Stress Testing an Economic Literature*

Vanitha Ragunathan
University of Queensland

Jared Stanfield
University of Oklahoma

Robert Tumarkin†
UNSW Sydney

November 18, 2026

Abstract

Researcher degrees of freedom, enabled by publication incentives, can produce a literature in which results are selected for statistical significance. We develop an approach to stress testing an economic literature using *Foghorn*, a declarative econometric framework that makes empirical choices explicit, concise, and modifiable. After independently replicating approximately 79% of the M&A studies meeting our sample criteria, we vary the construction of the dependent variable, control variable definitions, and sample selection using alternative design choices found within the literature. Across roughly 300 tests per paper (and more than 4,500 tests total) using alternative dependent variable constructions, baseline t-statistics for the key explanatory variable exceed alternatives in 80% of cases, and 37% of alternatives lose significance at 10%. This baseline advantage is significantly smaller for controls, consistent with specification selection associated with publication incentives. As LLMs lower the cost of searching for significant results, our approach enables the scrutiny of specification dependence.

JEL Classification: C10, C58, C80, G34

Keywords: Replication, Stress testing, Specification curve, Publication bias, Mergers and acquisitions

* We would like to thank Ran Duchin, Maryam Fathollahi, Cam Harvey, Mihai Ion, Adam Kolasinski, Dmitriy Muravyev, Mihail Velikov, Jason Zein, seminar participants at the RCF-ECGI Corporate Finance and Governance Conference (2025), Curtin University, Deakin University, Hong Kong Polytechnic University, Ludwig Maximilian University of Munich, the University of Illinois Chicago, the University of Melbourne, and the University of Oklahoma for helpful feedback and comments.

† Corresponding Author. UNSW Business School, UNSW Sydney, Sydney NSW 2052, Australia. Telephone: +(61) 02 9065 1656. E-mail: r.tumarkin@unsw.edu.au.

1. Introduction

Economists recognize that economically defensible yet discretionary empirical design choices can materially affect research findings. Leamer (1978) and Leamer (1983) emphasizes the fragility of estimates to the assumptions underlying econometric analysis. These discretionary choices, which Simmons, Nelson, and Simonsohn (2011) term "researcher degrees of freedom," generate uncertainty around reported estimates (Menkveld et al. 2024). In addition, the publication (and grant) process also creates strong incentives for researchers to utilize their degrees of freedom non-randomly. When researchers and journals reward statistically significant findings, published specifications may systematically occupy the favorable portion of the distribution of defensible empirical choices (Sterling 1959). This pattern need not arise from researchers searching across specifications. Even if each researcher commits to one defensible specification before observing the data, selecting papers for publication based on the significance can favor specifications that happen to produce stronger estimates. Therefore, researcher degrees of freedom generate dispersion in findings, while selection into publication can favor significant results even without specification search.

Conventional robustness tests (often reported in the body of a paper, in untabulated results, or in an internet appendix) partially address these concerns. However, these tests are necessarily idiosyncratic, and authors choose which alternatives to investigate and report. Preregistration constrains discretion after an analysis plan is specified. Yet, the significant researcher degrees of freedom in the empirical process make pre-registering each unique decision potentially cumbersome and unrealistic. Nor does it preclude researchers from finding significant specifications before the pre-registered report. The central problem remains: readers observe a published specification and selected robustness tests, but often cannot readily determine which assumptions drive the result or how it would change under other economically defensible choices. Addressing this problem requires making empirical choices both visible and inexpensive to evaluate.

The existing literature offers complementary approaches to quantify and address these important issues. Many-researcher studies ask different teams to perform the same empirical task. These studies typically, but not always, involve replicating, or independently recovering, existing findings. This "many-researchers, one-dataset" literature provides compelling evidence on the replicability of research and that degrees of freedom shape it.[1] This literature shows poor replicability and that research teams analyzing similar data can reach opposing conclusions. Extending these exercises across a literature, however, is costly: researchers must reconstruct paper-specific data processing

[1]This literature spans multiple disciplines and includes studies in psychology (Silberzahn and Uhlmann 2015; Silberzahn, Uhlmann, and Martin 2018), neuroscience (Botvinik-Nezer et al. 2020), social and behavioral sciences (Tyner et al. 2026; Miske et al. 2026; Aczel et al. 2026), economics and political science (Huntington-Klein et al. 2021; Breznau et al. 2022; Brodeur et al. 2026), and finance (Menkveld et al. 2024).

and estimation procedures before they can distinguish specification sensitivity from implementation differences. These costs are particularly substantial when studies combine multiple datasets (Whited 2023), and replication can remain difficult even with author-provided code and assistance (Chang and Li 2022; Tyner et al. 2026; Miske et al. 2026; Aczel et al. 2026). Given these challenges, many branches of academic research lack a culture of replication (Harvey 2017).

More recently, Muravyev (2026) demonstrates that large language models (LLMs) can substantially reduce the cost of large-scale replication. Using LLMs to replicate and cross-validate many published empirical finance results, he reports an in-sample replication rate of about 75%, with substantially lower out-of-sample replicability. Successful in-sample replication, however, does not establish that a finding generalizes beyond the selected specification. Conversely, out-of-sample performance can reflect specification selection, but also changes in economic relationships, sample composition, or institutional conditions. Importantly, it leaves an complementary question unanswered: conditional on recovering a published result, how sensitive is that result to the defensible empirical choices within its original setting?

LLM-based replication replaces the bottleneck of organizing and funding human research teams, but may introduce other costs when replicating or evaluating alternative designs, requiring careful validation. Evidence suggests cross-validation between LLMs introduces several biases and may be difficult due to the high correlation of errors between models.[2] Further, repeated generation and validation can add high computational costs (Bai et al. 2026; Oviedo et al. 2026). However, even accurately implemented replication and out-of-sample tests leave open the question of where a published specification lies relative to economically defensible alternatives.

More fundamentally, the same technology that lowers the cost of replication also lowers the cost of specification search. LLMs can help researchers generate, modify, and estimate alternative specifications at scale. Under publication incentives that reward significance, cheaper search can increase the opportunity to select favorable results from a larger set of empirical choices. This concern arises even when the resulting code executes correctly and reproducibly: replicating a selected specification does not reveal the alternatives considered or the sensitivity of the reported finding. Specifically, LLMs may enable researchers to search for specifications based on statistical significance, with little evidence that LLMs can identify such behavior because of their probabilistic nature and the difficulty of recreating final-product, LLM-generated code (Asher et al. 2026; Novy-Marx and Velikov 2026; Sun 2026; Thomas, Gligoric, and Shah 2026). Therefore, cheaper replication does not eliminate the need to systematically scrutinize researcher degrees of freedom.

[2]Kohli (2026) identifies a high correlation in errors among LLMs in judgment panels, with similar evidence for LLMs performing and evaluating coding tasks (Jiang et al. 2026; Jin and Chen 2026). Additionally, artifacts, ordering, self-preference, and other biases negatively affect LLM robustness (Chen et al. 2024; Chen and Goldfarb-Tarrant 2025; Shi et al. 2025; Wang et al. 2025; Wataoka, Takahashi, and Ri 2024).

Researchers have established the importance of looking beyond a single specification. Meta-analysis studies examine publication incentives using the distribution of published estimates and statistical significance across papers.[3] A related response is to require stronger statistical evidence. Harvey, Liu, and Zhu (2016) argue that extensive testing undermines conventional significance thresholds. Harvey, Sancetta, and Zhao (2026) argue that a minimum t-statistic of three is necessary for credible factor discoveries, while Ringgenberg and Samadi (2026) estimate minimum thresholds of 3.0 to 3.6 to control the false discovery rate in frequently reused natural experiments in corporate finance. While these studies examine the statistical evidence required to establish credible findings in the presence of multiple testing and publication selection, our approach directly evaluates the sensitivity of published findings to alternative empirical choices within their original settings. The importance of assessing specification sensitivity dates back at least to Leamer (1983). Related methods, including specification checks and specification curves, systematically evaluate alternative analytical choices (Brodeur, Cook, and Heyes 2020b; Simonsohn, Simmons, and Nelson 2020). In asset pricing, Novy-Marx and Velikov (2024) propose a standardized protocol and accessible tools for evaluating potential cross-sectional return predictors. Their approach shares our objective of making systematic scrutiny of empirical findings easier to implement. Our approach complements these efforts by independently reconstructing published studies across an empirical literature and making their underlying design choices explicit and systematically modifiable.

In this paper, we develop a scalable approach to literature-level stress testing using the declarative econometric framework introduced by Tumarkin (2026). We first independently replicate published studies from raw data. We then systematically replace selected design choices with alternatives drawn from the same literature while retaining the remainder of each specification. Conventional robustness tests evaluate alternatives selected around an individual paper's baseline. Literature-level stress tests instead evaluate choices that researchers actually make across studies, including assumptions underlying common practice. Our approach of replicating each paper's baseline results and then drawing alternatives from observed practice allows us to examine both the variation that those choices generate and where the published specification appears within that variation. The resulting distributions reveal whether a published finding is robust to those choices

[3]For example, Brodeur et al. (2016) identifies a "two-humped" pattern in published p-values that suggests authors systematically inflate marginally insignificant results to marginally significant; Brodeur, Cook, and Heyes (2020a) finds evidence of p-hacking in the use of causal inference methodologies; Brodeur, Kattan, and Musumeci (2026) finds evidence that publication bias leads to recruitment bias in the academic job market; Chang, Gao, and Li (2025) find larger discontinuities around significance thresholds among studies with more researcher degrees of freedom in the accounting literature; Andrews and Kasy (2019) propose adjusting estimates and confidence intervals for the estimated degrees of selection; Brodeur, Cook, and Heyes (2020b) and Simonsohn, Simmons, and Nelson (2020) propose analysis that plots the distribution of estimates and p-values generated by potential control variable inclusion or variable measurement; Chen and Zimmermann (2020) develops an estimator of publication bias; and Mitton (2022) documents significant variation in empirical methodologies within corporate finance, and discusses potential solutions proposed by the literature above.

and whether its baseline specification systematically favors the significance of the coefficient on its key economic explanatory variable. Importantly, we do this within each paper's original empirical setting.

Foghorn, a newly developed declarative econometric technique from Tumarkin (2026), makes this exercise tractable by representing an empirical design directly. In traditional empirical research, researchers implement empirical designs through a sequence of data-processing and estimation commands. In declarative econometrics, researchers do not write these steps. Instead, the code consists of the econometric specification and variable definitions.[4] *Foghorn* determines the data-processing and estimation steps from its specification, generating the code required to implement tests. Whereas the statistical nature of LLMs means that a single LLM can produce different econometric procedures from identical prompts, *Foghorn* is deterministic; identical inputs produce identical outputs. Thus, an econometric specification in *Foghorn* always produces the same implementation, ensuring reproducibility. Critically, because the specification is short and declarative, humans can review its assumptions at a glance without searching through implementation details.

Stress testing in *Foghorn* is materially different from doing so in standard econometric work. In *Foghorn*, changing a variable definition or sample restriction is a targeted design modification at the specification level. This has material benefits at the literature level. A declarative stress test will work for all papers in the literature, freeing researchers to focus on the test's content rather than its implementation. For example, with standard techniques, a stress test that changes all scaled variable definitions (e.g., normalizing by sales instead of assets) would need to be written for each paper in a literature since the data processing steps may differ. *Foghorn* can instead match appropriate variable definitions, apply a renormalization function, and determine data-processing steps with a single function that covers every paper. AI can even help write baseline declarative *Foghorn* specifications and stress-testing functions, but implementing alternatives does not require repeated LLM code generation.

The contribution extends beyond reducing replication and stress-testing costs. *Foghorn* provides a way for authors to communicate empirical assumptions and the relevant researcher degrees of freedom succinctly to referees, editors, and readers. These users can then inspect the design and evaluate whether a result survives defensible alternatives or depends on a narrow set of favorable choices. If specification dependence becomes easier to observe and influences publication decisions, selectively searching for significance becomes less rewarding. Declarative econometrics

[4]See Example 1 on page 17 for an example of declarative econometrics in *Foghorn*. *Foghorn* source code is available as a public preview at github.com/tumarkin/*foghorn*. Tutorials, demonstration videos, and code examples are in development. The transpiler that transforms declarative *Foghorn* code into data processing and estimation procedures will be released shortly. Please contact the author for access to this feature before its public release.

thus provides a mechanism that can counteract the exploitation of publication incentives: while LLMs lower the cost of specification search, *Foghorn* lowers the cost of specification scrutiny.

We demonstrate the feasibility and relevance of this mechanism by applying our approach to the mergers and acquisitions (M&A) literature. This setting is well-suited for stress testing. It combines a common empirical framework with substantial implementation discretion. Studies typically analyze announcement effects on shareholder value using cumulative abnormal returns (CARs), draw on a standard set of databases, and employ related event-study methods. Despite this, 80% of papers in our replicated sample use a unique CAR construction, differing in announcement windows, expected-return models, or estimation windows. This variation, consistent with the broader methodological heterogeneity documented by Mitton (2022), provides a set of observed alternatives with which to evaluate published specifications.

Our sample comprises papers in four leading finance journals that analyze CARs around M&A announcements. For each paper, we identify the key (focal) result stated by the authors and encode its estimator, variable definitions, and sample selection from the published description. We replicate from raw data without author-provided code, assessing success through similarity in focal estimates and statistical significance, sample size, and control-variable estimates. We do not impose editorial judgment as to what constitutes the best methodology (for example, Kolari, Pynnonen, and Tuncez (2021); Hou, Xue, and Zhang (2020)). Given the limited space allocated to describing the data manipulation process in research papers, authors may elide assumptions, particularly those that qualify as standard practices or that, prima facie, appear to be of minor relevance to the final results.[5] When the published description leaves implementation choices implicit, reconstructing the analysis using *Foghorn* requires making those assumptions explicit. The resulting declarative specification records the choices used to recover the baseline and provides an inspectable starting point for stress testing.

Our sample construction also illustrates the constraints on independently evaluating an empirical literature. Of 207 candidate M&A studies, 47 satisfy our data-access, methodological, and reporting requirements. Among the 19 eligible papers attempted to date, we independently reconstruct 15, yielding a replication rate of approximately 79%. This rate is similar to the approximately 75% in-sample rate reported by Muravyev (2026), although the samples and methods differ. In our experience, replication does not come easily, which may explain the mixed findings in the existing literature.[6] While replication of a published result is valuable, it does not establish the robustness

[5]For example, it is common practice in corporate finance research to substitute zero for missing values of capital expenditures, deferred taxes, research and development, and other fiscal variables (for example, Bryzgalova et al. (2024); Koh and Reeb (2015)). Yet, not all authors may mention that missing values have been replaced with zeros in their paper.

[6]Ioannidis (2005) and Chang and Li (2022) document significant replication challenges in science and economics, respectively — only 49% of economics papers in the latter study could be replicated even with author-provided code

of the result to alternative empirical choices or rule out every implementation error. Recent asset-pricing evidence likewise shows that replicable findings can be sensitive to plausible portfolio-construction decisions (Hasler 2023; 2025).[7] These findings motivate examining specification dependence beyond portfolio construction. Therefore, our stress tests begin where replication ends: they examine what happens when we modify the choices underlying a successfully replicated result.

We stress test the replicated M&A studies along three dimensions: the construction of the dependent variable (CAR), definitions of commonly used control variables, and sample selection. Similar to Hasler (2023) and Hasler (2025), we allow definitions to vary within the literature; that is, alternatives reflect definitions used within the sample of replicated papers. However, we take a more conservative approach than Brodeur, Cook, and Heyes (2020b) suggest. Specifically, we vary one dimension at a time, without searching over the inclusion or exclusion of controls or changing the original sample period. In the variable-definition tests, we retain the original empirical setting and vary the measurement choices. The sample-selection tests change inclusion criteria within that setting. These exercises complement out-of-sample validation by assessing sensitivity to researcher choices rather than moving the analysis to a new period.

Varying CAR windows, expected-return models, and parameter estimation periods generate more than 4,500 stress tests across 15 replicated specifications. These alternatives produce substantial dispersion in the magnitude of t-statistics of focal variables of interest, but the changes also systematically favor the published results. The replicated baseline t-statistic exceeds its stressed counterpart in magnitude in 80% of tests, with an average decline of approximately 0.8. In 56% of tests, the focal result moves to a lower conventional significance category; in 37%, it loses significance even at the 10% level. Individual paper analysis shows that the baseline advantage we observe in aggregate is broadly distributed across studies, although its magnitude and direction vary across papers. Our specification curve analysis indicate that the aggregate pattern is not attributable to one particular CAR construction. Additional analyses show that the sensitivity extends to estimated coefficients and their economic magnitudes. Similar patterns persist when we vary only CAR announcement windows while holding the expected-return model and estimation period fixed.

Dispersion alone, or bias towards a published empirical design, does not establish selective specification choice. While observed practice in the literature supplies a set of plausible alternatives, a choice used elsewhere need not be equally appropriate for every paper's economic question. If

and author assistance. In contrast, Jensen, Kelly, and Pedersen (2023) find stronger evidence of replicability in the finance asset pricing literature.

[7]Specifically, Hasler (2023) and Hasler (2025) vary the portfolio construction methodology of 92 cross-sectional stock return predictors and the Fama and French (1993) HML portfolio within the literature to construct plausible alternative portfolios. He shows that plausible alternative portfolio constructions materially affect return predictors and the HML factor, with published constructions producing stronger results than alternatives.

the results above were due to our methodology or the use of inappropriate alternatives, we would expect that non-focal control variables should exhibit similar behavior to focal variables. Therefore, we examine non-focal control coefficients under the same CAR stress tests. Their t-statistics also vary substantially. Inconsistent with the alternative explanation, the systematic advantage of the published design is significantly smaller. Baseline t-statistics exceed stressed alternatives in magnitude in only 61% of control-variable tests, compared with 80% for focal variables. Further, the average decline in the t-statistic (0.34) is significantly smaller than for focal variables (0.78), and their distributions differ significantly. This comparison shows that while changing CAR definitions affects the dispersion of statistical significance throughout a regression, we observe the baseline advantage more strongly in the coefficient central to the paper's hypothesis. This is consistent with specification selection associated with publication incentives.

