	0.037	0.028	-0.011	-0.30	-0.020	-0.54	-0.018	-0.50
Mean $\|z\|$	3.48	3.46	-0.184	-0.63	+0.083	+0.30	-0.075	-0.27

Notes. Each row reports JAR's mean before/after January 1, 2015. *Within-JAR* is the OLS β on Post_t in a JAR-only specification with topic FE and submission-year-clustered SEs; *Difference-in-Differences* columns report β_3 on JAR × Post from a two-way-FE specification (topic + submission-year FE; SEs clustered at journal × submission year). The DiD sample is empirical (archival + experimental) papers with policy assign-date in January 2012–December 2017, excluding those whose p_{10}/p_{90}-lag bounds straddle the cutoff. The sample is limited to papers whose *most-important* (headline) result carries a p-value — reported, $|t|$-derived, or imputed from significance stars — which drops only papers whose headline result is non-numeric (e.g. a machine-learning AUC, a structural confidence region, or a magnitude read from text). Panel A reports the indicators across *all* extracted results (the *comprehensive* extraction: every coefficient of every column of every table/panel). Panel B restricts to regression rows the extraction LLM flagged as discussed in the paper's introduction (the results an author chose to highlight; roughly six per paper); Panel C restricts to the paper's *primary* hypothesis-test coefficients (roughly two per paper); Panel D reports the same indicators for the paper's *most-important* (headline) result. Results whose p-value is reported only via significance stars (roughly 4% of extracted rows) are dropped because the imputation rule places stars-only rows into specific $|z|$ bands mechanically. Controls X = {PaperLength, MeanHindex, MeanYearsCareerScopus, NumAuthors, NumTables, Top20UTD}. PaperPrimaryMeanP is omitted because the outcome is itself the p-value distribution. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (two-sided tests).

Table OA7: Does the *JAR* mandate response scale with project opacity?

	Triple interaction JAR × Post × Opacity	
Outcome	Opacity index (z)	Proprietary-private (z)
Abstract readability (Fog)	-0.425*	-0.516**
	(0.253) [t=-1.68]	(0.217) [t=-2.37]
	RI p=0.333, N=3,480	RI p=0.500, N=3,480
Main-text readability (Fog)	-0.035	0.003
	(0.150) [t=-0.24]	(0.138) [t=0.02]
	RI p=1.000, N=3,480	RI p=0.833, N=3,480
Introduction hedging	-0.021**	-0.009
	(0.009) [t=-2.43]	(0.008) [t=-1.09]
	RI p=0.167, N=3,480	RI p=0.500, N=3,480
Reporting discipline (avg stars)	-0.063	0.056
	(0.074) [t=-0.86]	(0.063) [t=0.90]
	RI p=0.667, N=3,363	RI p=0.500, N=3,363
Novelty (intro-doc kNN-20)	0.004	0.005*
	(0.003) [t=1.24]	(0.003) [t=1.88]
	RI p=0.167, N=3,480	RI p=0.333, N=3,480
Topic FE, Pub-year FE	Yes	Yes
Controls (length, h-index, career, #authors, top-20)	Yes	Yes
SE clustered journal × year	Yes	Yes

Notes. Accounting-only empirical sample (JAR + TAR, JAE, CAR, RAS, MS), ±7-year window. Each cell is the coefficient on the triple interaction JAR × Post × Opacity from a DiD that includes all lower-order interactions, topic and publication-year fixed effects, and the paper-level controls. Opacity is standardized (per 1 SD). The additive index combines data ownership (private>mixed>public), access (restricted>mixed>public), a proprietary-private flag, methods-section length, an experimental bump, and Compustat-like relief. Expected signs: readability and reporting-discipline coefficients negative (Fog/stars fall more where opacity is higher), hedging negative, novelty positive. RI p re-assigns the treated journal across the six accounting journals. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (analytic, clustered).

Table OA8: Writing effects by the strength of the headline result

	JAR × Post split by headline strength		
Outcome	Strong papers	× Weak (triple)	Weak papers (net)
Introduction hedging	-0.011	-0.042*	-0.053***
	(0.023) [t=-0.48]	(0.022) [t=-1.92]	(0.019) [t=-2.78]
			RI p=0.167, N=1,010
Contribution hedging (CARS)	-0.029	-0.156**	-0.185***
	(0.039) [t=-0.76]	(0.076) [t=-2.05]	(0.049) [t=-3.78]
			RI p=0.167, N=982
Gap hedging (CARS)	-0.152***	0.087*	-0.065
	(0.036) [t=-4.24]	(0.049) [t=1.76]	(0.056) [t=-1.17]
			RI p=0.333, N=847
Abstract readability (Fog)	-0.80	-0.55	-1.34*
	(1.16) [t=-0.69]	(1.04) [t=-0.53]	(0.73) [t=-1.83]
			RI p=0.333, N=1,010
Introduction readability (Fog)	-0.67**	-0.01	-0.69
	(0.34) [t=-1.99]	(0.67) [t=-0.02]	(0.64) [t=-1.08]
			RI p=0.333, N=1,010
Topic FE, Submission-year FE	Yes	Yes	Yes
Controls (length, h-index, career, #authors, top-20)	Yes	Yes	Yes
SE clustered journal × year	Yes	Yes	Yes

Notes. Accounting-only sample, submission years 2012–2017. "Weak" papers have a headline test statistic (the paper's most-important result, $|t|$ or $|z|$) below the sample median (3.12); the split leaves roughly 32 treated post-mandate papers on the weak side. Each row is one regression of the writing outcome on JAR × Post × Weak with all lower-order terms, topic and submission-year fixed effects, and the paper-level controls. "Strong papers" is the JAR × Post coefficient; "Weak papers (net)" adds the triple interaction (delta-method standard error). Hedging outcomes are hedge rates (higher = more hedged); readability is the Gunning Fog index (higher = more complex). The signed prediction is negative throughout, so RI p is the one-sided cross-journal relabeling rank of the net weak-paper effect among the six accounting journals (floor $1/6 \approx 0.167$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (analytic, clustered).

Table OA9: Timing-imputation robustness: received-date-only re-estimation

	Full sample			Received-date only		
Outcome	Coef.	t	N	Coef.	t	N
Abstract readability (Fog)	-1.250**	-2.03	1,356	-1.398**	-2.48	959
Intro-document novelty (kNN-20)	0.013**	2.53	1,356	0.014***	2.84	959
Introduction hedging rate	-0.028**	-2.41	1,356	-0.025**	-2.11	959
Log citations	-0.277*	-1.67	1,356	-0.202	-1.28	959
Randomization-inference p (received-only, relabel each control journal as treated):						
Abstract readability (Fog)	$\hat{\beta} = -1.398$, RI $p = 0.400$ (5 relabelings)					
Intro-document novelty (kNN-20)	$\hat{\beta} = 0.014$, RI $p = 0.400$ (5 relabelings)					

Notes. DiD estimates of JAR×Post from $y \sim \text{JAR} + \text{JAR} \times \text{Post} + \text{Post} + C(\text{pub. year}) + C(\text{topic}) + \text{controls}$, empirical (archival+experimental) papers, assign-date Jan 2012–Dec 2017, all-accounting control set (TAR, JAE, CAR, RAS, MS). Controls: paper length, mean author h-index, mean career years, number of authors, top-20 UTD. Standard errors clustered by journal×year. "Full sample" uses received plus imputed-certain submission dates; "Received-date only" keeps papers with an observed submission date. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C Within-editor stability at *JAR* (Design D)

This appendix supports the within-editor test in Section **??**. Table **??** reports the post-versus-pre-2015 change on each outcome estimated within the three *JAR* editors who span the mandate—editor by editor and pooled with editor fixed effects—alongside the whole-journal benchmark and a falsification at the journal's actual 2019 masthead turnover. Figure **??** plots the same coefficients. Editor assignments and the received-to-accepted duration measure are constructed for published papers only, so every estimate here conditions on acceptance and speaks to the composition of what a given editor published, not to submissions or rejections; the design accordingly rules out that the 2015 shifts are an artifact of masthead turnover, but does not separate demand from supply.