We next examine whether stronger significance requirements address this sensitivity. Motivated by proposed increases in statistical hurdles (Harvey, Sancetta, and Zhao 2026; Ringgenberg and Samadi 2026), we first examine findings with absolute baseline t-statistics below three. These findings also exhibit substantial specification sensitivity, with 46% of stressed alternatives losing significance at the 10% level. We then use the observed distribution of baseline–alternative t-statistic differences to assess the scope for hypothetical threshold crossings under higher hurdles, including benchmarks calculated under an assumed publication-selection process following Andrews and Kasy (2019). Higher hurdles reduce the proportion of observed differences large enough to bridge these thresholds, but substantial scope for such crossings remains. Stronger significance requirements and systematic scrutiny of specification choices therefore provide complementary safeguards.

Our remaining stress tests show that this sensitivity extends beyond how the outcome variable is constructed. Varying the definitions of two common control variables, market-to-book and leverage, produces smaller but meaningful changes in focal significance that tend to favor the published design. Varying sample-selection criteria for deal form, deal status, and relative acquisition size also produces consistent results. Focal t-statistics decline in magnitude in about 75% of these tests, with an average decline of about 0.82; 53% decline to a lower significance category, and 39% lose significance at the 10% level. The systematic advantage of both is again significantly smaller for non-focal controls. Together, these findings show that seemingly innocuous choices about outcomes, controls, and sample inclusion can materially affect the significance of independently replicated results.

Our paper contributes to the general economics and social science literatures by documenting both substantial sensitivity to researcher degrees of freedom and a systematic advantage for the specifications that entered the published literature. This goes beyond showing that defensi-

ble choices can generate different results. We replicate published baselines and measure how alternatives used within the same literature affect their focal results, extending evidence on methodological discretion (Mitton 2022; Menkveld et al. 2024) and complementing large-scale replication (Muravyev 2026). The stronger baseline advantage for focal variables of interest is consistent with specification selection associated with publication incentives. Importantly, our design does not identify deliberate p-hacking by individual authors or causally isolate the publication process. Instead, we establish that high replicability can coexist with substantial specification dependence that disproportionately favors published findings. Therefore, our findings complement concerns that conventional significance thresholds provide insufficient protection against selection in published research (Harvey, Liu, and Zhu 2016; Harvey, Sancetta, and Zhao 2026; Ringgenberg and Samadi 2026). We show that successfully reconstructing a published result does not establish that its statistical significance is stable across alternative empirical choices.

Our paper also methodologically contributes to the empirical literature by making this specification dependence transparent, succinct, and systematically evaluable. Independent replication establishes a baseline and literature-level stress testing reveals how that baseline compares with observed alternatives. Finally, declarative econometrics makes the underlying choices concise and modifiable. Together, these elements give authors a means to communicate robustness and give referees, editors, and readers a means to assess it without reconstructing an entire implementation. By enabling scrutiny of a disclosed set of defensible choices, a knife-edge significant result becomes easier to distinguish from a finding supported across specifications. This process reduces the returns to selective specification search and becomes increasingly relevant as AI reduces the cost of empirical analysis.

Declarative econometrics is not intended to replace existing techniques. Writing bottom-up code builds invaluable knowledge of the nuances and idiosyncrasies in the data, and there will always be a role for it. Rather, declarative methods complement traditional approaches. Its comparative advantage lies in providing a transparent, succinct, scalable, and systematic way to encode, replicate, and stress test empirical analyses. The framework can extend beyond M&A as modules for additional data sources and methods are developed. Its value is therefore both empirical and institutional: it enables researchers to measure sensitivity within a literature and provides a practical way to make that sensitivity visible during research evaluation. Rather than relying on the reproducibility of a selected result alone, the research community can also examine its robustness to the choices available to its authors.

The remainder of the paper is organized as follows. Section 2. defines declarative econometrics and illustrates its relationship to stress testing. Section 3. describes our sample, defines replication

success, and presents the M&A replication results. Section 4. presents the stress tests. Section 5. concludes.

2. Declarative Econometrics: An Illustrative Model and Implementation

Declarative econometrics takes a novel approach to empirical research that is discussed in detail in Tumarkin (2026). In this section, we first begin by presenting an illustrative model to demonstrate the importance of this methodology to stress testing. We next discuss how this is implemented empirically by presenting a brief example.

2.1. A model of declarative econometrics

The declarative econometric language of Tumarkin (2026) enables researchers to write high-level, top-down specifications and convert them into data manipulation processes and estimation procedures. For our purposes, an econometric specification, $\mathcal{S}$, fully defines an empirical methodology, including variable definitions, the modeled function form, and the estimation technique. Notably, our definition of a specification does not include the steps required to transform raw data into a dataset used for estimation, which we collectively refer to as an empirical procedure, $\mathcal{P}$.

Formally, the declarative econometric language $\mathcal{L}$ consists of two components: (1) a notation system $\mathcal{N}$ that provides syntax and semantics, and (2) an implementation $\mathcal{I}$ that algorithmically converts a specification into an empirical procedure (i.e., the data processing and estimation procedure):

$$\mathcal{I} : \mathcal{S}^{\mathcal{N}} \longrightarrow \mathcal{P}, \tag{1}$$

where $\mathcal{S}^{\mathcal{N}}$ denotes a specification $\mathcal{S}$ written in notation system $\mathcal{N}$. Thus, the declarative econometric language $\mathcal{L}$ is the pair of the notation and the implementation:

$$\mathcal{L} = \{\mathcal{N}, \mathcal{I}\}. \tag{2}$$

For example, consider an ordinary least squares regression where the researcher wishes to examine how cumulative abnormal stock returns relate to the market capitalization of the acquiring firm. The specification $\mathcal{S}$ consists only of the variable definitions, the linear functional form, the choice of OLS as the estimator, the sample selection criteria, and the choice of standard errors. The corresponding procedure $\mathcal{P}$, by contrast, must define the steps that construct cumulative abnormal returns, market capitalization, and other control variables from raw data. The empirical procedure links observations across datasets, combining the data into a final dataset, and estimating the parameter vector on that dataset.

In standard research, the specification consists of the information described in the research paper. The procedure is the complete implementation code from raw data. Due to commercial and proprietary data restrictions, authors may not be able to provide the whole procedure to other

researchers, even in online form. Instead, code may start with pre-processed data to satisfy licensing restrictions.

While the specification and procedure are related, they have traditionally been expressed in fundamentally different formats. Specifications are described qualitatively in a paper's text, whereas procedures are implemented in code, leaving no formal link between the two. As a result, errors in the procedure are not visible in the specification or to readers of the paper. The absence of a programming language for specifications underlies this disconnect. A declarative econometric language resolves the problem by allowing the researcher to state the specification directly, with the implementation translating this high-level description into the complete empirical procedure.

2.1.1. Notation

The notation system $\mathcal{N}$ needs to allow researchers to describe a specification's variables, functional form, and estimation technique. It should form a complete algebra whereby different types of operations may be composed transparently through a uniform syntax. This complete algebra provides succinct, top-down semantics to define variables. Operations such as arithmetic, data aggregation, index conversion, data merging, time transformations, and other necessary data manipulation tasks work in harmony. Whereas existing packages let researchers manipulate data through a series of steps, they isolate different types of operations. By contrast, a declarative language provides a way to compose such operations, compactly describing every test variable in terms of raw data.

The notation system also provides semantics to describe estimators. Most software packages already use declarative syntax to define estimators. For example, in SAS, Stata, or R, the researcher uses a command to describe the specification but does not instruct the package on how to perform the required computations.

While the estimator description in a declarative language has a similar syntax to that in standard econometrics packages, it is functionally more powerful. A declarative language enforces a *whole-specification* approach to writing econometric tests. As detailed below, declarative econometrics elevates econometric information from the data to the language. The language's compiler can then reason about a specification logically as a whole, rather than evaluating a specification simply as a sequence of programming steps. A declarative econometric language validates that a specification is internally consistent before execution, ensures that all operations that comprise variable definitions are correct, and infers desired data transformations. The language's ability to ensure specification integrity has clear benefits for writing econometric tests, in general. However, the power of language-level integrity checking is particularly evident in stress testing.

2.1.2. Empirical procedure

The declarative language implementation $\mathcal{I}$ converts the empirical specification $\mathcal{S}^{\mathcal{N}}$ into the empirical procedure $\mathcal{P}$. The empirical procedure can be seen as a mapping from the universe of raw data, $\mathcal{D}_0$, to an estimated parameter vector, $\hat{\theta}$:

$$\mathcal{P} : \mathcal{D}_0 \longrightarrow \hat{\theta}. \tag{3}$$

Without loss of generality, we consider the implementation of empirical specification to consist of two broadly defined steps. First, a data manipulation (generation) process, $\mathcal{M}$, transforms raw data into the processed data required for estimation. Second, an estimation procedure, $\mathcal{E}$, computes the coefficient vector, standard errors, and other statistics of interest from the processed data.

Let the data manipulation process $\mathcal{M}$ have n steps where the data after i steps is $\mathcal{D}_i$. An individual step, $\mathcal{M}_i$, in that data process transforms the data from $\mathcal{D}_{i-1}$ to $\mathcal{D}_i$,

$$\mathcal{M}_i : \mathcal{D}_{i-1} \longrightarrow \mathcal{D}_i. \tag{4}$$

The cumulative process $\mathcal{M} : \mathcal{D}_0 \longrightarrow \mathcal{D}_n$, which converts the raw data into the final data needed for estimation, is the sequential composition of the individual steps:

$$\mathcal{M} = \mathcal{M}_n \circ \mathcal{M}_{n-1} \circ ... \circ \mathcal{M}_2 \circ \mathcal{M}_1. \tag{5}$$

We note that mathematical composition associates to the right – the rightmost transformation is applied first, followed by those to its left.

Finally, the estimation procedure $\mathcal{E}$ maps the final data into the coefficient vector:

$$\mathcal{E} : \mathcal{D}_n \longrightarrow \hat{\theta}. \tag{6}$$

Thus, we can view the empirical procedure $\mathcal{P}$ as the composition of the estimation procedure and the data manipulation process:

$$\mathcal{P} = \mathcal{E} \circ \mathcal{M}. \tag{7}$$

2.1.3. Stress testing

Stress testing an economic literature is a principal innovation from declarative econometrics. Stress testing a literature is fundamentally different than performing robustness tests on an individual paper. While a robustness test typically examines a paper's assumptions that differ from common practice, literature-level stress testing examines the assumptions that underlie common practice. A stress test $\mathcal{T}$ may be considered a transformation of a base empirical specification $\mathcal{S}$ to a new one, $\mathcal{S}'$:

$$\mathcal{T} : \mathcal{S} \longrightarrow \mathcal{S}'. \tag{8}$$

Researchers have not yet developed a systematic approach for literature-wide stress testing. Such stress tests must be scalable both in the range of tests they can accommodate and in their ability to apply those tests consistently across specifications. Equally important, they must ensure correctness, regardless of stress-test complexity.

Stress testing using existing methods

With existing technologies, researchers cannot manipulate the specification directly as per Equation 8. Instead, they must work with the data manipulation and estimation procedures that together comprise the empirical method using either a manual or procedural approach.

A manual approach follows traditional economic practice: the researcher applies a stress test transformation to an original procedure $\mathcal{P}$, derives the stressed procedure $\mathcal{P}'$, and constructs the corresponding data-management process $\mathcal{M}'$ and estimation procedure $\mathcal{E}'$. While feasible for a few stress data-generating and estimation processes, this strategy does not scale; the effort required to stress test a literature increases multiplicatively with both the number of tests and the number of projects, and remains vulnerable to human error.

Alternatively, a researcher may pursue a procedural approach, directly modifying the data-manipulation process and estimation procedures:

$$\mathcal{T}^{\mathcal{E},\mathcal{P}} : \mathcal{E} \circ \mathcal{M} \longrightarrow \mathcal{E}' \circ \mathcal{M}'. \tag{9}$$

However, like manual stress testing, the procedural approach is neither scalable nor reliable, underscoring the need for a systematic framework. This strategy will not be effective across empirical procedures. Since coding choices of data-manipulation processes and estimation procedures vary across papers, a modification that applies to one implementation is unlikely to translate to another.

The practical burden goes beyond locating the command that defines a variable. A change in its definition may require revising intermediate calculations, data merges, missing-value treatments, and sample restrictions that depend on it. Researchers must trace these dependencies and update the affected steps while preserving choices unrelated to the intended stress test. Across papers, the same empirical quantity may be constructed from different data sources, calculated in a different order, or embedded in different intermediate variables. Therefore, applying a common stress test requires understanding each implementation and, more importantly, verifying that the modifications represent the same intended design change. Even when every baseline implementation is correct, this repeated reconstruction makes literature-level stress testing costly and creates opportunities for unintended changes.

Furthermore, this process is unlikely to succeed even when procedural implementations share steps. Consider two procedures $\mathcal{P}_\mathcal{A}$ and $\mathcal{P}_\mathcal{B}$, with corresponding data manipulation processes $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{B}$. Suppose both data manipulation processes contain a common step $\mathcal{R}$ that appears some-

where in the data manipulation pipelines: $R \coloneqq \mathcal{M}_{\mathcal{A},i} = \mathcal{M}_{\mathcal{B},j}$. Function composition, however, suggests that the role of $\mathcal{R}$ within procedure $\mathcal{A}$ need not match its role with procedure $\mathcal{B}$. Consequently, knowing that a stress test $\mathcal{T}^{\mathcal{E},\mathcal{P}}$ correctly transforms $\mathcal{P}_{\mathcal{A}}$ into $\mathcal{P}_{\mathcal{A}'}$ does not imply that it will do the same for $\mathcal{P}_{\mathcal{B}}$, even if the two procedures share steps:

$$\mathcal{T}^{\mathcal{E},\mathcal{P}} : \mathcal{P}_{\mathcal{A}} \to \mathcal{P}'_{\mathcal{A}} \nRightarrow \mathcal{T}^{\mathcal{E},\mathcal{P}} : \mathcal{P}_{\mathcal{B}} \to \mathcal{P}'_{\mathcal{B}}. \tag{10}$$

Stress testing using declarative econometrics

The approach of Tumarkin (2026) offers an alternative, scalable approach to creating correct stress tests. Using this approach, a researcher is working with specification descriptions, not the data manipulation process and estimation procedure. It permits direct specification manipulation, providing a way to write stress tests in terms of modified specifications directly:

$$\mathcal{T}^{\mathcal{S}} : \mathcal{S} \to \mathcal{S}'. \tag{11}$$

This approach retain its descriptive style when transforming specifications. It focuses on variable and estimator definitions, and occurs within the standard language without needing to resort to macro-programming or similar techniques.

Since the language implementation converts a declarative stress test into the desired modified data manipulation process and estimation procedure, an implementation can take the specification that arises from a stress test and determine the correct procedure:

$$\mathcal{I} \circ \mathcal{T}^{\mathcal{S}} : \mathcal{S}^{\mathcal{N}} \to \mathcal{P}' = \mathcal{E}' \circ \mathcal{M}' \quad \text{for all } \mathcal{S}^{\mathcal{N}}. \tag{12}$$

This approach has numerous benefits. First, it is straightforward to conceptualize stress tests based on specifications. The researcher needs only to consider the econometrics of the test, determining the necessary changes to variable definitions and the estimator required for the stress test. By contrast, a stress test that manipulates bottom-up data processes and estimation procedures is more complicated to write, as one needs to consider the role of each step in the data pipeline.

At the specification level, the intended change can instead be expressed as a modification to a variable definition, sample restriction, or estimator. The language determines the corresponding implementation steps, including intermediate calculations that depend on the modification. Choices outside the transformation remain unchanged in the specification. This makes the scope of the stress test explicit even when the generated implementation requires several changes. Further, this distinction is important when the same transformation is applied across papers. researchers can define a common change in empirical design without writing a separate set of procedural edits for each set of code. Researcher will still need to determine whether an alternative is economically appropriate for each study. However, they do not need to reconstruct its entire implementation in order to do so.

Second, it emphasizes correctness of transformations, avoiding the function composition and bespoke code pitfalls associated with traditional methods. The whole-specification approach embedded within declarative econometrics means that, not only can the language reason about specifications, but it can also reason about stress test transformations. As a consequence, it can ensure that, given any declarative specification, the stress test transformation will return a valid modified version of the specification. Such transformations are called *total* mapping, also known as *total* function, and the language itself checks to ensure that all stress test transformations meet this criterion.[8]

Finally, this approach is scalable and can be applied across specifications. The implementation can transform any stressed specification into the corresponding stressed data manipulation process and estimation procedure. The stress tests themselves are scalable. As transformation functions, stress tests can be combined through functional composition. Thus, it is possible to build a complicated stress test from multiple simple tests. The language itself will guarantee that the derived, stressed specification is valid, and it will provide the corresponding bottom-up data processing process and estimation procedure.

2.2. Implementation

Traditional econometric code requires the researcher to specify each step of data manipulation explicitly. The researcher knows, for instance, that share prices are numbers and ticker symbols are text, but the programming language does not. A statement adding the two is syntactically valid in SQL, SAS, and other data manipulation languages; the error surfaces only at runtime. This disconnect between what the researcher knows and what the language knows limits how concise and expressive econometric code can be.

Declarative econometrics closes this gap. Instead of writing step-by-step instructions, the researcher writes a specification that describes the analysis: variable definitions and an estimator. The language itself determines the necessary data manipulation steps. This is possible because a declarative language can reason about the specification before execution, validating relationships among variables and inferring how to combine data from different sources. *Foghorn* separates logic from implementation, it is modular (separates its core language from dataset-specific information), and it is transparent.[9]

[8] A stress test may be inherently incompatible in certain empirical settings. For example, a stress that lags explanatory variable by a year would be incompatible with a cross-sectional dataset lacking a time dimension. Such incompatibility can be handled elegantly by total functions within a declarative language. Details are provided in Tumarkin (2026).

[9] *Foghorn* will be publicly available for use iby researchers. Technical details on the language design, including its type system and implementation in Haskell, are provided in the Internet Appendix and Tumarkin (2026).