Table OA10: The same editors show the same shifts: within-editor before/after changes at *JAR*

	Post − pre (2015), editor fixed effects					Falsification
Outcome	JAR-wide	Pooled (3 ed.)	Berger	Leuz	Skinner	Break at 2019
Abstract readability (Fog)	-1.166***	-1.125***	-0.723	-1.190**	-1.532**	-0.350
	(0.295)	(0.343)	(0.571)	(0.555)	(0.728)	(0.284)
Introduction hedging rate	-0.016*	-0.017**	-0.021	-0.029*	+0.002	-0.013
	(0.009)	(0.008)	(0.021)	(0.016)	(0.023)	(0.008)
Intro-doc novelty (kNN-20)	+0.008**	+0.007**	+0.010	+0.003	+0.006	+0.002
	(0.003)	(0.003)	(0.008)	(0.007)	(0.007)	(0.007)
Log review duration	+0.335***	+0.198***	+0.113	+0.250**	+0.228**	—
	(0.073)	(0.051)	(0.166)	(0.117)	(0.116)	
Data opacity (prop. either)	+0.112**	+0.067	+0.012	-0.033	+0.245**	+0.019
	(0.045)	(0.068)	(0.096)	(0.102)	(0.095)	(0.062)
Editor FE	Yes	Yes	—	—	—	Yes
SE clustered assign-year	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Design D. The sample is the three *JAR* editors who handled at least five papers on each side of the January-2015 mandate—Berger (45/43), Leuz (34/55), and Skinner (35/39); Sapra is excluded (pre-mandate $n = 5$). Each cell is the coefficient on a post-2015 indicator from a regression of the outcome on that indicator, estimated within the pooled spanning subsample with editor fixed effects (columns 2–3) or editor by editor (columns 4–6), with standard errors clustered on the assignment year; the *JAR*-wide column is the whole-journal benchmark. Adding topic fixed effects to the pooled specification leaves the estimates essentially unchanged (readability −1.172, hedging −0.016, novelty +0.009, log duration +0.203). The final column is a falsification at the journal's *actual* masthead turnover: among post-2015 *JAR* papers, the break at the 2019 editor transition (Hail, Wittenberg-Moerman, Verdi) on each outcome; log duration is untestable there because the received-by-2017 censoring rule leaves no computable post-2019 durations. Data opacity does not move within editor except for Skinner, whose desk's auditing share rises from 6% to 18% over the window—a composition shift, not a within-editor change (his opacity coefficient is more than twice the journal-wide change and survives topic fixed effects on his desk alone). This design rules out that *JAR*'s 2015 shifts are an artifact of who sits in the editor's chair; it does not separate the demand margin from supply, because the same editor accepts from a changed submission pool. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (analytic, clustered).

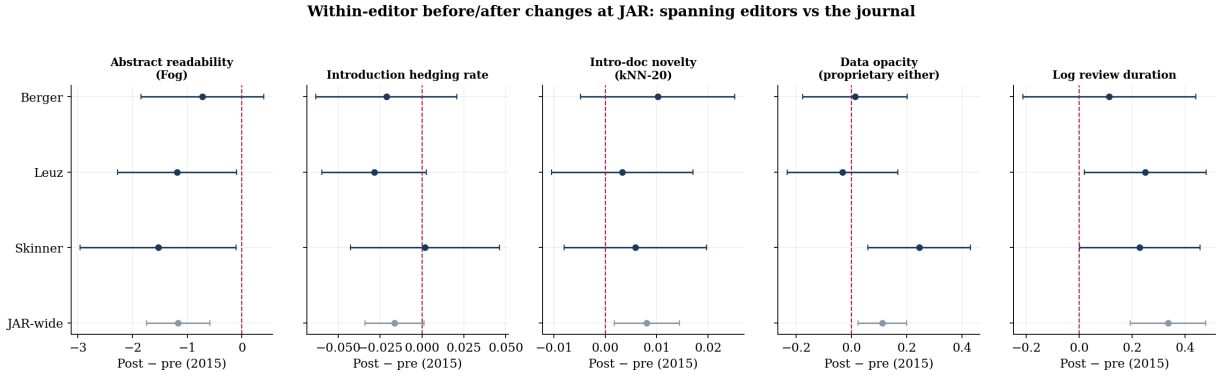

Figure OA1: The same editors show the same shifts. Post-versus-pre-2015 change at *JAR* on each outcome, estimated editor by editor for the three editors who span the mandate (Berger, Leuz, Skinner; navy) and for the whole journal (grey), with 95% confidence intervals. Points left of the dashed line mark declines (plainer, less hedged); points to the right mark increases (more novel, slower reviews, more opaque data). Data opacity moves only for Skinner, a desk-composition artifact discussed in Section **??**.

D Data construction and reproducibility

All estimates are computed from an aggregate corpus of accounting (and, for the general-equilibrium tests, finance) research articles. Text measures (section-level Gunning–Fog, hedging rates, and embedding-based novelty) and the extracted reported-results data are constructed at the paper level and merged to bibliographic and citation metadata. The difference-in-differences frame is restricted to archival and experimental research articles with an abstract and a classifiable introduction, with treatment timing assigned at the submission date. The design-based inference (randomization inference, the restricted wild cluster bootstrap, in-time placebos), the synthetic-control and synthetic difference-in-differences estimators, the Oster bounds, and the finance-benchmark general-equilibrium tests, the second-treated-journal (*Management Science*, 2019) and continuous policy-index dose-response designs, and the proprietary-data and weak-result heterogeneity are implemented in Python. The replication package (`code/`, documented in its README) is organized as a numbered pipeline: a shared data loader (`_engine.py`) followed by scripts `01`–`09` that build each channel's estimates, the design-based inference, the extensions, and every figure and table in the paper, with an `extensions/` folder holding the exploratory analyses and their result files. Every figure and table here is reproduced by that pipeline from the bundled aggregate data. The full package—code, the bundled aggregate data, and the README—will be posted publicly upon publication; until then it is available from the authors on request.

E Design-based inference

Before the inference battery, Table **??** collects in one place the design each headline result's interpretation rests on.

Because treatment is assigned to a single journal, we confirm the main estimates with inference that does not rely on large-cluster asymptotics. Table **??** reports, for one primary outcome per channel, the $JAR \times Post$ estimate together with the analytic clustered p-value and four design-based alternatives: the restricted wild cluster bootstrap, an in-time placebo, cross-journal randomization inference, and the synthetic-control RMSPE-rank test. The novelty effect clears the cross-journal randomization inference and synthetic-control tests at the floor these designs allow ($p \approx 0.17$) and is significant analytically ($p = 0.01$); the wild bootstrap (0.16) and in-time placebo (0.38) are weaker, reflecting the limited power of a single treated cluster, so we corroborate novelty with the convergent-validity evidence in Appendix **??**. The writing and citation effects keep their sign and magnitude but, in a single small journal, are estimated less precisely under the most conservative procedures.

Table OA11: Identification designs by outcome

Result	Headline estimate	Margin	Primary supporting design	Where
Readability improves	Abstract Fog −1.25	Behavioral	Second treated journal (*MS* 2019) within author; exempt-theory placebo; opacity and evidence-strength gradients	§??, §??
Hedging decline concentrates in weak-evidence papers	Contribution hedging −0.185 (weak papers)	Behavioral (mechanism)	Triple interaction with all lower-order terms; directional RI at the design floor	§??
More supporting coefficients; fewer headline claims	Multiplicity ratio +14.9; ≈15 more coefficients within author	Behavioral	The one channel that survives explicit author fixed effects; flat test-statistic distribution	§??, §??
Novelty rises	Intro novelty +0.013 ($\approx \frac{1}{3}$ SD)	Market (selection)	Directional RI and synthetic control at the floor; no-text reference-network corroboration; stable core topic mix	§??
Rare-topic share rises	+12.9pp	Market (selection)	Step at 2015; *JAR* first among all journal relabelings; threshold, leave-one-year-out, and pre-only-definition robustness	§??
Citation counts fall; citing-journal status rises	Count −0.28; top-6 share +2.5pp	Market (reception)	Tier shares and citing-SNIP robust; prestige gain concentrated in top accounting/business citers	§??

Notes. Treatment is a single journal, so analytic clustered t-statistics are descriptive throughout and no single design supports every result; this table summarizes the design each result's interpretation rests on. Cross-journal randomization inference (RI) with six accounting journals has a floor of $1/6 \approx 0.167$: "at the floor" means *JAR* ranks first among all journal relabelings, the strongest statement the design permits, not conventional 5% design-based significance. Full design-based inference for every estimate is collected in Appendix **??**.

One caveat cuts the other way: with a single treated cluster the wild cluster bootstrap is known to under-reject (**?**), so its large p-values understate rather than overturn the evidence. Figure **??** shows the corresponding synthetic-control gaps: novelty is the series in which *JAR* separates from the placebo donor cloud after the policy, while the other outcomes lie within the placebo spread.

Table OA12: Design-based inference for one treated cluster

Outcome	Coef.	Analytic cluster	Wild boot.	In-time placebo	Cross-jrnl RI	Synth. ctrl
Novelty (intro-doc)	+0.013	0.01	0.16	0.38	0.17	0.17
Abstract readability	−1.250	0.04	0.30	0.50	0.17	0.89
Introduction hedging	−0.028	0.02	0.43	0.31	0.17	0.67
Log citations	−0.277	0.09	0.55	0.69	0.17	0.22
Cite share top-6 acct.	+0.028	0.00	0.09	0.19	0.17	0.50

Notes. Each row reports the $JAR \times Post$ point estimate and five design-based p-values. *Analytic cluster*: cluster-robust on journal×year (the value behind Table **??**). *Wild boot.*: restricted wild cluster bootstrap-t on the journal (1,999 replications); conservative with one treated cluster (**?**). *In-time placebo*: placebo policy years among controls. *Cross-journal RI*: each control journal relabeled treated, one-sided in the direction predicted by P1–P5 (the floor with six journals is $1/6 \approx 0.17$); *JAR* is the most extreme journal in the predicted direction for every outcome, so each reaches the floor. *Synthetic control*: RMSPE-ratio rank p. With a single treated cluster these procedures are deliberately conservative and no outcome clears every one, but every effect passes the cross-journal RI at the floor, the writing and novelty effects clear the analytic cluster test, and novelty also passes the synthetic control (0.17). Novelty's remaining soft spot is the in-time placebo (0.38); we therefore corroborate it with the convergent-validity evidence in Appendix **??**, where reference-network measures that use no text move the same way and clear the RI floor.

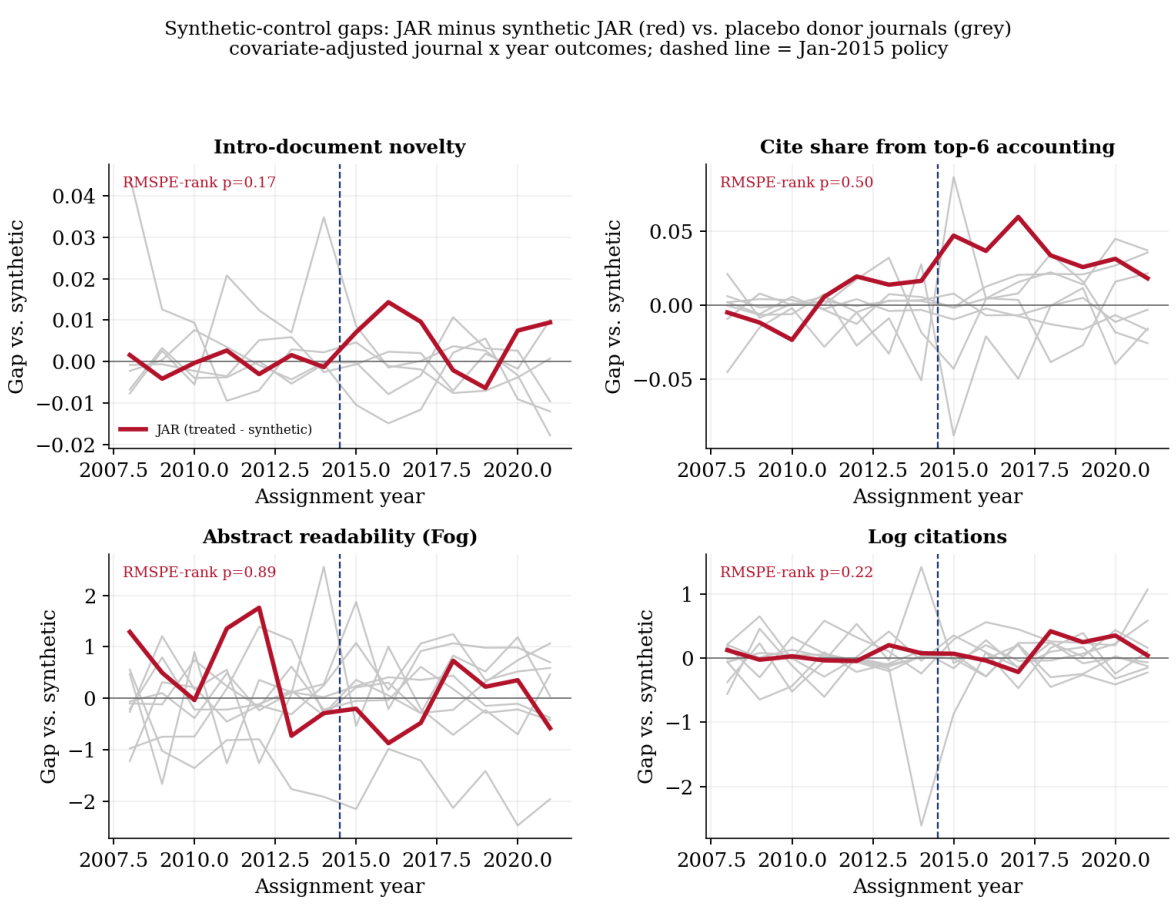

Figure OA2: Synthetic-control gaps. *JAR* minus synthetic *JAR* (solid) against placebo donor journals (grey), for covariate-adjusted journal×year outcomes. The vertical line marks the policy. RMSPE-rank p-values are reported in each panel.

F Supporting tables for the writing results

This appendix collects the coefficients and t-statistics that underlie the writing figures in Section **??**. Table **??** reports the section-level difference-in-differences for hedging and readability that Figure **??** plots, with a descriptive panel for the pooled sample. Table **??** reports the same estimates within the introduction, move by move, formalizing the rhetorical-move asymmetry in the right panel of Figure **??**: hedging falls in the gap move (-0.087, $t = -2.42$) and the contribution move (-0.091, $t = -3.46$) and is flat in the descriptive moves (motivation, results preview, present research). Both tables use the same specification and sample as the main estimates.

Table OA13: Paper-level sections: descriptives and difference-in-differences around the *JAR* January 2015 data- and code-sharing policy

Panel A: Descriptive statistics by section

	N	Mean	SD	Q1	Median	Q3
Whole paper						
Length (sents)	4,630	425	88	366	418	474
Hedging	4,630	0.35	0.06	0.30	0.34	0.39
Readability	4,630	16.7	1.5	15.7	16.7	17.7
Abstract						
Length (sents)	4,630	6	2	5	6	7
Hedging	4,630	0.63	0.22	0.50	0.62	0.80
Readability	4,630	19.1	3.0	17.1	19.0	21.0
Introduction						
Length (sents)	4,629	65	18	52	64	76
Hedging	4,629	0.50	0.10	0.44	0.51	0.57
Readability	4,629	17.7	1.8	16.5	17.6	18.9
Theory/predictions						
Length (sents)	4,282	69	31	47	64	85
Hedging	4,282	0.46	0.14	0.37	0.47	0.56
Readability	4,282	17.3	1.9	16.1	17.3	18.5
Methodology/data						
Length (sents)	4,623	74	39	46	69	95
Hedging	4,623	0.18	0.08	0.12	0.18	0.24
Readability	4,623	15.8	1.9	14.6	15.8	17.0
Results						
Length (sents)	4,629	195	78	139	187	243
Hedging	4,629	0.28	0.08	0.23	0.28	0.33
Readability	4,629	16.4	1.7	15.3	16.4	17.5
Conclusions						
Length (sents)	4,627	21	16	12	17	25
Hedging	4,627	0.62	0.17	0.50	0.62	0.74
Readability	4,627	18.1	2.3	16.7	18.1	19.6

Table ?? (continued). Difference-in-differences by section: sentence hedging and Gunning Fog readability as sub-rows under each section banner.