Language design affects how readily empirical choices can be observed and evaluated. Complete procedural code can document assumptions explicitly, but recovering an empirical design from that code often requires tracing choices across data preparation, variable construction, and estimation steps. *Foghorn* organizes these choices at the specification level, providing a concise representation of the analysis whose implementation is generated automatically. This structure allows authors to communicate the assumptions underlying a result directly and enables referees, editors, and readers to identify which choices distinguish it from economically viable alternatives. This also provides impetus to the author to explicitly describe why certain assumptions are economically important. Because those choices can be modified within the specification, their consequences can also be evaluated systematically. Therefore, the advantage is not simply shorter code, but a closer connection between the empirical design authors communicate and the analysis others can inspect and stress test.

This methodology offers several benefits for empirical research. It organizes empirical assumptions explicitly within the specification and makes them systematically modifiable, enabling stress testing across an entire literature. It also provides two complementary levels of transparency. The declarative specification presents a concise description of variable definitions, sample restrictions, and the estimator. This allows reviewers to inspect the empirical design without needing to reconstruct it from procedural code. The generated implementation exposes the corresponding data manipulation and estimation steps, allowing reviewers to examine how that design is executed. We call this bi-directional transparency: a reviewer can move from a stated assumption to its implementation or trace implementation back to the specification that generated it. Therefore, authors can communicate their empirical choices succinctly while retaining a detailed implementation for verification. This transparency can also help reviewers detect discrepancies between stated and implemented data construction. Referees, editors, and readers can assess both whether the implementation corresponds to the stated design and how the findings change when selected assumptions are varied.

Declarative econometrics complements rather than replaces traditional methods. Working directly with data builds invaluable knowledge of its nuances and idiosyncrasies. However, for replication and stress testing at the literature level, a declarative approach provides capabilities that procedural code cannot match.

2.3. Examples of declarative econometrics in Foghorn

The examples below are intentionally minimal to highlight the differences from a traditional bottom-up approach. While they may not fully showcase the potential of working declaratively or the variety of tools currently available in *Foghorn*, they still offer valuable insights. We remind

readers that, although these examples focus on corporate finance, declarative econometrics is a general framework applicable across all areas of the social sciences.

2.3.1. The declarative econometric statement

Assume that a researcher wants to compute a market-to-book ratio for research on financial markets. The ratio requires two variables from the Center for Research in Security Prices (CRSP) monthly stock data, shares outstanding (`Crsp.shrout`) and share price (`Crsp.prc`), both of which are indexed by a permno cross-sectional firm identifier and a trading-date time-series identifier.[10] It also requires the book value of total assets variable (`Comp.at`) as of the fiscal year-end from the Compustat Fundamentals Annual data, which is indexed by a gvkey cross-sectional identifier and a datadate fiscal-year-end time-series identifier. Computing a market-to-book ratio in a bottom-up data manipulation language requires writing code to convert gvkey – datadate assets to a corresponding permno – trading-date asset time series. Then, an algebraic step does the actual calculation. The resulting code emphasizes the steps, rather than the algebraic relationship. It would generally be verbose, requiring many lines of precise table joins to transform the underlying data indices before the final algebraic step that computes the market-to-book ratio.

In a declarative language, the index conversion and calculation may be written in a single step. This calculation in *Foghorn* is coded as

$$\texttt{Crsp.shrout} \times \texttt{Crsp.prc} \;/\; \texttt{reindex}(\texttt{Comp.at}). \tag{13}$$

This top-down calculation emphasizes the underlying economic meaning by using a simple reindex function to tag which panel variable should have its indexes converted. The benefits of such succinct coding syntax grow substantially with variable complexity.

Foghorn can convert this top-level equation into bottom-up data manipulation code because it knows the indices of each variable (i.e., cross-sectional, time-series, or panel) at the language level, a feature known as static typing in programming language design. In the expression above, *Foghorn* identifies that `Crsp.shrout` and `Crsp.prc` are both indexed by permno-trading-date. Thus, their product is also indexed by permno-trading-date. *Foghorn* also knows `Comp.at` is indexed by gvkey-datadate. Therefore, because `Comp.at` divides `Crsp.shrout` $\times$ `Crsp.prc`, *Foghorn* infers that the `reindex` function must convert `Comp.at` from gvkey-datadate to permno-trading-date using a programming language feature known as type-inference.

[10]The term “indexing” refers to index variables, unique identifiers for individual observations within a dataset. These identifiers may be cross-sectional and/or time-series units. The term “reindexing” describes the procedure of aligning identifiers across different datasets, accurately matching an observation from a source dataset with its corresponding observation in a target dataset.

2.3.2. The declarative econometric specification

Consider a research project to analyze the effect of entrenched corporate boards on stock returns around acquisition announcements. The study will control for both the acquirer's and the target's Tobin's *q*. To keep this example concise, we will not include additional controls, time-constant or firm-constant fixed effects, or sample selection criteria. We will also not adjust or cluster standard errors. All these features are available in *Foghorn*, which also allows for organizing empirical datasets by any cross-sectional and/or time-series indices.

Example 1 shows the complete econometric code for implementing this empirical study in *Foghorn*. For brevity, we have omitted boilerplate statements that import *Foghorn* libraries.

Example 1

Declarative Econometric Specification

```
study = estimate
  [regHdFe| car ~ bcfIndex + acquirerTobinsQ + targetTobinsQ |]

-- Cumulative abnormal returns
car = acquirer $ cumulativeAbnormalReturn (-2, 2)
        `overModel` singleIndexModel (-210, -11) CrspEqualWeighted

-- Entrenchment index (Bebchuk, Cohen, and Ferrell (2009))
bcfIndex = acquirer
  (Risk.cboard + Risk.labylw + Risk.lachtr + Risk.gparachute + Risk.supermajor +
  Risk.ppill)

-- Firm value
acquirerTobinsQ = acquirer tobinsQ
targetTobinsQ = target tobinsQ
tobinsQ = (Funda.at - Funda.ceq + reindex (Msf.shrout * Msf.prc)) / Funda.at
```

The sample code in Example 1 differs noticeably from traditional bottom-up data management and econometric estimation techniques. Perhaps most strikingly, it is concise and direct. The code comprises only nine lines of statements; the remaining six lines consist of white space and comments. The specification requires only variable definitions, a functional form between dependent and independent variables, and the estimation technique. This code emphasizes the "what" of this empirical study, providing only the econometric definitions required to describe the specification.

Unlike traditional bottom-up research techniques, the declarative specification is reasonably

self-evident.[11] An interested third party does not need to review a step-by-step data manipulation process to understand the econometric methodology. Instead, this declarative specification begins with a top-down statement indicating that this study estimates a single regression model. The model's functional form specifies that observed cumulative abnormal returns are related to a hypothesized linear relationship among the acquiring firm's corporate board entrenchment index, the acquiring firm's market-to-book value ratio, and the target firm's market-to-book value ratio.[12]

The code then defines the variables in the model. The acquirer's cumulative abnormal return (CAR) is calculated over a five-day window, starting two days before and ending two days following the announcement. This calculation utilizes the CRSP Equally Weighted Index as a single index market model, with parameters estimated over a window beginning 210 days before and ending 11 days before the announcement. The Entrenchment Index (Bebchuk, Cohen, and Ferrell 2008) for the acquirer is the total of six indices from RiskMetrics (`Risk`): classified boards (`cboard`), limited ability to amend bylaws (`labylw`), limited ability to amend charter (`lachtr`), golden parachutes (`gparachute`), supermajority requirements (`supermajor`), and poison pills (`ppill`). Finally, the acquirer's and target's Tobin's q (lines 10 and 11, respectively) are built upon a general definition in line 12. This uses the book values of assets (`at`) and common equity (`ceq`) from Compustat Fundamentals Annual (`Funda`) data as well as stock price (`prc`) and shares outstanding (`shrout`) from the CRSP Monthly Stock File (`Msf`).

The specification's required data sources, including Compustat, CRSP, RiskMetrics, and SDC, are prominent in the specification.[13] Each of these employs different indexing variables to uniquely identify observations. While verbose data manipulation steps that convert variables across indices often comprise a substantial portion of code in traditional econometric work, they are noticeably concise in Example 1. The `acquirer` and `target` functions attribute data to the respective parties of an acquisition.

2.3.3. Declarative econometrics and the bottom-up process

The high-level declarative code above may seem insufficient for implementing an econometric study, particularly because there does not appear to be any bottom-up steps. Recall that in declarative econometrics, the researcher defines the empirical specification, with the language determining the implementation. Thus, an econometric study is well-defined provided the language

[11] We recognize that the language has some notational idiosyncracies (e.g., the dollar signs and back ticks), which arise from our decision to implement *Foghorn* as an embedded domain specific language. More information is provided in Tumarkin (2026).

[12] We have borrowed notation from *R* for estimators as we believe the use of algebraic operators makes the functional form clear.

[13] Unlike Compustat, CRSP, and RiskMetrics which have the explicit prefixes `Comp`, `CRSP`, and `Risk`, respectively, SDC data is referenced implicity by the `cumulativeAbnormalReturn`, `acquirer`, and `target` functions.

has sufficient information to convert the specification into a database construction process. *Foghorn* translates top-down specifications into standard database code, a process known as transpiling.[14] *Foghorn* analyzes the specification, determines the order in which to perform algebraic, aggregating, reindexing, time-shifting, and other calculations, and generates all the necessary database code for the researcher. For example, in Equation 13, *Foghorn* checks if this conversion is allowed (user-defined permissible index conversions through modules), and, upon execution, generates the appropriate bottom-up code in SQL or SAS, for instance.[15] Transpilation converts functions in *Foghorn* into their data manipulation equivalents. As will be discussed later, the design decision to provide transpiling has numerous benefits, particularly in terms of transparency.[16]

2.4. AI-assisted analysis and transparent research evaluation

Advances in AI-assisted empirical analysis make the distinction between replication and stress testing increasingly important. Muravyev (2026) demonstrates that large language models can substantially reduce the cost of replicating published finance research. This capability can improve independent verification and broaden the scope of replication exercises. However, replicating a published result establishes that its specification can be implemented with comparable findings. Replicating does not establish that the finding survives other defensible empirical choices. Out-of-sample tests provide additional evidence about whether a relationship persists in a different sample, but their outcomes can also reflect changes in sample composition or the economic environment. Literature-level stress testing addresses a complementary question by systematically varying empirical choices within the setting used to replicate (and generate) the original result.

Validating AI-generated implementations also requires distinguishing agreement from independent cross-validation. When LLMs have correlated errors, agreement between a model that generates code and one that evaluates it may provide less assurance than their apparent independence suggests. Recent evidence on correlated model errors and model-based evaluation is consistent

[14]Transpiling is a new concept to econometrics, but is an established technique in programming languages to add static type constraints to non-statically typed languages. For example, JavaScript is a weakly typed language, making it relatively easy for programmers to include bugs in web applications. Languages such as TypeScript and PureScript provide type safe languages for web development. These languages transpile to JavaScript, eliminating many classes of errors. *Foghorn* employs transpilation in a similar fashion, eliminating many types of errors in econometric coding. *Foghorn* has a plug-in architecture for transpiling, allowing it target different data manipulation languages. It currently has a SQL and STATA transpiler. SAS transpiler is under development.

[15]We note that, while it would be labor-intesive to convert the the market-to-book ratio into a panel variable indexed by *gvkey-datadate* in a standard data manipulation language, it is trivial to do so in a declarative language. One simply applyies the reindex function to the CRSP variables instead: `reindex` (`shrout` × `prc`)/ `at`.

[16]We anticipate that many readers will understand the main ideas behind writing top-down specifications, but be curious about the specific mechanics underlying how declarative code operates. Tumarkin (2026) provides a detailed discussion of this as well as more details. These include how a declarative econometric language can conduct an econometric test without bottom-up data manipulation steps, how a specification can work without explicitly describing how observations are matched between datasets, and whether the brevity of declarative code limits its applicability across the social sciences.

with this concern (Kohli 2026; Jiang et al. 2026; Jin and Chen 2026). Repeated generation and evaluation can provide additional checks, but also require computational resources and a clear account of what those checks establish. *Foghorn* offers an inspectable specification and a generated implementation that can be examined separately. This structure does not eliminate the need to validate the language, its data modules, or the economic assumptions supplied by users. However, it allows reviewers to directly examine the stated design and its implementation in a transparent and succinct representation without relying solely on agreement among LLMs.

The incentive problem extends beyond implementation accuracy. By reducing the cost of generating and estimating alternatives, LLMs can also reduce the cost of searching over researcher degrees of freedom for significant results. When publication rewards statistically significant findings, cheaper search can make it easier to identify a favorable specification among many defensible alternatives. The selected analysis may be correctly coded and fully reproducible while remaining highly sensitive to the choices that produced it. Neither successful replication nor the absence of coding errors reveals that sensitivity. Therefore, it is important to scrutinize specification choices and verify their implementation to evaluate the reliability of empirical findings.

Declarative econometrics can reduce the cost of this scrutiny. A *Foghorn* specification allows authors to communicate variable definitions, sample restrictions, and estimation choices succinctly, while the generated implementation provides the corresponding procedural detail. Referees, editors, and readers can inspect those choices and systematically evaluate alternatives relevant to the economic question. If authors anticipate that specification dependence will be visible and will influence research evaluation, the expected return to selectively searching for significance may decline. This mechanism depends on disclosure, access to the necessary data, and meaningful scrutiny of economically appropriate alternatives. A disclosed baseline specification does not itself reveal every alternative an author considered, and authors can also selectively report stress tests. Therefore, the benefit of transparency is greatest when the alternatives and transformation rules are disclosed and others can extend the analysis. Subsequent research can also assist in specification scrutiny. However, the reconstruction costs associated with this research, combined with limited access to certain data and opaque choices can impede it. *Foghorn* potentially lowers those costs.

Stress tests inform editorial judgment rather than supply a mechanical acceptance threshold. The proportion of alternatives retaining significance depends on the alternatives examined, their economic relevance, and their dependence, so no universal survival rate determines whether a finding is credible. Referees should instead assess whether the direction and economic magnitude of the focal effect remain stable, which choices drive sensitivity, and whether those choices have a convincing economic justification. Therefore, sensitivity can motivate clarification or qualification of a claim rather than rejection.

Consequently, *Foghorn* and AI-assisted research can serve complementary roles. LLMs can help draft declarative specifications, identify methodological variation across papers, and propose alternatives for investigation. Researchers must assess those proposals for economic relevance. Finally, *Foghorn* provides a setting where the selected transformations can be implemented and inspected. The separation between proposing alternatives and executing disclosed design changes supports both productive stress testing and independent evaluation. Our M&A application demonstrates the feasibility and relevance of this scrutiny.

3. Replicating the M&A Literature

We employ our declarative econometric technique to stress test the merger and acquisition (M&A) literature. M&A provides an ideal setting for stress testing, given the importance of M&A to corporations. The literature is mature, having been extensively studied over an extended period, and has established results across numerous papers. Papers draw information from a common set of databases, and, although the event study methodology is an accepted framework, critical assumptions vary across papers.

3.1. Sample selection

Our sample selection process begins with papers published in the four journals with arguably the highest impact factors in academic finance: *The Journal of Finance*, the *Journal of Financial Economics*, *The Review of Financial Studies*, and the *Journal of Financial and Quantitative Analysis*. We identify those papers that analyze returns around M&A announcements, yielding 207 candidates for replication published between 1983 and 2023. The early 1980s start of this period coincides with the nascent academic empirical research into M&A documented by (Mulherin, Netter, and Poulsen 2017). We eliminate studies with purely theoretical contributions and those whose analysis is primarily descriptive.

The inclusion criteria derive from both our replication and stress testing objectives. We first impose a data requirement; we must have access to a paper's data to replicate its results. The candidate paper must use publicly available data, such as that from the Bureau of Labor Statistics and academic websites, or commercially available data to which we have access. As a result, we restrict our sample to papers that use the following commercial data providers: S&P Compustat, Center for Research in Security Prices, Refinitiv Thomson, MSCI RiskMetrics, and SDC Platinum.

Stress testing at the literature-level is the examination of assumptions that underlie common practice. Therefore, we select those papers that exemplify typical empirical specification design and methodology. We limit our sample to those papers where (1) the benchmark analysis, as stated in each paper, uses Cumulative Abnormal Returns (CARs) as the dependent variable and (2) the primary estimation technique involves ordinary least squares or panel regressions. As we cannot

stress test null results, we examine only those benchmark analyses that find a statistically significant result with a coefficient in the direction hypothesized.

Finally, we have practical considerations that eliminate some papers from our sample. The paper must describe the methodology well enough for us to replicate the key findings. In some cases, the paper may not adequately define a variable, requiring that we make assumptions to replicate the result. Examples include financial variable definitions and industry classification levels. In such cases, we revert to definitions used by other authors in the literature. These replication assumptions are documented in the online replication code.

After applying these criteria, our sample consists of 47 candidate papers. This dataset requirement restricts the candidate studies to those that examine M&A within the United States. It also eliminates those studies that use proprietary or hand-collected datasets in their baseline findings. The earliest papers in our sample use hand-collected data from the Wall Street Journal, Grimm's Mergerstat Review, and other sources. Thus, our candidate papers begin with those published after 2000, when data sources became more standardized.

3.2. Defining Replication Success

There is no universally accepted definition of successful replication. We define replication as independently reconstructing a published empirical result from raw data using the methodology described in the article, without author-provided code. This approach provides an independent implementation of the reported analysis, rather than a re-execution of the authors' implementation. Our objective is to replicate a sufficiently comparable baseline, which will allow us to assess sensitivity to alternative empirical choices.

It is probably not possible to replicate papers exactly. Data providers update information, often restating historic data to improve accuracy (Ljungqvist, Malloy, and Marston 2009). Lyle, Siano, and Yohn (2025) find that Compustat's periodic "standardizations significantly alter key financial figures such as sales and earnings, among many others, leading to material differences in research findings." Researchers also may elide essential steps in their data manipulation process, making precise replication difficult. For example, authors generally do not precisely describe the process to match observations across databases, especially with text-based matching. Authors may also omit their process for dealing with missing financial data, such as research and development expenses. Lyle, Siano, and Yohn (2025) assert that "precise replication of prior studies using common Compustat products is nearly impossible."

Reproduction, as opposed to replication, is the independent execution of a study performed using data and code provided by the authors. Reproduction should be easier than replication, yet researchers have experienced difficulty reproducing results. Chang and Li (2022) attempt to

reproduce 67 macroeconomic papers. Defining success as one that produces the "key qualitative results of the paper," they can reproduce 33% of the sample without contacting the authors. Fišar et al. (2024) assess the reproducibility of nearly 500 articles in *Management Science*. A "fully reproduced" paper is one in which the reproduction generates the exact numerical results as the published article. A "largely reproduced with minor issues" paper is one in which the reproduction has small differences from the published article. They find that only 45% of articles that voluntarily provided data and code were reproducible before the journal's 2019 institution of a data and code disclosure policy.

We assess replication success using two qualitative criteria: (i) the coefficient on the focal economic variable has the same sign and a similar magnitude and statistical significance as the published estimate, and (ii) the replicated sample contains a similar number of observations. Agreement in control-variable estimates provides an additional diagnostic, but is not a separate requirement for replication success because differences can arise from the data revisions and implementation ambiguities discussed above.[17] These criteria establish whether we have recovered a baseline comparable to the published result, but do not establish that the result is robust to alternative specifications. We assess this directly through stress tests in Section 4.

3.3. Replication Process

We replicate within *foghorn*, focusing on a single result examining cumulative abnormal returns in the target paper. The specifications selected for replication are typically identified by the author as a key result supporting the paper's hypotheses. We use a single, principal result to keep the scope of our replication manageable and to ensure that stress testing is comparable across papers.

We use the empirical methodology and assumptions as described by the authors to reimplement a paper. In some cases, a paper may not provide sufficient detail. Common issues include the handling of missing data, variable definitions, and sample selection. For example, many well-known financial variables are missing from Compustat (e.g., research and development expenses (Koh and Reeb 2015)), and the authors may not indicate if they drop observations or replace missing values with zero or industry averages. In other cases, a paper may not provide the individual variables used to create a composite financial control (e.g., the components used to create a total debt control variable). Papers may also not include the market index or model estimation window used when computing cumulative abnormal returns. Papers may also cite other research as a methodology source without clarifying whether sample selection and other decisions from the source apply. *Foghorn* takes supplied data as inputs, so its results remain conditional on any

[17] Our inability to replicate a result does not prove that it is not replicable. Rather, it suggests that our interpretation of the empirical approach did not yield similar results. Consequently, we do not list those papers whose results we were unable to replicate.

prior manual cleaning. Where raw and cleaned versions are available, their consequences can be compared.

These omissions create a barrier to independent verification even when the underlying databases are available. Access to the same commercial data sources does not uniquely determine the estimation sample or variable construction, and researchers must resolve ambiguities before assessing whether published findings can be recovered. Therefore, making these reconstruction choices explicit is an important contribution of *Foghorn*, separate from whether the resulting estimates satisfy our replication criteria.

Fortunately, these elided assumptions generally have precedent in the literature. We use common missing variable treatments, common variable definitions, sample selection, and other criteria as necessary, selecting the assumptions that yield the replication closest to the published result. Declarative econometrics is very helpful in efficiently systematizing this search process. For example, acquirer market capitalization is a common control in the M&A literature and may be sourced from Compustat, CRSP, and SDC in our sample of papers. If not specified, we can quickly switch the source for market capitalization without having to worry about integrating the data into the overall data management process.[18]

3.4. Results

Table 1 summarizes the results for the papers that were successfully replicated. For each study, it lists the key economic variable of interest and identifies the table and column of the replicated specification. The table reports both the published and replicated coefficients, with significance levels indicated by stars at the 1%, 5%, and 10% levels, and compares the number of observations in the published and replicated tests.

Because we replicate a number of papers, we also provide a separate section, "Replication Tables," located after the standard tables, that presents direct comparisons of all explanatory variables between the published and replicated results. Within each table, coefficients are reported in the same order as tabulated in the original publication, with the primary economic variable highlighted in bold. We also include t-statistics, p-values, or standard errors as appropriate, consistent with the original study.

Finally, many of these papers employ fixed-effect indicator variables, which require omitting one unspecified category to estimate a constant. Consequently, published and replicated constants are not directly comparable, and we do not report constants even when they appear in the original tables.

[18] In *foghorn*, the `acquirer` function links source data to deal observations. Thus, we can simply slot in `acquirer Sdc.market_cap` , `acquirer (Crsp.prc * Crsp.shrout)`, and `acquirer (Comp.prccF * Comp.csho)` and the transpiler creates the correct empirical procedure across database sources.

The findings in Table 1 suggest that key results in published papers are generally replicable. In 15 of the 19 papers we have attempted to date, we generate coefficients on key economic variables with comparable economic and statistical magnitudes.[19] These replications suggest that important findings in the literature exist in a common data-processing framework. Thus, the results can be generated simultaneously; researchers do not need to impose paper-specific methodologies and assumptions to reproduce each result.

We do not tabulate papers for which replication was unsuccessful. Failure to replicate does not imply that the original results are not replicable. By applying a declarative approach to economics, we aim to make all assumptions explicit, but the possibility of misinterpreting a paper's methodology remains.

Literature-level replications can significantly enhance transparency. Beyond offering bi-directional transparency for individual papers, generating results within a shared data-processing framework clarifies the empirical practices and assumptions applied across the literature. We now assess the robustness of these results to variations in standard practice through stress testing.

4. Stress Testing the M&A Literature

We now build off of the work of Mitton (2022) and utilize a methodology similar to that suggested by Brodeur, Cook, and Heyes (2020b) and Simonsohn, Simmons, and Nelson (2020) and perform three types of within-literature stress tests on the replicated results presented in the previous section. These tests examine the robustness of each replicated paper's coefficient of interest by varying (1) the dependent variable definition, (2) definitions of control variables common across most of the papers, and (3) the sample selection criteria. We utilize the fact that, consistent with the findings of Mitton (2022), the definitions of the dependent variable, common control variables, and sample selection criteria vary significantly within the M&A literature. We ask whether these choices generate substantial variation in focal statistical significance and whether that variation systematically favors the published specification. In other words, does the significance of the published results tend to be greater than when using plausible alternative definitions.

Our empirical objective is to test the sensitivity of an independently replicated focal result to alternative design choices observed within the same literature. Each stress test begins with the fixed replication baseline established in Section 3 and modifies one dimension of the design while retaining the remaining specification choices. CAR stress tests vary announcement windows, expected-return models, and estimation periods while retaining the baseline control definitions and sample-selection rules. Control-variable stress tests vary the definitions of market-to-book or leverage while retaining the baseline CAR construction and sample-selection rules. Sample-

[19] We continue to work on replication and expand our sample of papers.

selection stress tests vary eligibility rules based on deal form, deal status, and relative acquisition size while retaining the baseline variable definitions. Across these stress tests, we retain the original sample period, estimator, and set of included explanatory variables. Holding selection rules fixed does not necessarily preserve identical observations, because alternative variable constructions can change data availability. Therefore, the stress tests measure the sensitivity of design changes within the original empirical setting, rather than in a new period or an unrestricted search across specifications.

Our approach combines independent replication with systematic specification analysis. Similar to the specification checks of Brodeur, Cook, and Heyes (2020b) and the specification curves of Simonsohn, Simmons, and Nelson (2020), we examine a distribution of results rather than a single preferred specification. Our use of alternative constructions observed within the literature also parallels the analysis of portfolio-construction choices of Hasler (2023) and Hasler (2025) in asset pricing. We apply this logic to independently replicated M&A findings, using a deliberately restricted set of transformations: we vary one design dimension at a time without changing the sample period or searching over the inclusion and exclusion of explanatory variables. These tests complement Muravyev (2026)'s out-of-sample replication by evaluating sensitivity to empirical choices within the original setting. Therefore, they stress test specific specification choices while limiting changes in the economic environment that can accompany out-of-sample validation.

For all stress tests, we study how the significance (t-statistics) of the variables of interest in each paper vary across these stress tests. We define our primary measure, *tstat_diff* for each stress test as the following:

$$tstat_diff_{i,j,k} = \begin{cases} base_tstat_{i,j} - stress_tstat_{i,j,k} & \text{if} \quad base_tstat_i > 0 \\ stress_tstat_{i,j,k} - base_tstat_{i,j} & \text{if} \quad base_tstat_i < 0 \end{cases}$$

where $base_tstat_{i,j}$ is the replicated t-statistic for variable *i* in paper specification *j*, and $stress_tstat_{i,j,k}$ is the t-statistic for variable *i* in paper specification *j* and stress *k*. *tstat_diff* will be positive when the magnitude of the replicated t-statistic is greater than the magnitude of the t-statistic of the same variable in the stress test (i.e. the replicated t-statistic is more significant than the stress test's t-statistic). *tstat_diff* less than zero can be thought of as conservative (the alternative definition in the stress test produces a larger, or more significant t-statistic than the replicated t-statistic of the specification found in the published paper) and *tstat_diff* greater than zero can be thought of as aggressive (the replicated t-statistic of the reported specification in the published paper produces a larger, more significant t-statistic than the stress test).

4.1. Dependent variable stress testing

BEN-DAVID et al. (2026) question the interpretation of acquisition announcement CARs as measures of expected value creation, finding that they largely reflect information about the stand-alone acquirer rather than the transaction. Our analysis addresses a complementary issue: retaining CARs as the outcome measure, we examine how alternative constructions used within the literature affect focal findings and ask whether modest changes in measurement choices can materially alter statistical significance.

To perform stress testing on dependent variable definitions we rely on the chosen sample of papers outlined in the section above and allow the definition and estimation of CARs to vary along three dimensions. First, we vary the size of the CAR announcement window for each test along the following dimensions relative to the announcement day: $[-1, +1]$, $[-2, +2]$, $[-3, +3]$, $[-5, +5]$, $[-1, +5]$, $[-5, +1]$, $[0, +5]$, and $[-5, 0]$. Second, we vary how abnormal returns are estimated: subtracting the expected return from the daily raw return utilizing a single-index model (SIM) with equally weighted returns, an SIM with value weighted returns, a Fama-French 3-factor model (FF3), a Fama-French 4-factor model (FF4), and the equally weighted (OEW) and value weighted (OVW) market return (i.e. assuming the beta of the single-index model is equal to one following Brown and Warner (1980)). Third, we vary the expected return estimation period (for tests estimating the market return using the SIM, FF3, and FF4) along the following daily windows relative to the announcement day: $[-205, -6]$, $[-210, -11]$, $[-245, -45]$, $[-252, -20]$, $[-272, -20]$, $[-300, -91]$, and $[-370, -253]$. These variations generate 238 to 336 stress tests per replicated paper, depending on the specification.

4.1.1. Impact on the significance of variables of interest

We complete the stress tests described in Section 4.1. and compute the *tstat_diff* for each stress test on each paper specification variable of interest. We focus on baseline replicated results from our sample of published papers where the coefficient on the variable of interest was predicted to be significant. From our sample, this yields 15 specifications and 4,543 stress tests for variables of interest. We report the results of these tests in Figure 1 and Table 6.

Figure 1 shows that even slight within-literature variations in how the dependent variable (*CAR*) is defined produce significant dispersion in the t-statistics of variables of interest in the replicated papers. A positive *tstat_diff* indicates the reported/replicated t-statistic is larger in magnitude than the corresponding t-statistic for a given stress test. Table 6, Panel A reports a *tstat_diff* standard deviation of 1.05. Figure 1 also shows that this distribution is significantly skewed in favor of the published specification. As reported in Panel B of Table 6, the reported t-statistic is 0.67 larger than the median stress test t-statistic (i.e. this is consistent with the reported t-statistic being a significant

2.20, while the stress test t-statistic is an insignificant 1.53). This evidence is consistent with the exploratory findings of Menkveld et al. (2024).

As reported in Table 6, Panel C, we find the replicated/reported t-statistic is larger in magnitude than the stress test t-statistic (*tstat_diff* was positive) in 80% of stress tests (3,642 of 4,543). As further reported in the panel, 56% of stress tests result in a drop in significance (i.e. moving from one conventional significance level to a lower one) and 37% of stress tests lose significance (a t-statistic of less than 1.65). Finally, only 3% of stress tests increase in significance.

Our focus is on t-statistics, since research suggests that statistical significance significantly increases the likelihood of publication (Andrews and Kasy 2019). However, we document similar evidence for changes in estimated coefficients. In Figure IA1 and Table IA16 in the Internet Appendix, we find that the coefficient estimates for reported results are significantly larger in magnitude than stressed alternatives. Specifically, Panel A of Figure IA1 displays a histogram of the percentage change in the magnitude of coefficient estimates from stressed alternatives to the reported coefficient. As reported in Table IA16 the average percentage change in the magnitude of the coefficient estimate relative to the stressed alternative for the variable of interest is 42% larger (with a median increase of 15%). We also study the economic significance of these effects by reporting the change in the estimated coefficient multiplied by a within-replicated-sample standard deviation increase (or 1, for indicator variables) in the variable of interest in Figure IA1, Panel B and Table IA16. As seen in the figure and table, the economic change in CARs resulting from a standard deviation increase in the variable of interest is material. Specifically, the average difference in magnitude between the reported result and stress test generates CARs 8 percentage points higher. Taken together, these results demonstrate that reported results tend to be advantaged in terms of statistical and also economic significance.

4.1.2. Impact on the significance of control variables

The higher significance we find in published specifications relative to corresponding stress tests that we document above may be an artifact of our methodology, or simply reflect the randomness induced by non-standard errors (Menkveld et al. 2024). Alternatively, the specific definition of the dependent variable may be driven by the unique underlying economic setting of the paper. In this case, it may not be surprising that the reported results are more significant than stress tests, since the definition itself is more relevant than stressed alternatives.[20] To test these alternative explanations, we calculate *tstat_diff* for control variables rather than variables of interest using the same stress tests. If our methodology is driving the results, we would expect that the signifi-

[20]Varying CAR models and estimation windows may contain implicit assumptions regarding the nature of abnormal returns that may be vary across studies. In the Internet Appendix, we find similar results when we restrict our analysis to only variation of CAR windows in Figure IA2 and Table IA17.

cance of published control variables would behave similar to the variables of interest, relative to their stress test alternatives. Similarly, if the unique underlying economic setting chosen for the published specifications is the most relevant, we would also expect the significance of published control variables to behave like variables of interest. However, if these results are instead driven by publication incentives (arising either from specification search or publication selection), we expect the significance of control variables will not exhibit the same tendency towards using definitions of the dependent variable that yield more significant coefficients on the variable of interest.

The non-focal controls provide an economically meaningful comparison. These variables are included because of their expected relevance to announcement returns, and many are focal explanatory variables in other studies of M&A announcement returns. Therefore, the distinction between focal and control variables reflects their role in a paper's principal hypothesis, rather than a general distinction between economically relevant and irrelevant predictors. Comparing their sensitivity allows us to examine whether the baseline advantage extends broadly across economically relevant coefficients or is stronger for the coefficient central to the paper's contribution. We report these results in Panel B of Figure 1 and Table 6.

Figure 1 and Panels A and B of Table 6 show significant variation in the t-statistics of control variables in stress tests relative to the replicated t-statistics. Table 6, Panel A shows that the standard deviation of *tstat_diff* remains large (1.39) and the skewness (1.35 vs. 1.26) and kurtosis (16.59 vs. 7.80) are even larger. This supports the notion that the choice of dependent variable definition can introduce significant variation in estimated coefficients on all variables. This is consistent with the work of Mitton (2022) and Menkveld et al. (2024). However, we no longer observe the degree of advantage with the significance of control variables that we observed with variables of interest. As reported in Panel B of Table 6, the median difference between a replicated/reported t-statistic of a control variable and the corresponding t-statistic from a stress test is 0.22 (relative to 0.67 for variables of interest). We find that 61% of stress tests of control variables (versus 80% of variables of interest) were lower than the replicated/reported t-statistic. Only 16% of stress tests result in a drop in significance (relative to 56% of stress tests for variables of interest) and 12% of stress tests lose significance (relative to 37% of stress tests for variables of interest). Finally, 12% of stress tests for control variables increase in significance (relative to 3% of stress tests for variables of interest).

A t-test comparing the mean *tstat_diff* between variables of interest (0.775) and control variables (0.336) is highly significant (*t-statistic* of -20.720) and as reported in Table 7, a two-sample Kolmogorov-Smirnov test for the equality of distributions shows that the distribution for *tstat_diff* is significantly smaller for the control variables than for variables of interest. Taken together, these results document evidence of significant sensitivity of variables of interest to within-literature variation in CAR definitions. We document that t-statistics are significantly more skewed towards

significance in the published results relative to corresponding stress tests, where incentives for significant results in publication are the strongest. We do not observe this same behavior (in terms of skewness favoring published results) in the significance of control variable estimates, where these publication incentives are less severe. If instead the results were driven by our methodology comparing the published results with inappropriate alternatives, we would expect similar behavior in non-focal variables. Thus, the evidence is most consistent with these differences arising from publication and journal incentives.

4.1.3. Individual paper and stress test analysis

We have reported stress tests thus far in aggregate across all of the replicated studies. However, aggregate results can conceal heterogeneity across papers. Specifically, our >4500 stress tests are made up of roughly 300 stress test per replicated specification. Therefore, we examine whether the focal-variable baseline advantage is broadly distributed across the replicated specifications. To report this variation, we randomly order the individual paper specifications and report histograms of *tstat_diff* for each specification individually in Panel A of Figure 2. The figure shows that the published specification of most of the replicated papers are advantaged towards higher t-statistics relative to plausibly alternative definitions. However, several are centered around zero and even slightly disadvantaged. Importantly, we observe significant variation in t-statistics for all specifications. Thus, while the selection of dependent variables tends to favor the published version of specifications more in some papers than others, all exhibit significant variation and suggest that our evidence is not being driven by one outlying paper-specification. Panel B of Figure 2 similarly displays the distribution of *tstat_diff* for control variables for each of the individual papers. Similar to in aggregate, we observe significant differences between the distributions of *tstat_diff* for variables of interest and control variables at the individual paper level. In addition, Table 8 reports the mean *tstat_diff* for each paper's variable of interest, control variables, their differences, as well as the significance (p-value) of the differences. As can be seen in the table, the mean *tstat_diff* of focal variables are significantly larger than the mean *tstat_diff* of control variables in 11 of the 15 specifications. Interestingly, 4 of the papers' focal variables have lower average *tstat_diff* relative to control variables, with two of these significant so. Taken together, while we observe heterogeneity among papers in our sample, on average published specifications are significantly more advantaged towards significance relative to control variables, with the large majority of individual papers demonstrating similar characteristics and significance.