Panel B: Hedging and Readability DiD by section

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 84$)	Post-Policy ($N = 78$)	Within-JAR ($N = 162$)		JAR vs. TAR + JAE ($N = 788$)		JAR vs. All Top ($N = 1,356$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Whole paper								
Hedging	0.366	0.362	-0.012	-1.43	-0.013	-1.64	-0.009	-1.09
Readability	17.33	16.87	-0.570**	-2.56	-0.768**	-2.35	-0.533*	-1.82
Abstract								
Hedging	0.664	0.610	-0.068*	-1.73	-0.028	-0.73	-0.032	-0.93
Readability	20.52	19.36	-1.221**	-2.32	-1.352***	-2.60	-1.250**	-2.03
Introduction								
Hedging	0.523	0.503	-0.026*	-1.81	-0.027**	-2.02	-0.028**	-2.41
Readability	18.74	17.98	-0.835**	-2.29	-1.079***	-2.74	-0.959**	-2.52
Theory/predictions								
Hedging	0.448	0.417	-0.038*	-1.88	-0.018	-0.85	-0.024	-1.07
Readability	18.22	17.40	-0.950**	-2.51	-1.017***	-2.90	-0.669*	-1.96
Methodology/data								
Hedging	0.194	0.186	-0.016	-1.00	-0.001	-0.07	-0.004	-0.29
Readability	16.26	16.06	-0.359	-0.91	-0.539	-1.18	-0.311	-0.73
Results								
Hedging	0.320	0.323	-0.009	-0.73	-0.012	-1.42	-0.008	-0.84
Readability	16.96	16.73	-0.280	-1.25	-0.547	-1.62	-0.313	-1.05
Conclusions								
Hedging	0.664	0.605	-0.070***	-5.12	-0.048**	-2.18	-0.042**	-2.48
Readability	18.53	17.78	-1.038***	-4.45	-1.006***	-2.83	-0.823**	-2.31

Notes. Panel A pools all journals and years in the empirical sample of archival and experimental research articles ($N = 4{,}630$; see Table **??**). *Length* is the section sentence count (or the whole paper for *Whole paper*). *Hedging* is the fraction of section sentences flagged as uncertain by the DistilRoBERTa sentence classifier (binary label at $t = 0.5$). *Readability* is Gunning Fog index. Each section row is conditional on the section being non-empty. Panel B reports difference-in-differences for the same sections, with two sub-rows per section: *Hedging* (sentence-level DistilRoBERTa rate at $t = 0.5$) and *Readability* (Gunning Fog). Each row reports JAR's mean before/after January 1, 2015. *Within-JAR* is the OLS β on Post_t in a JAR-only specification with topic FE and submission-year-clustered SEs; *Difference-in-Differences* columns report β_3 on JAR × Post from a two-way-FE specification (topic + submission-year FE; SEs clustered at journal × submission year). The DiD sample is empirical (archival + experimental) papers with policy assign-date in January 2012–December 2017, excluding those whose p_{10}/p_{90}-lag bounds straddle the cutoff. Controls $X = \{\text{PaperLength}, \text{MeanHindex}, \text{MeanYearsCareerScopus}, \text{NumAuthors}, \text{Top20UTD}\}$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (two-sided tests).

Table OA14: Introduction paragraphs by CARS discourse-move category: descriptives and difference-in-differences around the *JAR* January 2015 data- and code-sharing policy

Panel A: Descriptive statistics by discourse category

	N	Mean	SD	Q1	Median	Q3
Motivation						
Length (paragraphs)	4,504	2.4	1.4	1.0	2.0	3.0
Hedging	4,504	0.46	0.22	0.33	0.46	0.60
Readability	4,504	19.0	3.9	16.6	18.6	20.8
Gap						
Length (paragraphs)	3,891	1.8	1.0	1.0	2.0	2.0
Hedging	3,891	0.62	0.21	0.50	0.62	0.75
Readability	3,891	18.8	3.8	16.4	18.5	20.6
Present research						
Length (paragraphs)	4,612	3.2	1.7	2.0	3.0	4.0
Hedging	4,612	0.44	0.18	0.31	0.43	0.56
Readability	4,612	19.1	3.1	17.2	18.9	20.8
Results preview						
Length (paragraphs)	4,586	3.7	2.0	2.0	3.0	5.0
Hedging	4,586	0.59	0.16	0.50	0.60	0.70
Readability	4,586	19.2	2.7	17.4	19.0	20.8
Contribution						
Length (paragraphs)	4,488	2.6	1.3	2.0	3.0	3.0
Hedging	4,488	0.61	0.18	0.50	0.62	0.73
Readability	4,488	18.6	3.1	16.6	18.4	20.1
Other						
Length (paragraphs)	3,710	2.1	1.5	1.0	2.0	3.0
Hedging	3,670	0.17	0.24	0.00	0.00	0.29
Readability	3,710	12.9	4.2	10.3	12.6	15.2

Table ?? (continued). Difference-in-differences by CARS discourse category: sentence hedging and Gunning Fog readability as sub-rows under each category banner.

Panel B: Hedging and Readability DiD by discourse category

			Difference		Difference-in-Differences			
	Pre-Policy ($N = 82$)	Post-Policy ($N = 76$)	Within-JAR ($N = 158$)		JAR vs. TAR + JAE ($N = 765$)		JAR vs. All Top ($N = 1,313$)	
	Mean	Mean	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Motivation								
Hedging	0.502	0.462	-0.043	-1.09	-0.016	-0.61	-0.005	-0.22
Readability	20.16	18.64	-1.645**	-2.50	-2.046***	-2.64	-1.135	-1.30
Gap								
Hedging	0.673	0.582	-0.096***	-4.38	-0.076*	-1.89	-0.087**	-2.42
Readability	20.53	19.40	-1.327	-1.56	-0.946	-1.19	-0.491	-0.52
Present research								
Hedging	0.439	0.397	-0.054*	-1.74	-0.022	-0.77	-0.017	-0.59
Readability	19.84	19.09	-0.941	-1.40	-1.264	-1.64	-0.788	-1.06
Results preview								
Hedging	0.596	0.575	-0.025	-0.92	-0.017	-0.71	-0.023	-0.98
Readability	20.05	18.87	-1.197**	-2.37	-1.498***	-3.22	-1.330***	-3.11
Contribution								
Hedging	0.664	0.577	-0.085***	-2.71	-0.094***	-3.76	-0.091***	-3.46
Readability	19.71	18.72	-1.043**	-2.29	-1.133***	-2.65	-0.947**	-2.00

Notes. Introduction paragraphs are classified into one of six CARS discourse-move categories (Motivation, Gap, Present research, Results preview, Contribution, Other) using a CARS-style prompt with a Qwen LLM. Both panels use the canonical primary-or-secondary CARS convention (a paragraph counts toward a category if either its primary or secondary label matches). Panel A reports paragraph-count per paper, fraction of sentences flagged as uncertain (DistilRoBERTa, $t = 0.5$), and Gunning Fog readability for the concatenated text of the category. Panel B reports difference-in-differences for the same five primary categories, with two sub-rows per category: *Hedging* (sentence-level DistilRoBERTa rate at $t = 0.5$) and *Readability* (Gunning Fog). Each row reports JAR's mean before/after January 1, 2015. *Within-JAR* is the OLS β on Post_t in a JAR-only specification with topic FE and submission-year-clustered SEs; *Difference-in-Differences* columns report β_3 on JAR $\times$ Post from a two-way-FE specification (topic + submission-year FE; SEs clustered at journal $\times$ submission year). The DiD sample is empirical (archival + experimental) papers with policy assign-date in January 2012–December 2017, excluding those whose p_{10}/p_{90}-lag bounds straddle the cutoff. Controls $X = \{$PaperLength, MeanHindex, MeanYearsCareerScopus, NumAuthors, Top20UTD$\}$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (two-sided tests).

G Threshold bunching: showcased versus incidental results

Section ?? reports that we find no robust evidence the mandate reduced threshold-driven inference. This appendix shows why. We split every extracted coefficient into the results authors showcase—the primary hypothesis tests, the results discussed in the introduction, and the single headline result—and the incidental coefficients that fill out the tables, and we estimate the difference-in-differences on the caliper share just above and just below each

conventional cutoff. Figure **??** collects the picture. The just-above-one-percent share falls, and it falls most in the showcased results (-0.077 for the primary results, -0.015 for the incidental ones), which on its own looks like deterred specification search. But the displaced mass does not move into the non-significant region. It reappears just above the five-percent cutoff, where the showcased share rises by roughly 0.10, while the share of insignificant showcased results does not rise and the overall share of significant showcased results, if anything, increases. With a single treated journal the one-percent decline is also fragile: relabeling each control journal as treated places *JAR* at the extreme of the placebo distribution but at a one-sided randomization-inference p of 0.17. The pattern is that of reported statistics sorting one threshold down, not of less specification search, so we treat the test-statistic channel as an unresolved null.

H Novelty: robustness and convergent validity

The novelty result in Section **??** is the paper's most inference-robust finding, so we document that it does not depend on a single measure or a single embedding choice. Table **??** collects the evidence. Panel A re-estimates the $JAR \times Post$ effect across nearest-neighbor distances at $k \in \{1, 5, 10, 20, 50\}$, a five-year reference window, an abstract-only variant, an "originality" variant, and a distance-to-field-centroid variant. The effect is stable between one-third and one-half of a standard deviation everywhere except the centroid measure, which is weak—the signature of a *local* shift, in which papers move away from their nearest predecessors while the field's center holds still. Panel B turns to measures built from the reference list rather than the text: post-mandate *JAR* papers cite a wider range of fields (higher entropy, a larger non-accounting share, more distinct journals), and because these use no text they are an independent check on the embedding measure. Panel C confirms that the text and citation measures correlate positively at the paper level, while the rhetorical novelty-claim indicator is nearly orthogonal to all of them and is therefore reported as a distinct construct. The one measure that runs the other way is Uzzi journal-pair atypicality, which *falls*: the mandate sorts in papers that reach across more fields but combine *established* journals in more conventional pairs, so the novelty here is one of breadth and textual distance, not of unusual recombination.

Two descriptive figures make concrete what kind of novelty the mandate sorts in. Figure **??** plots the density of introduction novelty in *JAR* papers before and after the policy: the distribution does not shift uniformly but grows a thicker upper tail, with more mass past the pre-mandate ninetieth percentile. Figure **??** traces the mechanism on the citation side, showing that *JAR* papers come to draw a markedly larger share of their references from

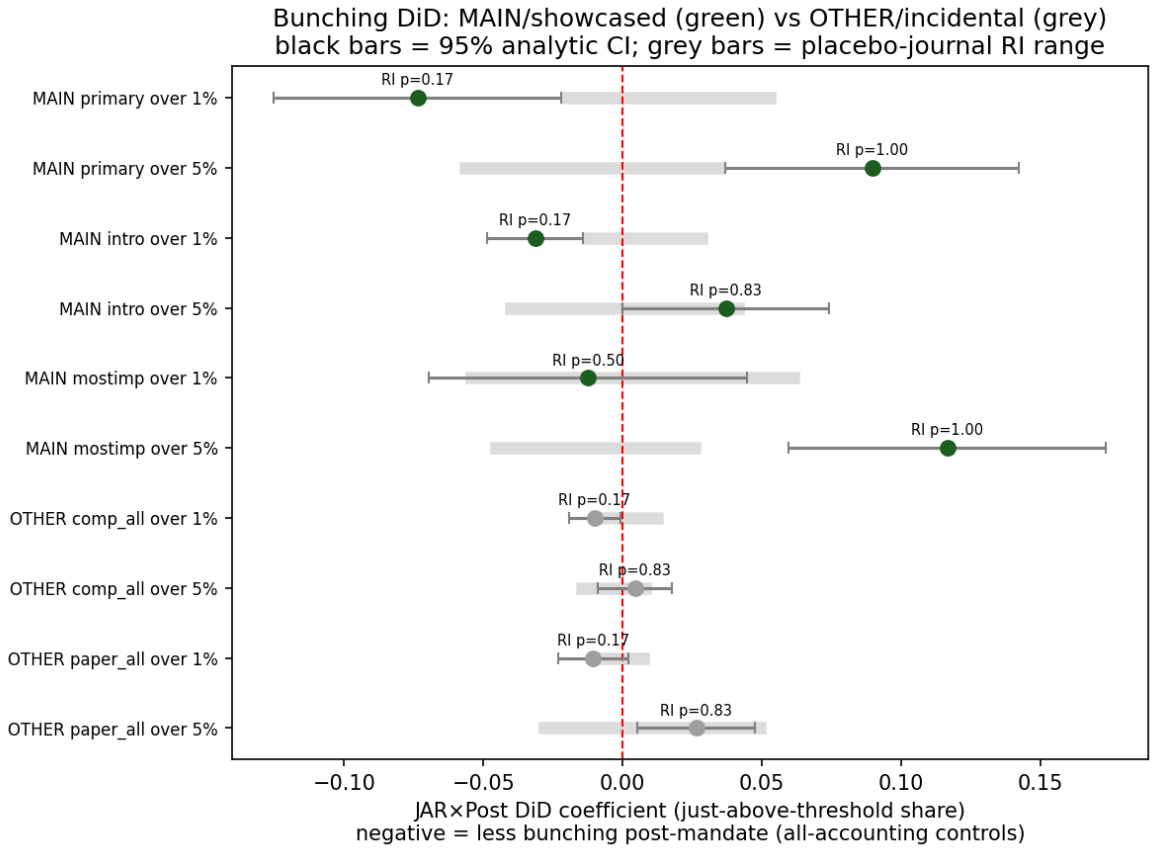

Figure OA3: Just-above-threshold bunching: showcased versus incidental results. $JAR \times Post$ difference-in-differences estimates on the share of a paper's reported results in the just-above caliper of each conventional threshold, split by whether the result is one the author showcases (primary, introduction-discussed, headline) or an incidental coefficient (the full universe of extracted results). The one-percent just-above share falls most in showcased results, but the mass relocates just above the five-percent threshold rather than into the non-significant region. See Section **??** and the discussion in this appendix for the corresponding randomization-inference p-values.