Next, we study whether the skewness or dispersion we observe in the reported significance of variables of interest is determined by specific stress tests. To do so we employ a specification curve methodology similar to Simonsohn, Simmons, and Nelson (2020) in Figure 3. Specifically,

we report the *tstat_diff* of stress tests between the 1st and 99th percentile. Additionally, we also include vertically aligned horizontal rows below the figure designating whether a given empirical choice was included in a given stress test. As can be seen in the figure, we observe that no single specification type tends to explain large (or small) *tstat_diff*. In other words, it does not appear that papers avoid a specific definition literature-wide. Rather, we observe consistent results across all specification choices.

Consistent with these results, we do not find that our baseline results are driven by a specific stress test or specific type of stress test in untabulated robustness tests. We find consistent evidence when we drop any specific stress test (i.e. no stress tests have a $[-5, 0]$ announcement window) or stress test type (i.e. do not allow stress test CAR windows to vary but allow the estimation period or method to vary).

4.1.4. Do higher hurdle rates help?

A potential path to prevent specification search is adjusting hurdle rates. Harvey, Sancetta, and Zhao (2026) motivate a minimum t-statistic hurdle of approximately three for new factor discoveries to avoid spurious results due to publication bias. Although their threshold is derived for factor research, it provides a useful reference point for examining our results. Further, Ringgenberg and Samadi (2026) estimate minimum thresholds of 3.0 to 3.6 to reduce the false discovery rate in frequently reused natural experiments in corporate finance. Both of these hurdle rate recommendations roughly correspond to cutoffs following Andrews and Kasy (2019) implying that the likelihood of publication is 100× higher for a when a result is significant at 95%, relative to results <95% significant. Two natural questions arise from these recommendations. First, do we observe increased publication advantage for results with replicated t-statistics < 3? Second, do higher hurdle rates prevent specification search?

To examine the first question, we study whether specification sensitivity occurs among findings with weaker baseline statistical evidence. Therefore, in Figure 4, we restrict our analysis to replicated t-statistics that are less than 3 (where a reduction of 1.05 would reduce the significance of an estimate from 1% significance to 10% significance at best). We find similar evidence of dispersion and loss of significance in stress tests where the baseline reported t-statistic is less than 3 for variables of interest. Table 9, Panels A and B, reports the summary statistics of *tstat_diff* in this subsample, with a median of 0.56, a mean of 0.60, and standard deviation of 0.76. Given the proximity to cutoffs for significance we see an even greater proportion of stress tests with lower significance in Panel C. While 78% of stress tests are smaller in magnitude than the replicated/ reported t-statistic, 66% of stress tests dropped significance, 46% lost significance, and only 4% increased in significance. Thus, it appears that specifications under the proposed t-statistic hurdle

of 3 significantly benefit from the reported result relative to stressed alternatives. Further, it does not appear that our findings are only found among highly significant results.

We examine the second question in Table 10. Simply, we assess how frequently observed gaps in t-statistics between baseline–alternative definitions are large enough to bridge the distance between conventional significance cutoffs and higher proposed hurdles. This measures the scope for threshold crossings under an illustrative benchmark. Simply, we assume the distribution of *tstat_diff* is the variation available to authors that would allow a move up in the level of significance. For a higher hurdle (h) and starting cutoff (c), we report the share of observed gaps satisfying *tstat_diff* $>$ h $-$ c. For example, a positive *tstat_diff* would allow a result just below the 99% to become significant at the 99% level at standard significance thresholds. Similarly, a *tstat_diff* $>$ 0.62 would allow a result just below the 95% significance level (1.96) to become 99% significant at standard thresholds. In Panel A of Table 10, we document the proportion of tests that would allow these movements in significance levels using standard thresholds. As can be seen in the Panel A, 80% of stress tests would allow a result just below 99% significance (significant at the 95% level) to increase to 99% percent significance, while 52% and 40% of stress tests would allow a similar move from just below 95% significance (90% significant) and just below 90% significance (insignificant) to 99% significance, respectively.

In Panel B, we next report the impact of increasing the t-statistic threshold to 3.02, which is motivated by Andrews and Kasy (2019) where we assume a selection process leads to an estimated result with a 95% significance level being 100$\times$ more likely to be published relative to results less than 95% significant (t-statistic = 3.02). This increase significantly reduces the proportion of stress tests that would allow a move up in significance. Specifically, 60% of stress tests would allow moving from just below 99% significant (95% significant) at standard thresholds ($<$ 2.58) to a t-statistic above 3.02. Similarly, 35% of stress tests and 21% of stress tests would allow a move from 90% significance ($<$1.96) and insignificance ($<$1.65) at standard thresholds, respectively. We observe an even greater reduction when using the 99% significance threshold (t-statistic = 3.48) implied by the same assumptions motivated by Andrews and Kasy (2019) above, with 41%, 16%, and 11% of stress tests allowing increases from 95% significance, 90% significance, and insignificant at standard thresholds, respectively.

These results demonstrate that higher significance thresholds can materially reduce the proportion of stress tests that would allow an otherwise insignificant result at standard thresholds to be significant at higher thresholds. However, the ability to specification search remains. Specifically, a material number of variations still remain that would allow otherwise insignificant results to meet higher hurdles. For example, 11% the 300 choices available per specification in the CAR stress tests imply 33 choices that would allow an otherwise insignificant results at standard levels (t-

statistic < 1.65) to increase in significance to a t-statistic > 3.48. Thus, while higher hurdles help, especially in the case of randomness, it does not fully constrain specification search. We now turn to studying whether and how the definitions of independent control variables materially impact the significance of variables of interest.

4.2. Independent (control) variable stress testing

To perform stress testing on independent variable definitions we rely on the chosen sample of papers outlined in the section above and allow the definition and estimation of two controls common to the majority of the published studies in our sample: (1) market-to-book (Tobin's Q) and (2) leverage. As identified by Mitton (2022), both of these measures vary significantly across the general corporate finance literature. Similar to our analysis in Section 4.1., we vary the definition of these control variables only within the more narrow M&A literature, since these represent defensible alternative definitions of these variables. We then study the impact that varying only these definitions has on the significance of the variables of interest. We utilize the nine alternative definitions of market-to-book provided, along with references, in Table 3. These definitions are used within the literature as controls from both acquirers and targets. We also utilize the nine definitions of leverage provided, along with references to the appropriate literature, in Table 4.[21] We allow these variables to vary for those specifications in our sample in which they appear. These variations generate 80 to 96 stress tests per specification, dependent on the specification. Unlike varying the dependent variable, changing a control variable definition does not mechanically alter the left-hand side of the regression and can only affect the coefficient on the variable of interest through correlation.

4.2.1. Independent (control) variable stress test results

From our sample, 12 of the 15 specifications contain either market-to-book or leverage, yielding 1,136 stress tests for variables of interest. As we are not stress testing the inclusion/exclusion of variables, if a given control variable does not appear in a specification, we do not vary the definition of that variable for stress tests of the specification. We report the results of these tests in Figure 5 and Table 11. Figure 5 shows that, while dispersion in *tstat_diff* persists, there is much larger clustering around zero. Table 11, Panels A and B confirm this. The mean, median, and standard deviation of *tstat_diff* is smaller than what we observed in the dependent variable stress tests, at 0.13, 0.03, and 0.53, respectively. However, we see even higher skewness and kurtosis, 2.34 and 11.69, respectively.

[21]We conservatively rely on the variation of definitions of these variables that occurs only within the replicated M&A literature. However, Mitton (2022) identifies 25 unique definitions of Tobin's Q and 96 unique definitions of leverage across all areas of corporate finance within his sample.

Importantly, we continue to see a slight skewness favoring the published specifications in Table 11, Panel C. Specifically, 63% of stress tests were lower than the reported/replicated t-statistic. 20% of stress tests resulted in a drop in significance, 11% resulted in a loss in significance, and only 2% resulted in an increase in significance. These results are roughly consistent when we restrict the sample to those reported/replicated results with t-statistics less than 3, as reported in Panel B of Figure 5 and Table 11. In these cases we see roughly similar summary statistics, and slightly higher proportion of stress tests leading to a drop in significance (25%) and a loss in significance (14%).

To fully ascertain whether these results may be an artifact of our methodology, we again compare the results to the variation in t-statistics for control variables. Importantly, since we are varying control market-to-book and leverage, we exclude these variables from this sample when they are varied (to avoid creating a direct effect on the t-statistics). We present these results in Figure 6 and Table 12.

The figure shows even higher clustering of stress tests near a *tstat_diff* of zero. Table 12, Panel A shows that the average *tstat_diff* for control variables (excluding market-to-book or leverage when these are allowed to vary in a given stress test) is 0.04 (compared with 0.13). This difference is similarly significant with a t-statistic of -3.45 and a p-value of 0.0006. Further, as reported in Table 13, a two-sample Kolmogorov-Smirnov test for the equality of distributions shows that the distribution for *tstat_diff* is significantly smaller for the control variables than for variables of interest when we vary market-to-book and/or leverage definitions. Similar to the CAR analysis, we also report specification curve analysis on *tstat_diff* for these control variables in Figure 7. Similar to our CAR results, we do not find evidence that the skewness (or dispersion) of reported t-statistics (relative to their within-literature alternatives) are driven by particular stress test definitions.

Taken together, these findings suggest that, while variation in control variables does not have the same size of impact on the significance of variables of interest, it still generates a meaningful effect. This is unsurprising given its indirect influence on the estimation of the coefficients of interest. However, again we find evidence that the definitions chosen for control variables tend to favor the significance of the variable of interest in published results.

4.3. Sample selection stress testing

We now perform stress testing on the sample selection categories common in the M&A literature that vary significantly within this literature. To do so, we rely on the chosen sample of papers outlined in the section above and vary three sample selection criteria: (1) the inclusion into the sample based on the deal form reported in SDC Platinum, (2) the inclusion into the sample based on the deal status reported in SDC Platinum, and (3) the inclusion into the sample based on the

relative size (the size of the acquisition relative to the size of the acquirer). While these specific sample selection details are relevant to the M&A settings, similar sample selection variation exists across the literature (Di Tillio, Alfredo, Ottaviani, and Sørensen 2021). Similar to our analysis above, we vary these sample selection criteria only within the more narrow M&A literature, since these represent defensible alternative definitions of these variables. We then study the impact of that varying these decisions has on the significance of variables of interest, and compare those to the variation in significance of non-focal controls.

We take advantage of the variation of sample selection criteria in this literature as highlighted in Table 5 to perform stress tests on these decisions. Panel A highlights the variation of cutoffs based on the relative size of the deal within the literature. Using these as a baseline, we disaggregate and vary these decisions in 13 ways that highlighted in subsequent figures. SDC Platinum's deal form classifies the structural form of transactions, or what ownership or assets are actually being acquired. We take advantage of and disaggregate the variation within this literature in the inclusion into the sample by deal form reported in Table 5, Panel B in seven ways. Briefly, these deal forms include variations of whole mergers and acquisitions, majority-interest acquisitions, and acquisitions of assets (without acquiring the ownership of the firm). We also include stress tests without restrictions as well as those disaggregating some of these above. Finally, we also vary the sample selection criteria in stress tests by deal status in the three ways highlighted in Table 5, Panel C. These variations generate 292-293 stress tests per specification, dependent on the specification.

4.3.1. Sample selection stress test results

Our sample of the 15 specifications yields 4,386 stress tests for the sample selection criteria. We report the results of these tests in Figure 8 and Table 14 In both Figure 8 and Table 14, we continue to find similar dispersion and skewness favoring published specifications as we observed in our analysis of dependent variables in Section 4.1., and greater than was found for control variables in Section 4.2.. The mean, median, and standard deviation of *tstat_diff* for the sample selection stress tests are 0.82, 0.60, and 1.39, respectively. In Table 14, Panel C, we find that the t-statistics of our variables of interest were lower in our 75% of stress tests relative to the reported/replicated t-statistic. Further, and similar to what we observed for the dependent variable stress tests, 53% of stress tests resulted in drop in significance, 39% resulted in a loss of significance, and only 7% resulted in an increase in significance. We also find consistent evidence when we restrict the sample to those reported/replicated results with t-statistics less than 3, as reported in Panel B of Figure 8 and Table 14.

Similar to our discussion in Section 4.1., these results could be an artifact of our methodology. For example, it is likely that certain sample selection decisions, such as including deals that involve

the acquisition of assets, but not of the ownership of the firm, are only appropriate for certain economic questions. Therefore, including the acquisition of assets in the sample of a stress test where it is not economically important may inappropriately bias the significance of the variable of interest downward. As we argue above, if the sample selection criteria were based on the economic question unique to the particular setting of one paper, but not of the literature as a whole, we would expect a discussion of this criteria. Unfortunately, a limited minority of papers describes their sample selection criteria justification. Further, if this is the source of the positive bias in published t-statistics and not publication incentives, we would expect similar behavior in our non-focal control variables. Therefore, we compare the results to the variation in t-statistics for control variables in Figure 9 and Table 15.

The differences in the distributions of t-statistics for variables of interest in Figure 8 and non-focal control variables in Figure 4 in our stress tests are clear. We observe significantly more even dispersion, as well as bunching near zero in the *tstat_diff* for control variables. Table 15, Panel A shows similar evidence. The average *tstat_diff* for non-focal control variables is 0.43 (compared with 0.82). This difference is highly significant with a t-statistic of -15.27 and a p-value of <0.001. In addition, we again compare these distributions formally in Table 16. In the table, we using a two-sample Kolmogorov-Smirnov test for the equality of distributions, we find that the distribution for *tstat_diff* is significantly smaller, with a p-value of <0.001 for the control variables than for variables of interest when we vary sample selection decisions.

Finally, we report specification curve analysis on *tstat_diff* for the sample selection decisions in Figure 10. Similar to the dependent variable and control variable definition stress tests above, we do not find evidence that the sample selection stress test results are driven by a particular stress test. However, we do observe some interesting findings. Specifically, it appears that the *tstat_diff* for stress tests including the acquisition of assets tend to lower, while those excluding acquisition of assets tend to be higher. We observe a similar, albeit more limited, relationship with deals including acquisitions of majority interest. As discussed above, deals classified as an acquisition of assets by SDC Platinum involve an acquirer purchasing the assets or business operations of a target, rather than ownership in the target. Therefore, these deals differ from others because they do not seek majority ownership in the target firm. Further, deals classified as acquisitions are for ownership of the target, but SDC Platinum does not have additional information as to whether it is for majority ownership, for the whole firm, or some other outcome.

To further study these issues, Figure 11 presents an additional specification curve that classifies subsamples of stress tests by whether the baseline published test or the stress test includes: acquisitions (A), acquisitions of majority interest (AM), and acquisitions of assets (AA). As can be seen in the figure, we do not observe evidence that decision to include or exclude acquisitions materially

impacts the *tstat_diff*. However, we observe important difference when considering the inclusion/ exclusion decision of deals classified as AA (and to a lesser extent those classified as AM). As seen in the figure, we find that when a baseline published test excludes deals classified as AA, adding these deals in stress tests does not tend to be much less significant than the baseline results. However, when these deals are included, and a stress test excludes them, stress test t-statistics tend to be much lower (and *tstat_diff* tends to be much higher) than the published results. These results are consistent with AA deals being included when it increases the reported t-statistics of variables of interest, as excluding them significantly reduces significance. However, including these deals when they are excluded does not materially impact significance.

These findings suggest that variation in sample selection requirements generates significant positive skew in reported/replicated t-statistics. In addition, certain deal forms, such as acquisition of assets, tend to be included when it increases the significance of reported results. Thus, the chosen sample selection requirements tend to favor the significance of the variable of interest in published results.

5. Conclusion

This paper utilizes a newly-developed novel declarative econometric technique from Tumarkin (2026). This scalable technique allows us to build on the existing literature by combining the replication and meta-study approaches to stress test an economic literature. Specifically, applying this framework to the mergers and acquisitions literature, we show the *actual*, rather than the *potential*, impacts that seemingly innocuous empirical decisions, or "researcher degrees of freedom", have on the significance (and publishability) of coefficient estimates. We do this by replicating the results of existing studies and then stress testing these replicated results.

We show that most published results in this literature are replicable once implicit assumptions are made explicit. However, we find that within-literature stress testing reveals significant sensitivity to specification choices. By utilizing the same dataset used to replicate the original results, we vary the definitions of dependent variables (cumulative abnormal returns), common independent control variables (Tobin's Q and leverage), and sample selection criteria (deal form, deal status, and relative size of deal cutoffs) only within the replicated M&A literature. This creates a series of conservative stress tests using plausible and defensible alternative definitions that could have been used during when designing the empirical specifications of the paper. We find that even among these conservative stress tests, reported results are significantly more likely to be above standard thresholds for statistical significance (e.g., 10%, 5%, and 1%) than stressed alternatives. Specifically, reported results tend to have significantly higher t-statistics, be much more likely to be significant at at least the 10% level, and be more likely to be significant at higher thresholds than plausible alterative specifications from the literature. This suggests that researcher discretion,

as well as the incentives of the publication process, plays a meaningful role in shaping published findings. At the same time, the framework highlights that many conclusions remain robust under alternative specifications, reinforcing the value of accumulated knowledge in this field.

The additional analyses clarify the breadth and implications of this sensitivity. The focal-variable baseline advantage occurs across most individual papers, although some exhibit little advantage or the opposite pattern. Specification-curve analyses show that the aggregate patterns are broadly distributed across alternative empirical choices rather than attributable to a single construction. Finally, higher significance hurdles reduce the share of observed differences between baseline and alternative t-statistic that are large enough to bridge conventional and higher thresholds. However, a material share remains, indicating that stronger statistical hurdles retain value while leaving a role for systematic evaluation of specification dependence.