Table OA15: Novelty: robustness across measures and convergent validity

Measure ($JAR \times Post$)	SD units	t	RI p
Panel A: Embedding-based measures			
Intro-document kNN-20 (headline)	+0.36	2.52	0.17
Abstract kNN-20	+0.43	2.83	0.17
kNN-5 (intro)	+0.39	3.69	0.17
kNN-50 (intro)	+0.41	2.75	0.17
Originality (1 − max similarity)	+0.33	3.27	0.17
kNN-20, five-year reference window	+0.44	3.22	0.17
Distance to field centroid	+0.17	1.41	0.50
Panel B: Reference-network measures (no text input)			
Non-accounting reference share	+0.48	3.73	0.17
Reference field entropy	+0.40	3.12	0.17
Distinct cited journals	+0.40	2.49	0.17
Mean reference age	+0.18	1.69	0.33
Uzzi journal-pair atypicality	−0.40	−3.44	0.17
Panel C: Convergent validity (correlation with headline kNN-20)			
Abstract kNN-20	$r = 0.70$		
Non-accounting reference share	$r = 0.35$		
Distinct cited journals	$r = 0.29$		
Reference field entropy	$r = 0.27$		
Uzzi journal-pair atypicality	$r = -0.33$		
Rhetorical novelty claim (indicator)	$r = 0.01$		

Notes. Panels A and B report the $JAR \times Post$ coefficient from (**??**) on each novelty measure, expressed in standard-deviation units of the measure, for the All-Accounting comparison set and the headline 2012–2017 window; the Top-3 set gives the same picture. "RI p" is the cross-journal randomization-inference p-value—each control journal relabeled treated, one-sided in the predicted direction (novelty *rises*), for which the floor with six journals is $1/6 \approx 0.17$. *JAR* is the most novelty-increasing journal in every embedding and reference-diversity measure, so each reaches the floor; the two discriminant checks (distance to the field centroid and mean reference age, where no directional prediction is made) are reported two-sided and, as expected, do not clear it. The embedding effect is stable across every k, reference window, and variant (SD magnitudes 0.33–0.44). It is *local*: distance to nearest predecessors moves but distance to the field centroid does not. Most tellingly, the reference-network measures use no text yet move the same way and clear the RI floor (0.17), and they correlate positively with the text measure (Panel C)—an independent corroboration that does not depend on the embedding pipeline; the exception is Uzzi journal-pair atypicality, which *falls* (200-shuffle degree-preserving null, seed 42), indicating that post-mandate papers reach into more fields through more *conventional* journal pairings rather than through unusual recombination. The rhetorical novelty-claim indicator is nearly orthogonal to every distance measure ($r \approx 0.01$), so we treat it as a distinct self-report construct.

outside accounting after the mandate while the control journals hold flat.

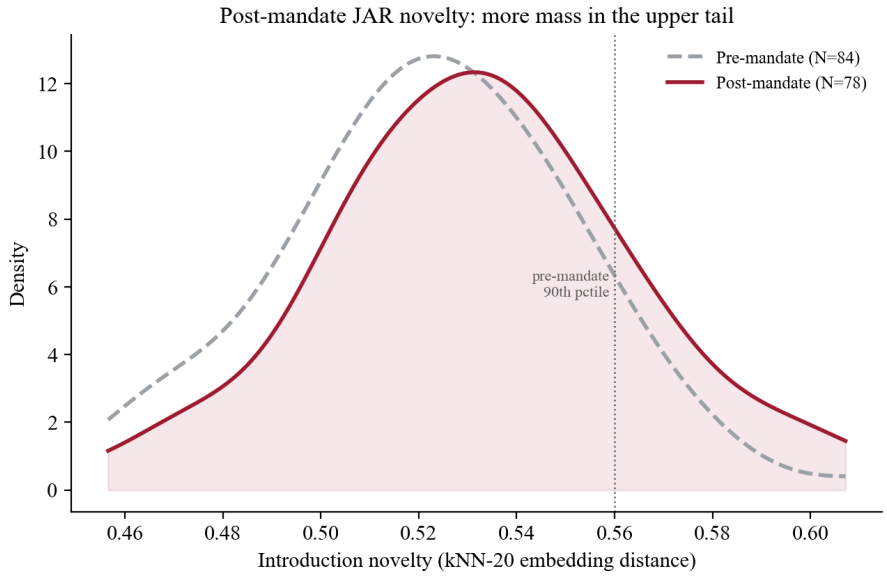

Figure OA4: Distribution of introduction novelty, before versus after the mandate. Kernel density of introduction novelty (kNN-20 embedding distance) in *JAR* papers before and after the mandate. Mass moves past the pre-mandate ninetieth percentile (dotted line), the distributional counterpart of the quantile estimates in Figure **??**.

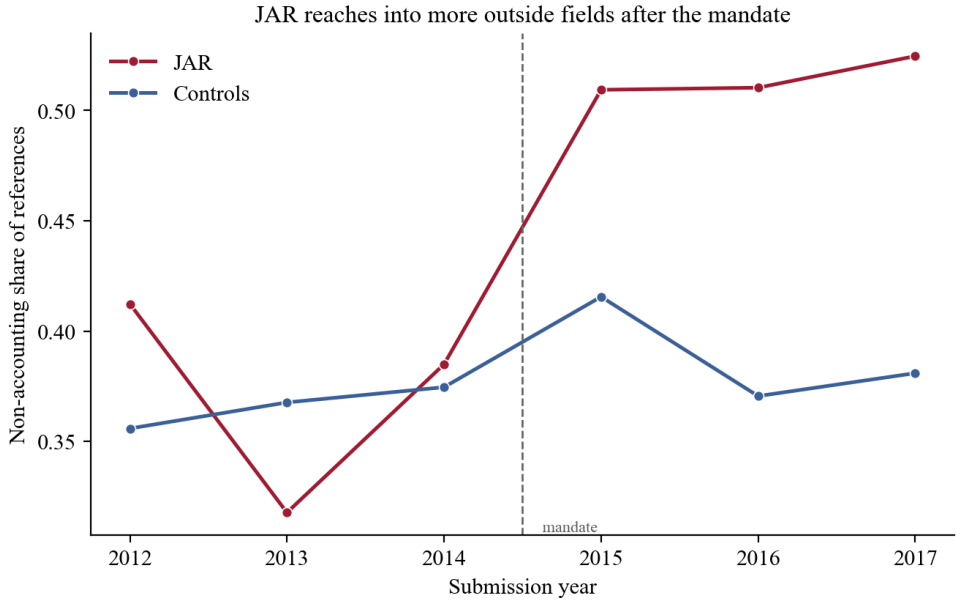

Figure OA5: Non-accounting share of references, by submission year. Mean non-accounting share of a paper's references by submission year, for *JAR* and the pooled accounting controls; the dashed line marks January 2015. The share jumps for *JAR* after the policy while the controls hold flat, a visual analogue of the reference-based difference-in-differences in Table **??**.

Finally, two figures document the topic-composition evidence behind Section **??**. Figure **??** plots each coarse topic's share of *JAR* papers before and after the mandate against the same movement in the controls: the shares track, and no per-topic difference-in-differences survives a Benjamini–Hochberg correction (Figure **??**; smallest $q = 0.20$). The one exception, plotted separately in the main text, is the rare-topic margin (Figure **??**), the only topic outcome that is BH-significant ($q < 0.001$). Distribution-level tests agree with the per-topic view: the Jensen–Shannon divergence between *JAR*'s and the controls' topic mixes moves by +0.007 bits at the mandate (relabeling RI $p = 0.60$; within-*JAR* permutation $p = 0.67$; exact χ^2 $p = 0.69$), and *JAR*'s granular-topic entropy rises by about half a nat relative to controls (from 11 to 20 distinct topics on a nearly constant paper count), which is the concentration-side expression of the same rare-topic entry, not an independent fact.

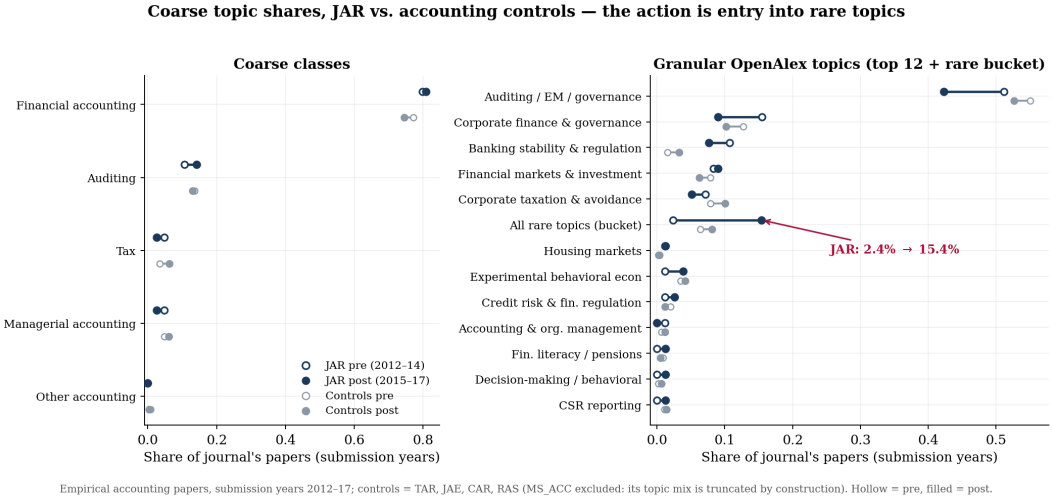

Figure OA6: *JAR*'s core topic mix is stable at the mandate. Coarse topic shares of *JAR* papers (navy) and the accounting controls (grey), pre (2012–14) versus post (2015–17) submission years. No share movement survives multiple-testing correction.