More broadly, our application demonstrates how declarative econometrics can make empirical choices easier to communicate and their consequences easier to evaluate (Tumarkin 2026). Extending this approach to other literatures requires suitable data and methodology modules, whose development costs can be shared across replication, stress testing, and new research. As AI reduces the cost of specification search for significant results, an inexpensive scrutiny of the dependence of a significant result on a given specification becomes increasingly valuable. By allowing authors to disclose their designs succinctly and others to evaluate defensible alternatives systematically, this framework provides a practical mechanism for strengthening research evaluation and potentially reducing the returns to searching for and reporting specifications based on statistical significance. In doing so, declarative econometrics provides a path toward more resilient empirical literatures.

References

Aczel, Balazs, Barnabas Szaszi, Harry T Clelland, Marton Kovacs, Felix Holzmeister, Don Van Ravenzwaaij, Hannah Schulz-Kümpel, et al. 2026. "Investigating the Analytical Robustness of the Social and Behavioural Sciences". *Nature* 652 (8108): 135–42.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias". *American Economic Review* 109 (8): 2766–94.

Asher, Samuel GZ, Janet Malzahn, Jessica M Persano, Elliot J Paschal, Andrew CW Myers, and Andrew B Hall. 2026. "Do Claude Code and Codex P-Hack? Sycophancy and Statistical Analysis in Large Language Models". *Sycophancy and Statistical Analysis in Large Language Models (February 19, 2026)*.

Bai, Longju, Zhemin Huang, Xingyao Wang, Jiao Sun, Rada Mihalcea, Erik Brynjolfsson, Alex Pentland, and Jiaxin Pei. 2026. "How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks". *Arxiv Preprint Arxiv:2604.22750*.

Bates, Thomas W., and Michael L. Lemmon. 2003. "Breaking up is hard to do? An analysis of termination fee provisions and merger outcomes". *Journal of Financial Economics* 69 (3): 469–504.

Bebchuk, Lucian, Alma Cohen, and Allen Ferrell. 2008. "What Matters in Corporate Governance?". *Review of Financial Studies* 22 (2): 783–827.

Becher, David A, Thomas P Griffin, and Greg Nini. 2021. "Creditor Control of Corporate Acquisitions". *The Review of Financial Studies* 35 (4): 1897–1932.

BEN-DAVID, ITZHAK, UTPAL BHATTACHARYA, RUIDI HUANG, and STACEY JACOBSEN. 2026. "The (Missing) Relation between Acquisition Announcement Returns and Value Creation". *The Journal of Finance* 81 (3): 1265–1320.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the analysis of a single neuroimaging dataset by many teams". *Nature* 582 (7810): 84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty". *Proceedings of the National Academy of Sciences* 119 (44): e2203150119.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020b. "A Proposed Specification Check for P-Hacking". In *AEA Papers and Proceedings*, 110:66–69.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020a. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics". *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Lamis Kattan, and Marco Musumeci. 2026. "Job Market Stars: Statistical Significance and Academic Hiring in Economics". *European Economic Review*, 105353.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back". *American Economic Journal: Applied Economics* 8 (1): 1–32.

Brodeur, Abel, Derek Mikola, Nikolai Cook, Lenka Fiala, Thomas Brailey, Ryan Briggs, Alexandra De Gendre, et al. 2026. "Reproducibility and Robustness of Economics and Political Science Research". *Nature* 652 (8108): 151–56.

Brown, Stephen J, and Jerold B Warner. 1980. “Measuring security price performance”. *Journal of Financial Economics* 8 (3): 205–58.

Bryzgalova, Svetlana, Sven Lerner, Martin Lettau, and Markus Pelger. 2024. “Missing Financial Data”. *The Review of Financial Studies* 38 (3): 803–82.

Burch, Timothy R. 2001. “Locking out rival bidders: The use of lockup options in corporate mergers”. *Journal of Financial Economics* 60 (1): 103–41.

Chang, Andrew C., and Phillip Li. 2022. “Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say “Often Not”. *Critical Finance Review* 11 (1): 185–206.

Chang, Xin, Huasheng Gao, and Wei Li. 2025. “Discontinuous Distribution of Test Statistics Around Significance Thresholds in Empirical Accounting Studies”. *Journal of Accounting Research* 63 (1): 165–206.

Chen, Andrew Y, and Tom Zimmermann. 2020. “Publication Bias and the Cross-Section of Stock Returns”. *The Review of Asset Pricing Studies* 10 (2): 249–89.

Chen, Guiming Hardy, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. “Humans or Llms as the Judge? A Study on Judgement Bias”. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–27.

Chen, Hongyu, and Seraphina Goldfarb-Tarrant. 2025. “Safer or Luckier? Llms as Safety Evaluators Are Not Robust to Artifacts”. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 19750–66.

Di Tillio, Alfredo, Marco Ottaviani, and Peter Norman Sørensen. 2021. “Strategic Sample Selection”. *Econometrica* 89 (2): 911–53.

Fama, Eugene F, and Kenneth R French. 1992. “The Cross-Section of Expected Stock Returns”. *The Journal of Finance* 47 (2): 427.

Fama, Eugene F, and Kenneth R French. 1993. “Common Risk Factors in the Returns on Stocks and Bonds”. *Journal of Financial Economics* 33 (1): 3–56.

Fišar, Miloš, Ben Greiner, Christoph Huber, Elena Katok, and Ali I Ozkes. 2024. “Reproducibility in Management Science”. *Management Science* 70 (3): 1343–56.

Fuller, Kathleen, Jeffry Netter, and Mike Stegemoller. 2002. “What Do Returns to Acquiring Firms Tell Us? Evidence from Firms That Make Many Acquisitions”. *The Journal of Finance* 57 (4): 1763–93.

Golubov, Andrey, Alfred Yawson, and Huizhong Zhang. 2015. “Extraordinary acquirers”. *Journal of Financial Economics* 116 (2): 314–30.

Gorton, Gary, Matthias Kahl, and Richard J. Rosen. 2009. “Eat or Be Eaten: A Theory of Mergers and Firm Size”. *The Journal of Finance* 64 (3): 1291–1344.

Harford, Jarrad, Mark Humphery-Jenner, and Ronan Powell. 2012. “The sources of value destruction in acquisitions by entrenched managers”. *Journal of Financial Economics* 106 (3): 247–61.

Harvey, Campbell R, Alessio Sancetta, and Yuqian Zhao. 2026. “What Threshold Should Be Applied to Tests of Factor Models?”. *National Bureau of Economic Research Working Paper*.

Harvey, Campbell R. 2017. “Presidential Address: The Scientific Outlook in Financial Economics”. *The Journal of Finance* 72 (4): 1399–1440.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "… and the Cross-Section of Expected Returns". *The Review of Financial Studies* 29 (1): 5–68.

Hasler, Mathias. 2023. "Looking under the Hood of Data-Mining". *Available at SSRN 4279944*.

Hasler, Mathias. 2025. "Is the Value Premium Smaller Than We Thought?". *Critical Finance Review* 14 (3): 357–86.

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

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The influence of hidden researcher decisions in applied microeconomics". *Economic Inquiry* 59 (3): 944–60.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False". *Plos Medicine* 2 (8): e124.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen. 2023. "Is There a Replication Crisis in Finance?". *The Journal of Finance* 78 (5): 2465–2518.

Jiang, Hongchao, Yiming Chen, Yushi Cao, Hung-yi Lee, and Robby T Tan. 2026. "Codejudgebench: Benchmarking Llm-as-a-Judge for Coding Tasks". In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 19416–48.

Jin, Haolin, and Huaming Chen. 2026. "Are Llms Reliable Code Reviewers? Systematic Overcorrection in Requirement Conformance Judgement". *Automated Software Engineering* 33 (3): 90.

John, Kose, Anzhela Knyazeva, and Diana Knyazeva. 2015. "Employee rights and acquisitions". *Journal of Financial Economics* 118 (1): 49–69.

Koh, Ping-Sheng, and David M. Reeb. 2015. "Missing R&D". *Journal of Accounting and Economics* 60 (1): 73–94.

Kohli, Guneet. 2026. "Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels". *Arxiv Preprint Arxiv:2605.29800*.

Kolari, James W., Seppo Pynnonen, and Ahmet M. Tuncez. 2021. "Further evidence on long-run abnormal returns after corporate events". *The Quarterly Review of Economics and Finance* 81:421–39.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Vol. 53. Wiley, New York.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics". *The American Economic Review* 73 (1): 31–43.

Li, Kai, Buhui Qiu, and Rui Shen. 2018. "Organization Capital and Mergers and Acquisitions". *Journal of Financial and Quantitative Analysis* 53 (4): 1871–1909.

Ljungqvist, Alexander, Christopher Malloy, and Felicia Marston. 2009. "Rewriting History". *The Journal of Finance* 64 (4): 1935–60.

Lyle, Matthew R., Federico Siano, and Teri Lombardi Yohn. 2025. "Re-Adjusted Financial Statement Data: Challenges in Replicating Research". *SSRN Electronic Journal*.

Ma, Qingzhong, David A. Whidbee, and Wei Zhang. 2019. "Acquirer reference prices and acquisition performance". *Journal of Financial Economics* 132 (1): 175–99.

Masulis, Ronald W., Cong Wang, and Fei Xie. 2007. "Corporate Governance and Acquirer Returns". *The Journal of Finance* 62 (4): 1851–89.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Michael Kirchler, Sebastian NeusüSS, Michael Razen, and Utz Weitzel. 2024. "Nonstandard Errors". *The Journal of Finance* 79 (3): 2339–90.

Miske, Olivia, Anna Lou Abatayo, Mason Daley, Mirka Dirzo, Nicholas Fox, Noah Haber, Krystal M Hahn, et al. 2026. "Investigating the Reproducibility of the Social and Behavioural Sciences". *Nature* 652 (8108): 126–34.

Mitton, Todd. 2022. "Methodological Variation in Empirical Corporate Finance". *The Review of Financial Studies* 35 (2): 527–75.

Moeller, Sara B, Frederik P Schlingemann, and René M Stulz. 2004. "Firm size and the gains from acquisitions". *Journal of Financial Economics* 73 (2): 201–28.

Mulherin, J. Harold, Jeffry M. Netter, and Annette B. Poulsen. 2017. "The Evidence on Mergers and Acquisitions: A Historical and Modern Report". In *The Handbook of the Economics of Corporate Governance*, edited by Benjamin E.Hermalin and Michael S. Weisbach, 1:230–90.

Muravyev, Dmitriy. 2026. "Does Empirical Finance Replicate?". *Available at SSRN 7439099*.

Nguyen, Nam H., and Hieu V. Phan. 2017. "Policy Uncertainty and Mergers and Acquisitions". *Journal of Financial and Quantitative Analysis* 52 (2): 613–44.

Novy-Marx, Robert, and Mihail Velikov. 2024. "Assaying Anomalies". *Available at SSRN 4338007*.

Novy-Marx, Robert, and Mihail Velikov. 2026. "Artificial Intelligence–Powered (Finance) Scholarship". *Journal of Economic Literature* 64 (1): 5–37.

Oviedo, Felipe, Fiodar Kazhamiaka, Esha Choukse, Allen Kim, Amy Luers, Melanie Nakagawa, Ricardo Bianchini, and Juan M Lavista Ferres. 2026. "Energy Use of AI Inference, Efficiency Pathways, And Test-Time Scaling". *Joule*.

Ringgenberg, Matthew C, and Mehrdad Samadi. 2026. "Publication Bias in Reused Natural Experiments". *Available at SSRN 6919600*.

Roosenboom, Peter, Frederik P Schlingemann, and Manuel Vasconcelos. 2013. "Does Stock Liquidity Affect Incentives to Monitor? Evidence from Corporate Takeovers". *Review of Financial Studies* 27 (8): 2392–2433.

Shi, Lin, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. "Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge". In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 292–314.

Silberzahn, R., E. L. Uhlmann, and D. P. Martin. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results". *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56.

Silberzahn, Raphael, and Eric L. Uhlmann. 2015. "Crowdsourced Research: Many Hands Make Tight Work". *Nature* 526 (7572): 189–91.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology”. *Psychological Science* 22 (11): 1359–66.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. “Specification curve analysis”. *Nature Human Behaviour* 4 (11): 1208–14.

Sterling, Theodore D. 1959. “Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa”. *Journal of the American Statistical Association* 54 (285): 30–34.

Sun, Yan. 2026. “The Garden, Measured: Loading, Exploitation, And the False-Positive Rate of LLM-Assisted Specification Choice”. *Measured: Loading, Exploitation, And the False-Positive Rate of LLM-Assisted Specification Choice (July 11, 2026)*.

Thomas, Maria, Kristina Gligoric, and Nihar B Shah. 2026. “Mitigating LLM-Based P-Hacking by Preregistering for the Next Llm”. *Arxiv Preprint Arxiv:2606.27687*.

Tumarkin, Robert. 2026. “Declarative Econometrics: From Data to Beta”. *University of New South Wales Working Paper (Available on Request)*.

Tyner, Andrew H, Anna Lou Abatayo, Mason Daley, Samuel Field, Nicholas Fox, Noah A Haber, Krystal M Hahn, et al. 2026. “Investigating the Replicability of the Social and Behavioural Sciences”. *Nature* 652 (8108): 143–50.

Wang, Cong, and Fei Xie. 2009. “Corporate Governance Transfer and Synergistic Gains from Mergers and Acquisitions”. *The Review of Financial Studies* 22 (2): 829–58.

Wang, Ruiqi, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025. “Can Llms Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering”. *Proceedings of the ACM on Software Engineering* 2 (ISSTA): 1955–77.

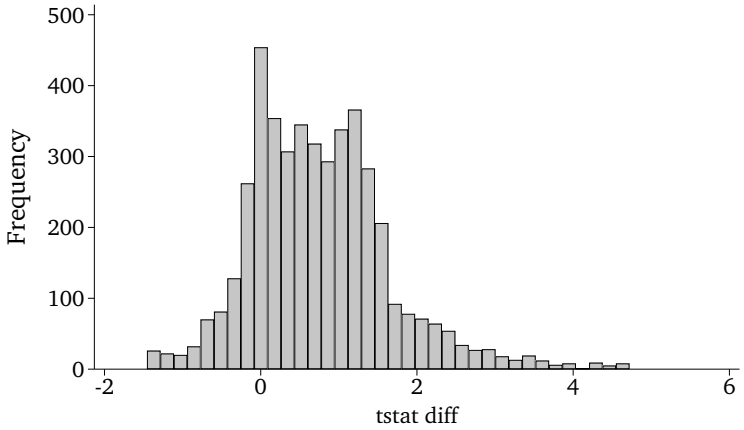
Wataoka, Koki, Tsubasa Takahashi, and Ryokan Ri. 2024. “Self-Preference Bias in LLM-as-a-Judge”. *Arxiv Preprint Arxiv:2410.21819*.

Whited, Toni. 2023. “Costs and Benefits of Reproducibility in Finance and Economics”. *Harvard Data Science Review* 5 (3).

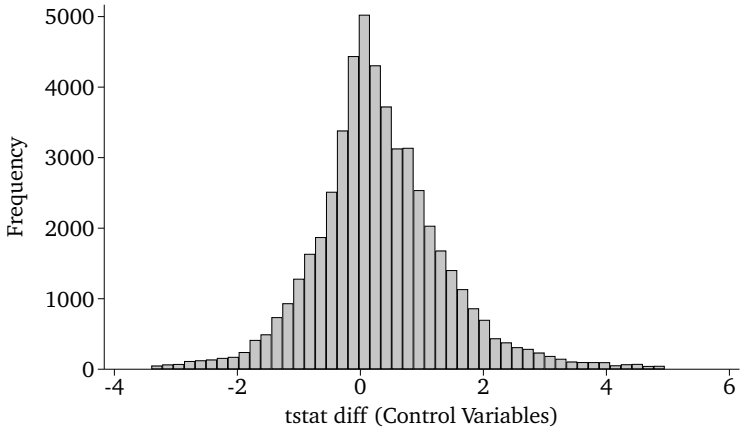
Figure 1

CAR Stress Tests Histograms: Variables of Interest and Control Variables

This histogram plots *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest and control variables by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Stress tests allow for within-literature variation on (1) the CAR announcement window ($[-1, +1]$, $[-2, +2]$, $[-3, +3]$, $[-5, +5]$, $[-1, +5]$, $[-5, +1]$, $[0, +5]$, and $[-5, 0]$), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ($[-205, -6]$, $[-210, -11]$, $[-245, -45]$, $[-252, -20]$, $[-272, -20]$, $[-300, -91]$, and $[-370, -253]$). *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel A plots the 1st through 99th percentiles of all stress tests for variables of interest resulting in 4,543 stress tests. Panel B plots the 1st through 99th percentiles of all stress tests for control variables resulting in 52,234 stress tests (one *tstat_diff* for each control variable-specification-stress test).