I Citations and selection: supporting figures

This appendix collects the supporting evidence for the market-margin results in Sections **??** and **??**. Figure **??** reports the $JAR \times Post$ effect on every citing-journal prestige and tier measure in standard-deviation units, including the versions that drop the top-six accounting journals as citers; the effect is positive and significant across the board, so the rise in citing-journal status is not an artifact of any single measure or of the field's own leaders. Figure **??** shows the accompanying broadening in the fields that cite *JAR*: the business-and-accounting share of citations falls after the mandate while economics, finance, and the social and decision

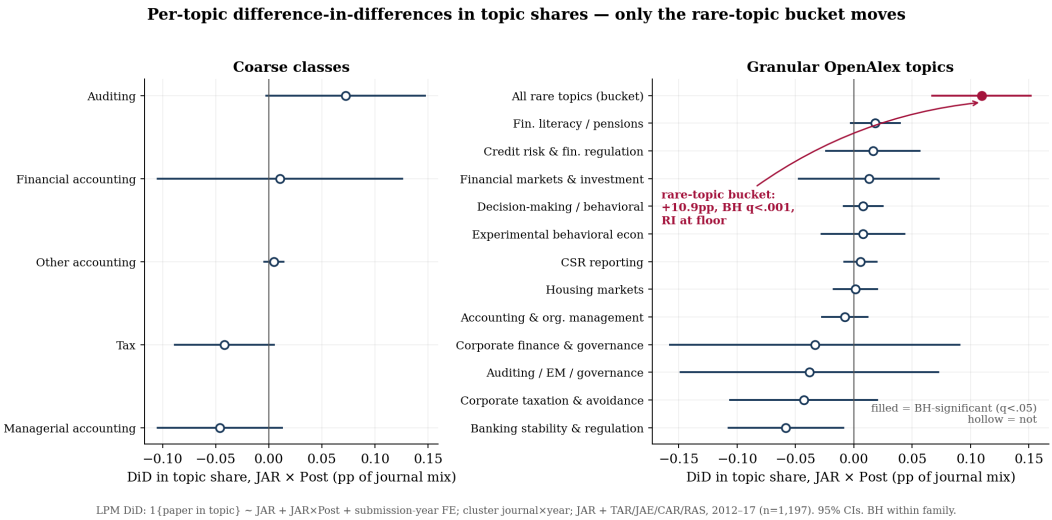

Figure OA7: Per-topic difference-in-differences. *JAR* × *Post* estimates for each topic-share outcome with 95% confidence intervals (analytic clustering by journal×year), sorted by coefficient; filled markers survive Benjamini–Hochberg correction, hollow markers do not. Only the rare-topic margin survives.

sciences pick up the difference.

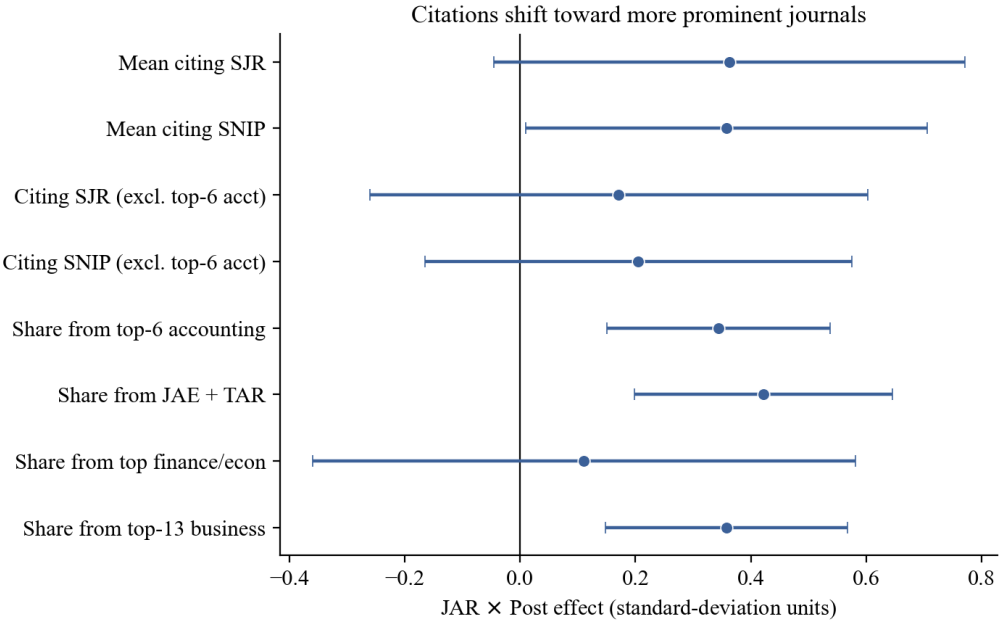

Figure OA8: Citing-journal status rises across the main prestige and tier measures. *JAR* × *Post* difference-in-differences on each citation-prestige and citation-tier measure, in standard-deviation units, with 95% confidence intervals (All-Accounting sample; analytic clustering by journal×year). The "excl. top-6 accounting" rows drop the six leading accounting journals from the citing pool; those estimates remain positive but are not statistically distinguishable against the accounting controls, so the prestige gain is concentrated in the field's leading outlets.

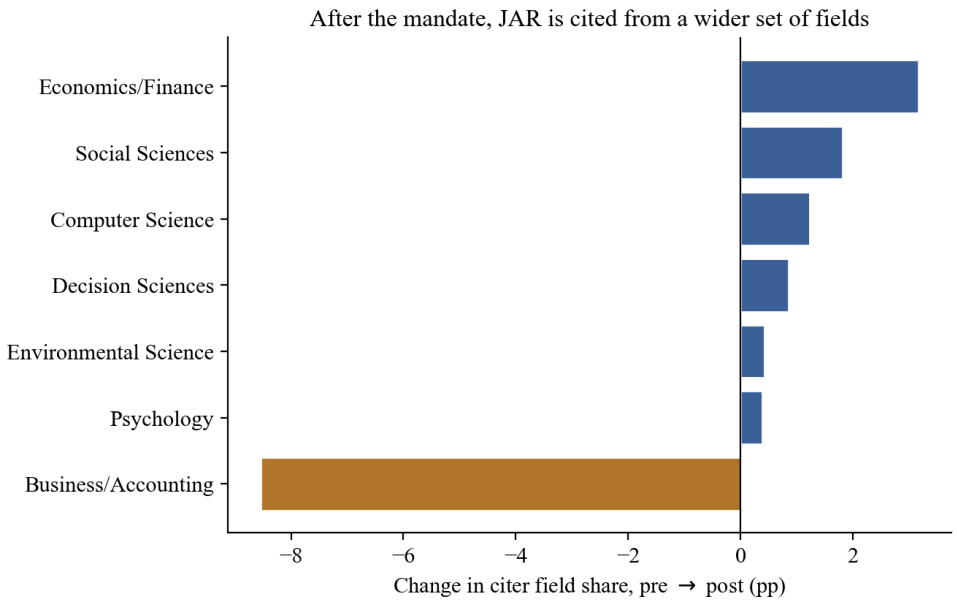

Figure OA9: After the mandate, *JAR* is cited from a wider set of fields. Change in the share of *JAR*'s citations originating in each field, pre- versus post-mandate. Business and accounting fall; economics, finance, and the social and decision sciences rise.