Panel A: Variables of interest



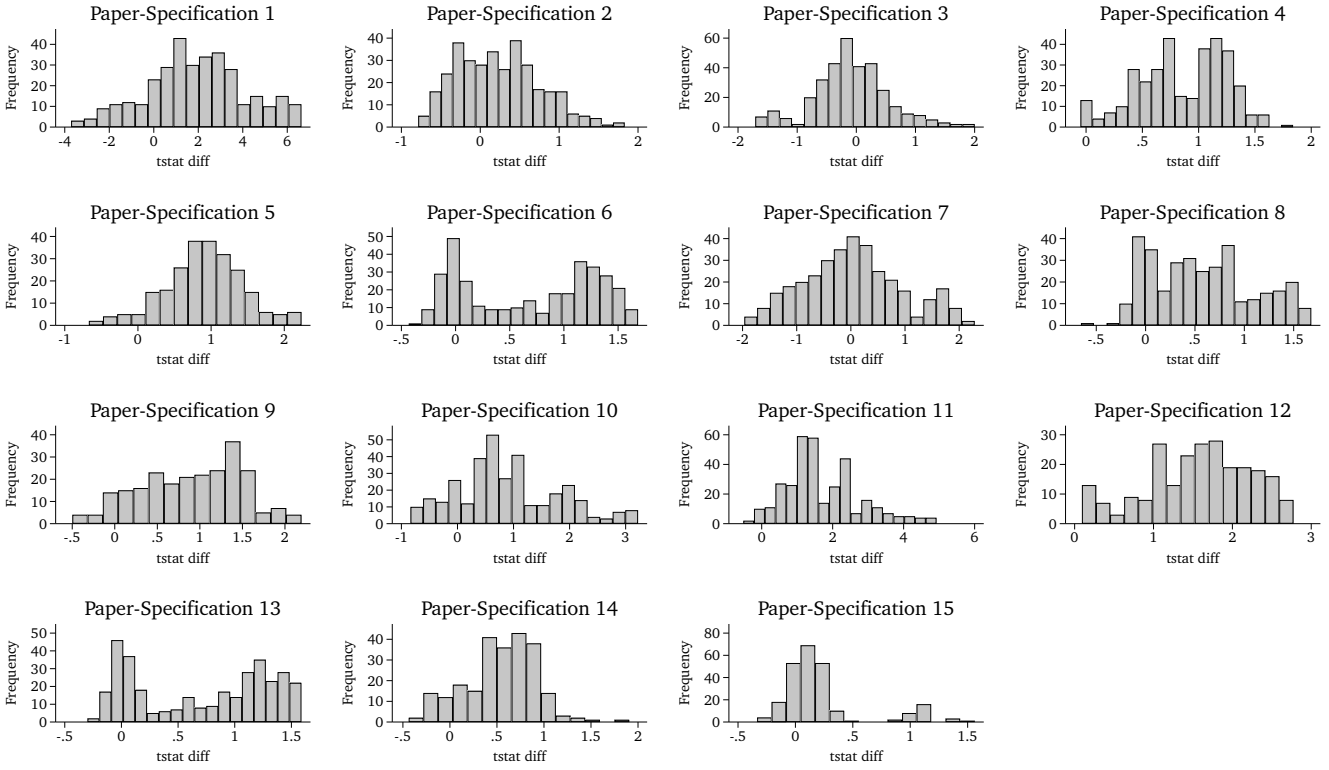
Panel B: Control variables

,

Figure 2

CAR Stress Tests Histograms: Individual Paper Variables of Interest and Control Variables

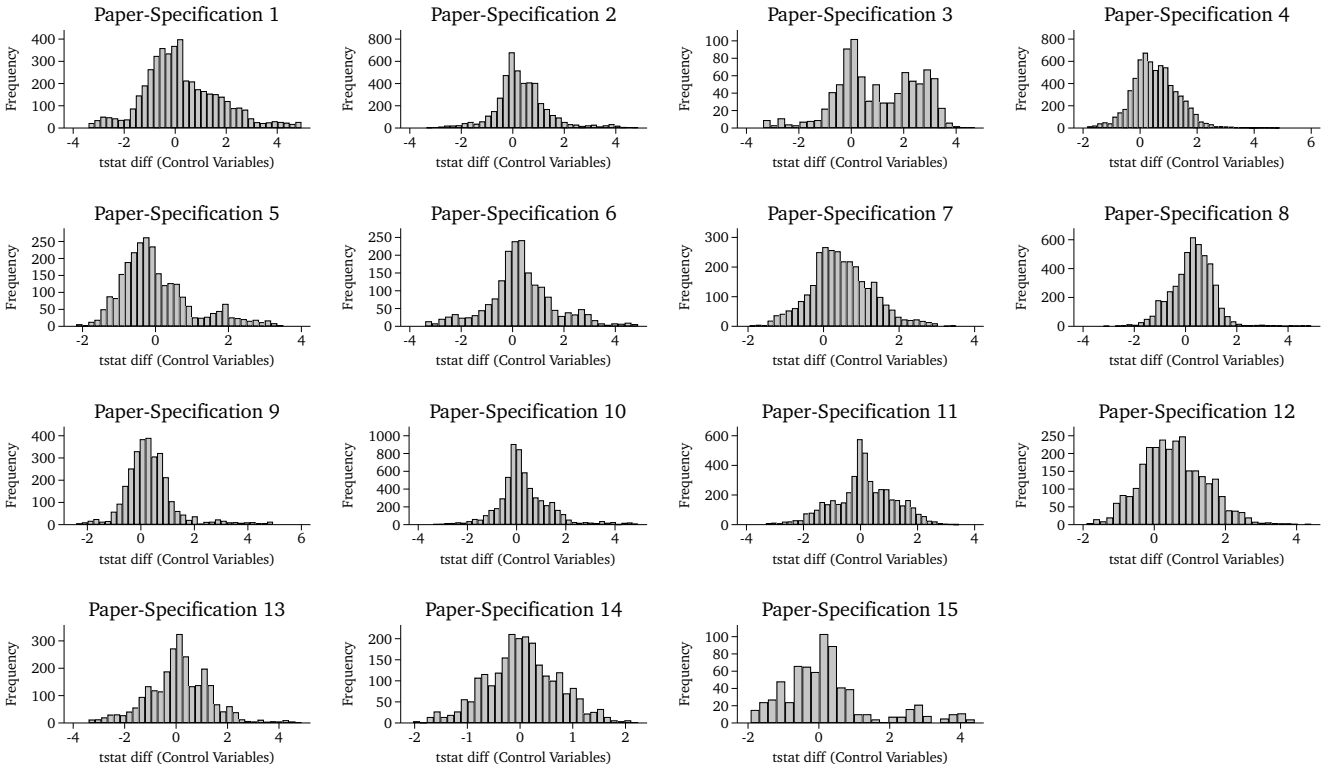
This figure plots the histogram of *tstat_diff* by specification for each of the 15 specifications from replicated papers. Each histogram presents stress tests on the significance of the coefficient on the variable of interest in Panel A, and of the coefficient on control variables in Panel B, by varying the definition of Cumulative Abnormal Returns (CARs). Stress tests allow for within-literature variation on (1) the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ([−205, −6], [−210, −11], [−245, −45], [−252, −20], [−272, −20], [−300, −91], and [−370, −253]). *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Specification numbers are assigned randomly to each specification. Each histogram plots the *tstat_diff* for all stress tests for that specification. Panel A provides histograms for each paper showing the 1st through 99th percentiles of all stress tests for variables of interest. Panel B provides histograms for each paper showing the 1st through 99th percentiles of all stress tests for control variables for each control variable-specification-stress test.



Panel A: Variables of interest

Figure 2

Continued

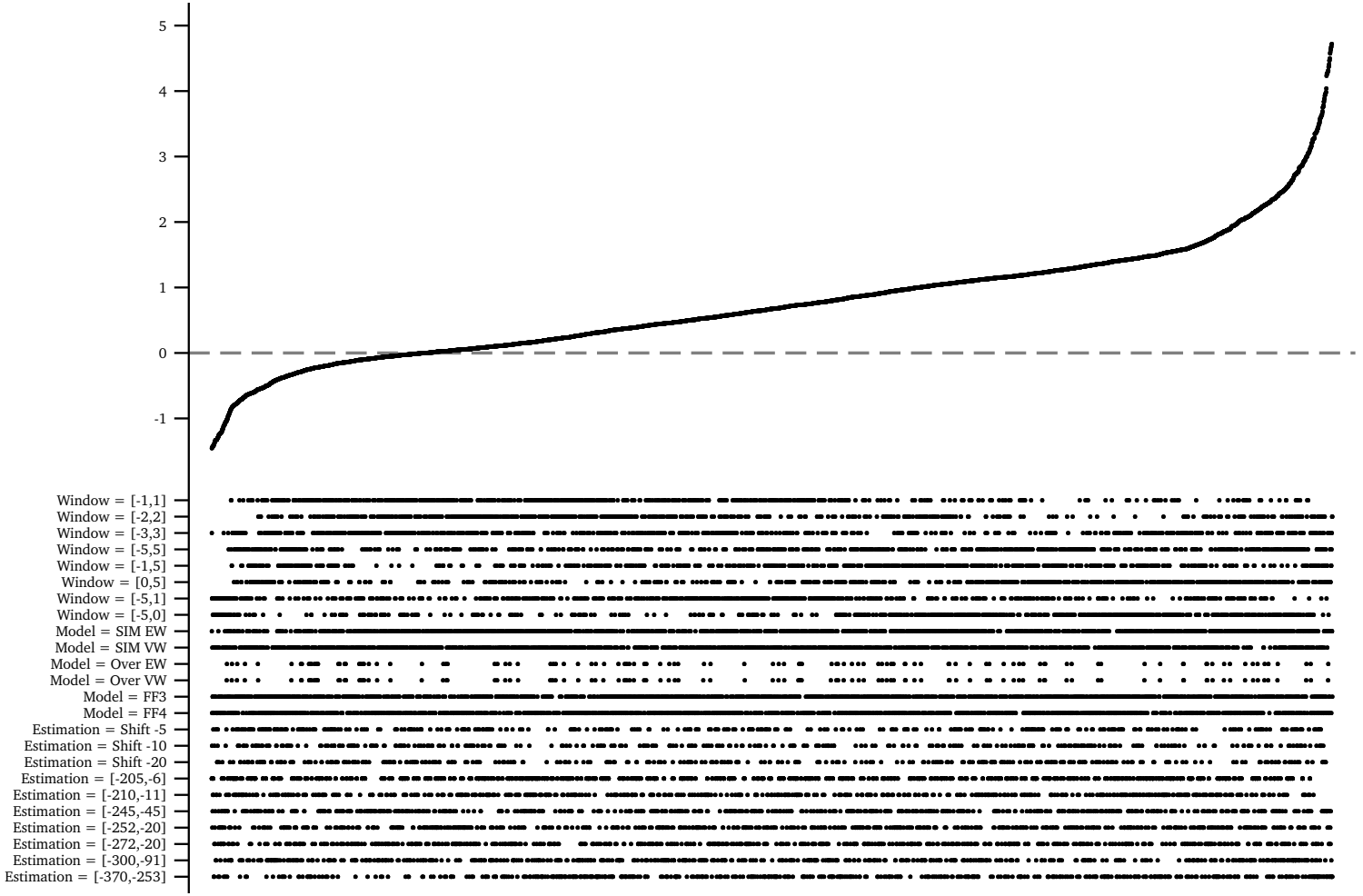


Panel B: Control variables

Figure 3

CAR Stress Test Specification Curve

Each dot in the top portion of the figure depicts the *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers, matching the sample reported in Figure 1, Panel A. The figure also reports a dashed line representing where *tstat_diff* is equal to zero. The vertically aligned dots below this plot indicate the combination of specifications within the given stress test used to compute *tstat_diff*. Thus, a positive *tstat_diff* (to the right of the figure) shows that the stress test, including the combination of specifications reported, yields a lower t-statistic than the t-statistic reported in a given replicated paper.

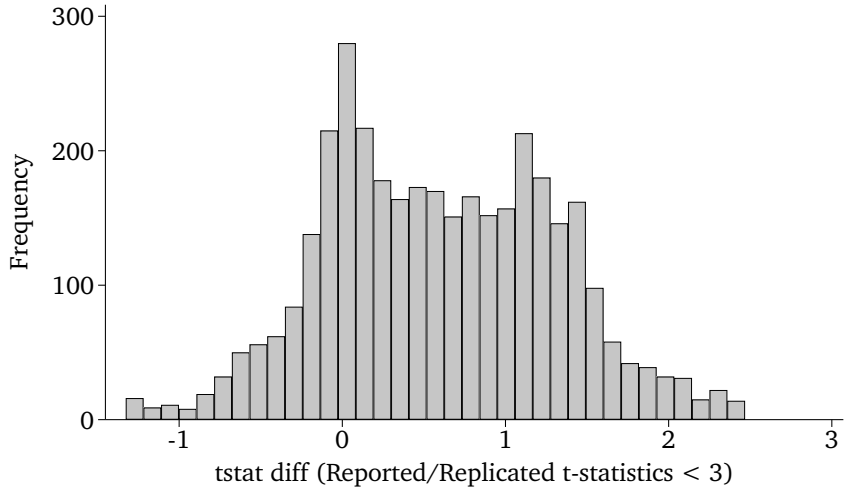


,

Figure 4

CAR Stress Tests Histogram: Variables of Interest with a Replicated t-Statistic < 3

This histogram plots *tstat_diff* for stress tests on the significance of the coefficients on variables of interest by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Stress tests allow for within-literature variation on (1) the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ([−205, −6], [−210, −11], [−245, −45], [−252, −20], [−272, −20], [−300, −91], and [−370, −253]). *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). The figure plots the 1st through 99th percentiles of a subsample of 3,633 stress tests where the reported/replicated t-statistic is < 3.



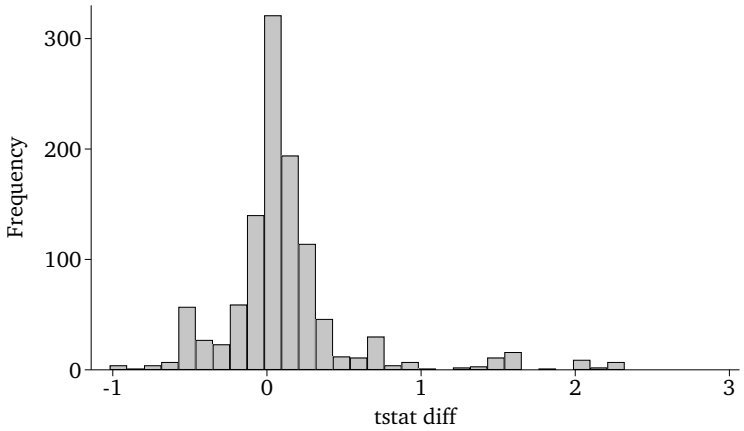
,

Figure 5

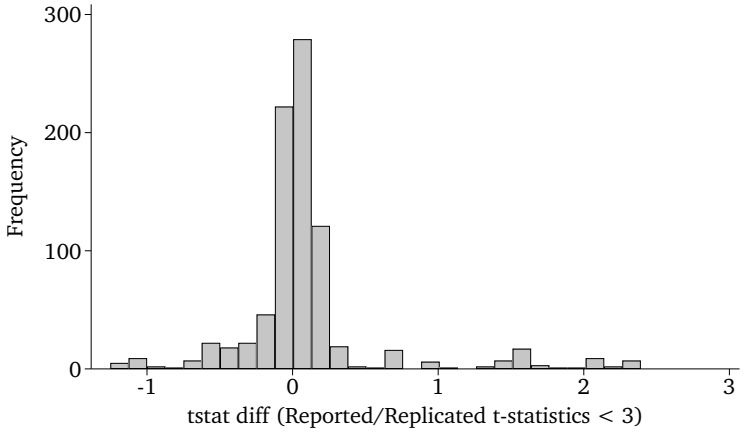
Control Variable Stress Tests: Histogram of the Change in t-statistics of Variables of Interest

This histogram plots *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest by varying the definition of two independent control variables common across the majority of replicated specifications: market-to-book (Tobin's Q) and leverage. These are performed on all specifications that contain these variables (12) from replicated papers. Stress tests allow for within-literature variation based on the definitions presented in Table 3 and Table 4. *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel A plots the 1st through 99th percentiles of all stress tests resulting in 1,136 stress tests. Panel B plots the 1st through 99th percentiles of a subsample of 848 stress tests where the reported/replicated t-statistic is < 3.

,))



Panel A: All original specifications



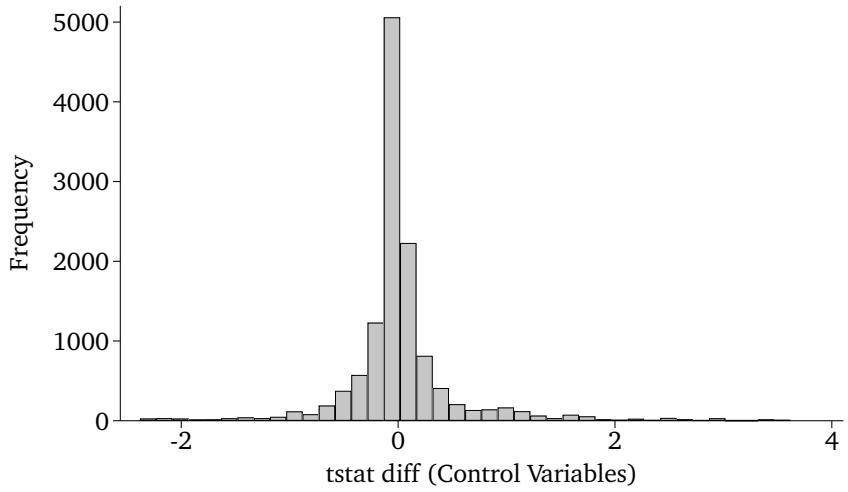
Panel B: Original specifications with t-statistic < 3

,

Figure 6

Control Variable Stress Tests: Histogram of the Change in t-statistics of Control Variables

This histogram plots *tstat_diff* for stress tests on the significance of the coefficients on control variables by varying the definition of two independent control variables common across the majority of replicated specifications: market-to-book (Tobin's Q) and leverage. These are performed on all specifications that contain these variables (12) from replicated papers. Stress tests allow for within-literature variation based on the definitions presented in Table 3 and Table 4. *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). The figure plots the 1st through 99th percentiles of all stress tests resulting in 12,374 stress tests (one *tstat_diff* for each control variable-specification-stress test).



,

Figure 7

Control Variable Stress Test Specification Curve

Each dot in the top portion of the figure depicts the *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest by varying the definition of two independent control variables common across the majority of replicated specifications: market-to-book (Tobin's Q) and leverage. These are performed on all specifications that contain these variables (12) from replicated papers, matching the sample reported in Figure 5, Panel A. The figure also reports a dashed line representing where *tstat_diff* is equal to zero. The vertically aligned dots below this plot indicate the combination of specifications within the given stress test used to compute *tstat_diff*. Thus, a positive *tstat_diff* (to the right of the figure) shows that the stress test, including the combination of specifications reported, yields a lower t-statistic than the t-statistic reported in a given replicated paper.