J Readability: robustness across indices

The readability result in Section **??** does not depend on the choice of the Gunning–Fog index. Figure **??** re-estimates the $JAR \times Post$ effect for the abstract and the introduction using both Fog and the SMOG grade, a separate readability index built entirely from the count of polysyllabic words. The two indices give the same answer—each significant, each between roughly one-third and one-half of a standard deviation—and they track each other at the paper level ($r = 0.98$ for the abstract, 0.88 for the introduction), so the estimated gain is a property of the prose, not an artifact of one formula. The two remaining components of Fog behave as Section **??** describes: the complex-word share moves, while average sentence length does not. Character-based indices (Flesch, Coleman–Liau) require the raw text, which our aggregated corpus does not retain, so we do not compute them. A quantile version of the readability effect shows broad-based simplification with a modest gradient—the hardest-to-read abstracts simplify somewhat more than the median—rather than a change confined to the tail.

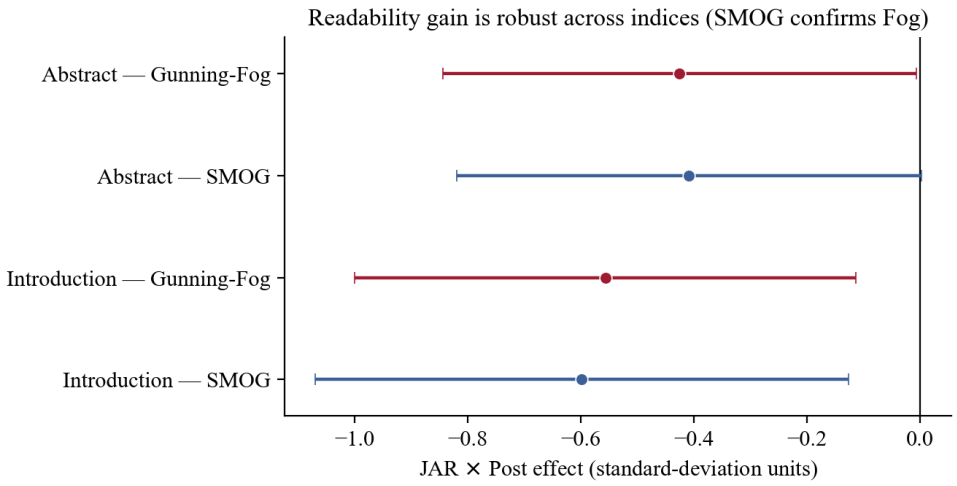

Figure OA10: The readability gain is robust across indices. $JAR \times Post$ effect on abstract and introduction readability, in standard-deviation units, measured by Gunning–Fog (crimson) and the SMOG grade (blue), with 95% confidence intervals (All-Accounting sample; clustering by journal×year). SMOG, an index based only on polysyllabic-word counts, confirms the Fog result.

K Differential measurement error

Because our outcomes are produced by a text pipeline and our treatment is a calendar cutoff, a natural concern is that the measurement instrument itself behaves differently across treated and control journals or across the pre and post periods, which would load onto the $JAR \times Post$ coefficient and masquerade as a treatment effect. We check this by regressing each pipeline primitive—text-extraction yield, the classifier's predicted probabilities, the mass of near-threshold classifications, and the share of statistics whose p-value is imputed from significance stars—on $JAR \times Post$ with journal and year fixed effects (Table **??**). The reassuring news is on the extensive margin: the share of papers with a usable abstract, introduction, and at least one extracted result is near one in every cell and shows no differential change (all $p > 0.15$), so the estimates are not driven by differential sample censoring, and the per-paper readability measures, which are deterministic transforms of the text rather than classifier outputs, are not exposed to classifier drift. The cautions are on the intensive margin and are concentrated in two measure families. First, the introduction yields about three more extracted sentences per JAR paper post-mandate and the hedging classifier labels those introductions slightly less decisively, so the introduction-hedging rate is the one writing outcome whose differential measurement we cannot fully separate from a true effect—which is one more reason we do not rest any behavioral claim on the *average* hedging decline. The weak-paper gradient of Section **??** is a within-JAR contrast between papers scored by the same pipeline in the same period, so a level drift in the classifier differences out of it. Second, JAR's post-mandate

share of star-imputed p-values is differentially higher, which compounds the caution on the reported-statistics distribution and is why we treat the test-statistic channel as an unresolved null (Section **??**) and report it on the stars-excluded aggregates. The readability and novelty headlines are not implicated by either caution. We note one limitation that the authors will close in revision: the frozen analysis files do not contain the held-out human-labeled validation set, so classifier F_1, precision, recall, and inter-rater agreement—ideally reported separately for the treated-post cell—belong in this appendix and are reported in the supplementary validation materials.

Table OA16: Differential measurement-error and coverage balance: *JAR* vs. controls, pre vs. post (empirical sample, January 2012–December 2017)

	JAR		Controls		DiD (JAR×Post)	
Measure	Pre	Post	Pre	Post	Estimate	p
Panel A: Text-extraction yield						
Abstract sentences extracted	5.940	5.859	7.743	7.547	-0.4955*	0.055
Introduction paragraphs extracted	13.917	15.487	15.245	16.767	+0.1201	0.464
Introduction sentences extracted	60.286	68.846	65.488	72.036	+2.1360***	0.000
Results extracted (all)	11.619	11.026	11.482	11.954	-1.0920***	0.000
Results extracted (primary)	2.583	1.962	2.096	1.893	-0.4452***	0.000
Share w/ usable abstract ($\geq$3 sent)	1.000	1.000	0.996	1.000	-0.0039	0.156
Share w/ usable intro ($\geq$5 sent)	1.000	1.000	0.999	1.000	-0.0016	0.340
Share w/ $\geq$1 extracted result	1.000	1.000	0.983	0.985	+0.0008	0.931
Panel B: Classifier confidence (decisiveness)						
Abstract mean hedge prob.	0.639	0.600	0.575	0.538	-0.0049	0.607
Intro mean hedge prob.	0.517	0.500	0.494	0.495	-0.0158***	0.000
Abstract mean claim prob.	0.361	0.400	0.425	0.462	+0.0049	0.606
Intro mean claim prob.	0.483	0.500	0.506	0.505	+0.0158***	0.000
Panel C: Near-threshold ambiguity (0.5 cutoff)						
Abstract near-0.5 ambiguity index	0.591	0.680	0.639	0.653	+0.0725***	0.000
Intro near-0.5 ambiguity index	0.867	0.863	0.859	0.858	-0.0051	0.231
Abstract hedged-sentence share	0.664	0.610	0.586	0.545	-0.0178	0.100
Intro hedged-sentence share	0.523	0.503	0.496	0.499	-0.0204***	0.000
Panel D: Star-imputed statistics						
Share primary p-values star-imputed	0.000	0.038	0.039	0.037	+0.0346***	0.008
Share all results lost if stars dropped	0.023	0.035	0.029	0.020	+0.0175***	0.008

Notes. Empirical (archival/experimental) sample, JAR plus accounting and finance controls, treatment-certain papers, assignment years 2012–2017. Cells are group×period means. DiD is the JAR×Post coefficient from OLS with journal and assignment-year fixed effects, standard errors clustered by journal. Panel A measures whether extractable text/statistics differ across the cutoff; Panel B whether the classifier labels sentences more/less decisively (mean predicted probabilities); Panel C the mass of near-0.5 (ambiguous) classifications and the realized hedged-sentence share; Panel D the share of extracted coefficients whose p-value is imputed from significance stars. A significant DiD indicates the *instrument* (extraction or classifier) behaves differently for JAR post-2015, so the corresponding language outcome is not cleanly comparable across the cutoff. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Held-out classifier F1/precision/recall and inter-rater agreement are author inputs and are not contained in the frozen data package.