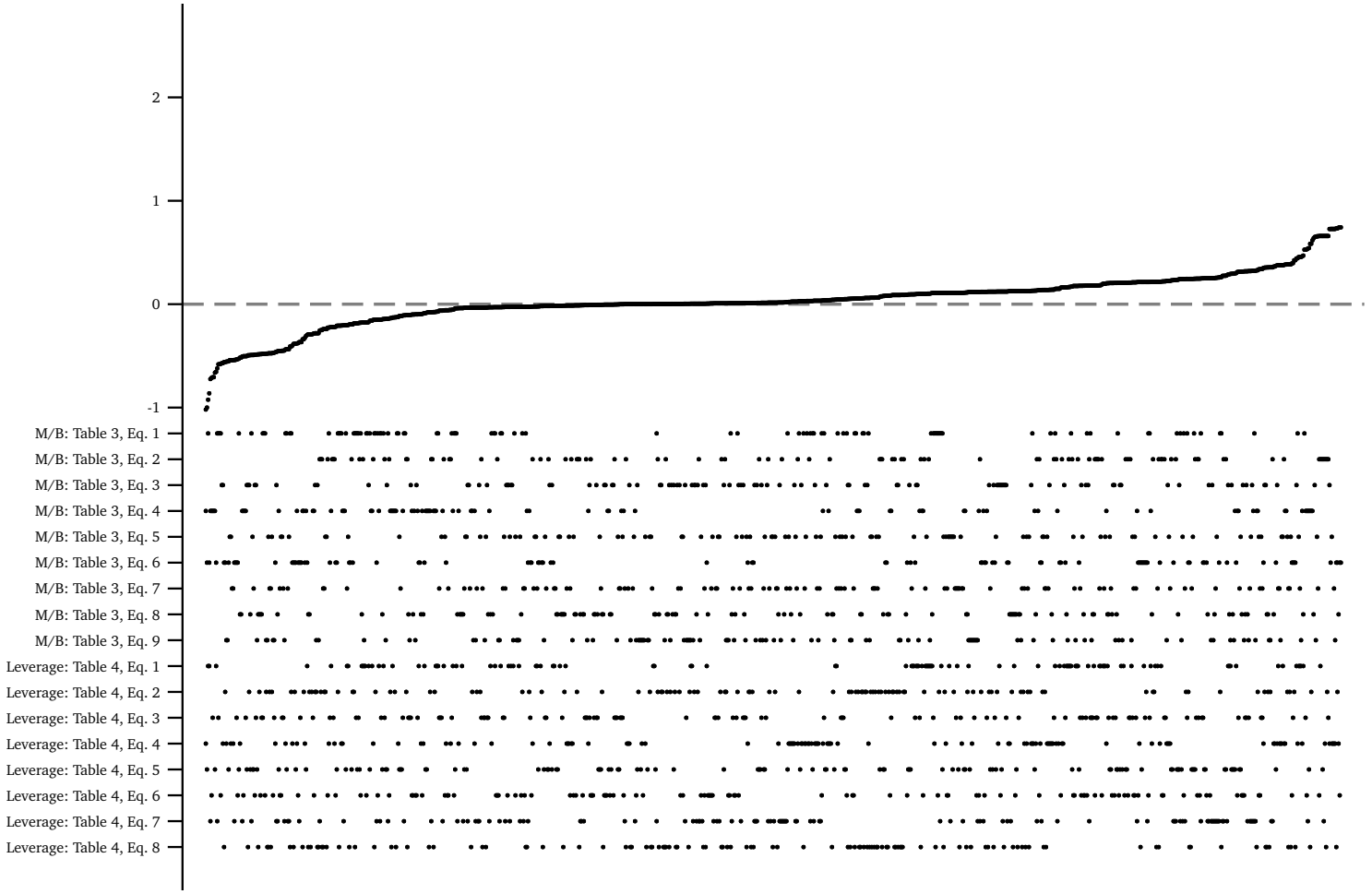
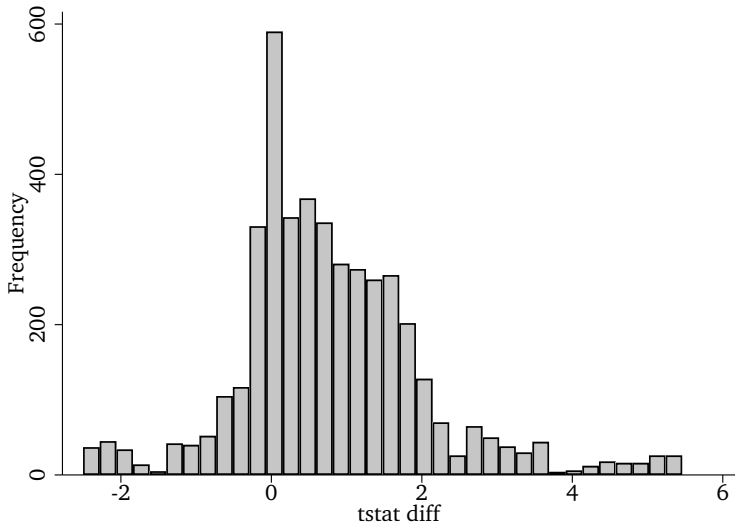


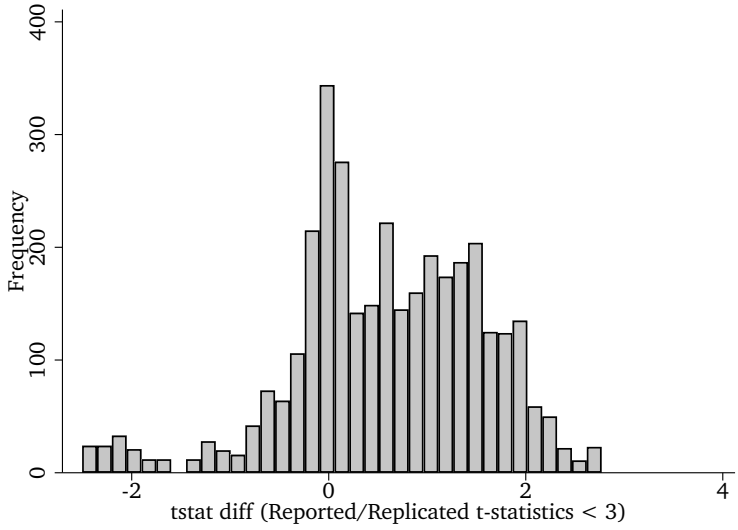
Figure 8

Sample Selection Stress Tests: Histogram of the Change in t-statistics of Variables of Interest

This histogram plots *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest by varying the sample selection criteria for 15 specifications from replicated papers. Sample selection tests allow for within-literature variation on (1) the minimum required relative deal size (none, 1%, or 5%), (2) the maximum allowed relative deal size (none, 150%, or 200%), (3) the deal form (six variations based on of acquisitions, merger, acquisition of majority interest, and acquisition of assets), and (4) deal status (completed or completed and withdrawn) Relative deal size uses different values of the firm using either lagged 11 day CRSP daily market cap, prior month end CRSP monthly market cap, or prior fiscal year Compustat book value of assets. *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel A plots the 1st through 99th percentiles of all stress tests resulting in 4,386 stress tests. Panel B plots the 1st through 99th percentiles of a subsample of 3,508 stress tests where the reported/replicated t-statistic is < 3.



Panel A: All original specifications



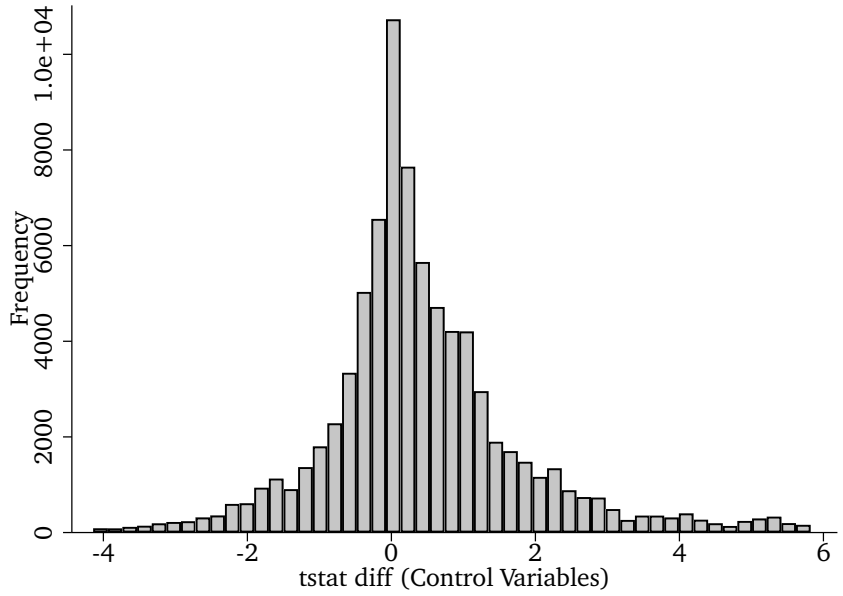
Panel B: Original specifications with t-statistic < 3

,

Figure 9

Sample Selection Stress Tests: Histogram of the Change in t-statistics of Control Variables

This histogram plots *tstat_diff* for stress tests on the significance of the coefficients on control variables by varying the sample selection criteria for 15 specifications from replicated papers. Sample selection tests allow for within-literature variation on (1) the minimum required relative deal size (none, 1%, or 5%), (2) the maximum allowed relative deal size (none, 150%, or 200%), (3) the deal form (six variations based on of acquisitions, merger, acquisition of majority interest, and acquisition of assets), and (4) deal status (completed or completed and withdrawn) Relative deal size uses different values of the firm using either lagged 11 day CRSP daily market cap, prior month end CRSP monthly market cap, or prior fiscal year Compustat book value of assets. *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). The figure plots the 1st through 99th percentiles of all stress tests resulting in 82,042 stress tests (one *tstat_diff* for each control variable-specification-stress test).

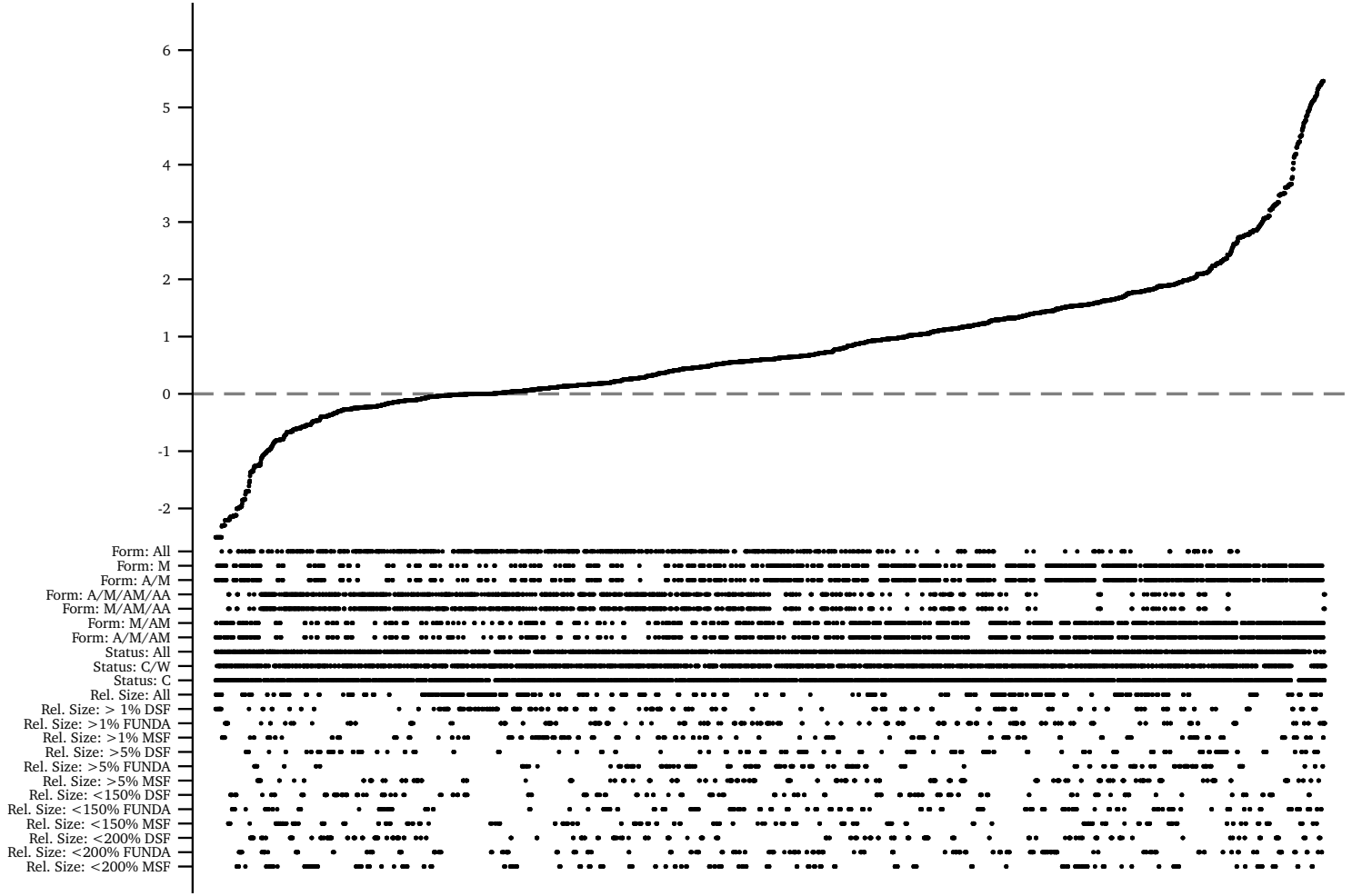


,

Figure 10

Sample Selection Stress Test Specification Curve

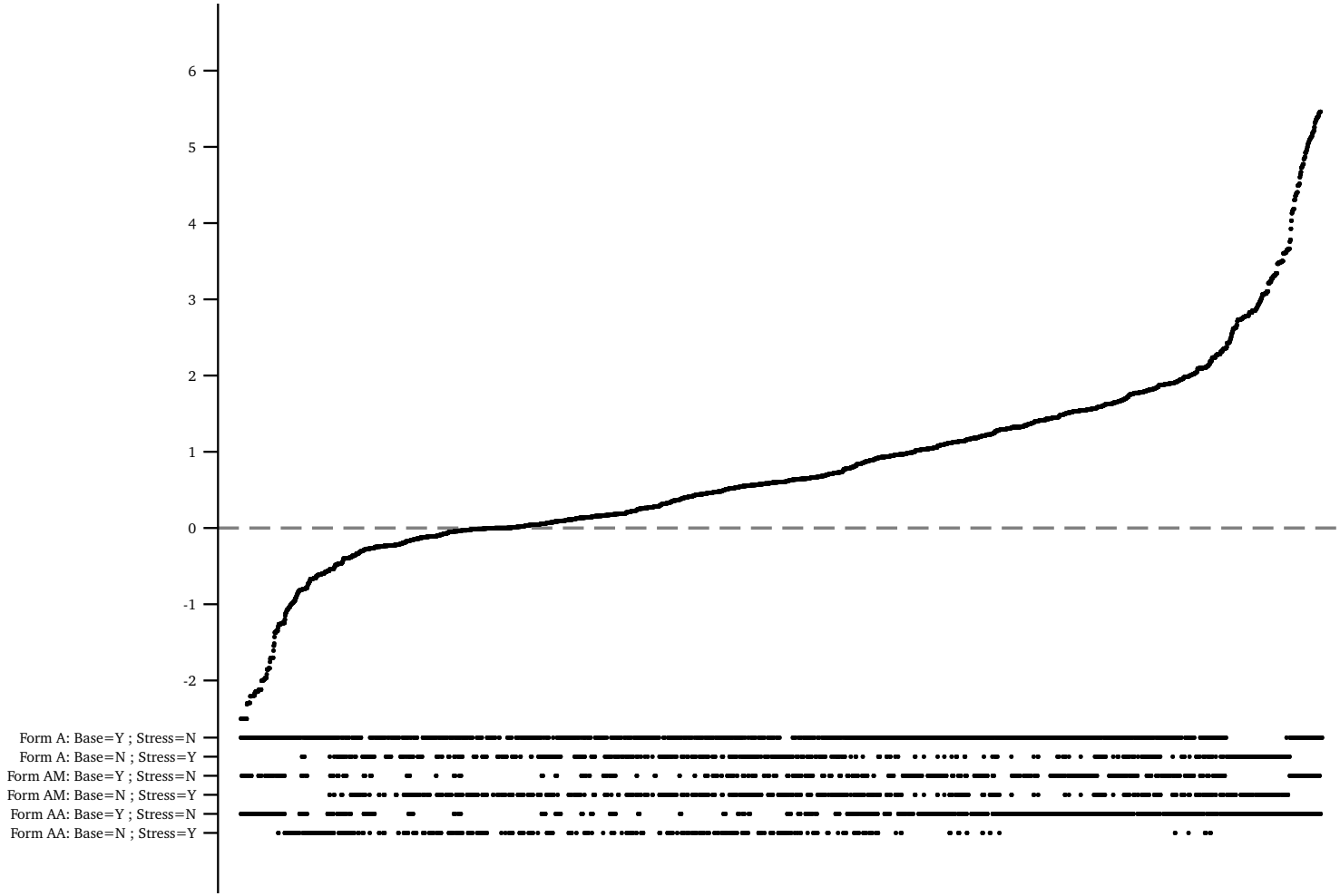
Each dot in the top portion of the figure depicts the *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest by varying the sample selection criteria on 15 specifications from replicated papers, matching the sample reported in Figure 8, Panel A. The figure also reports a dashed line representing where *tstat_diff* is equal to zero. The vertically aligned dots below this plot indicate the combination of specifications within the given stress test used to compute *tstat_diff*. Thus, a positive *tstat_diff* (to the right of the figure) shows that the stress test, including the combination of specifications reported, yields a lower t-statistic than the t-statistic reported in a given replicated paper.



,

Figure 11

Sample Selection Stress Test Specification Curve



,

Table 1

Replication summary.

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and target lockup options. Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

Paper	Key Economic Variable, Table, and Column	Coefficient		Observations	
		Published	Replicated	Published	Replicated
Bates and Lemmon (2003)	Bidder Termination Fee Indicator Table 8, Column (1)	−0.030**	−0.022**	3037	3203
Becher, Griffin, and Nini (2021)	Financial Covenant Violation Table 6, Column (1)	1.860***	1.093**	7191	7299
Burch (2001)	Lockup (0/1) Table 6, Column (6)	−0.012*	−0.017***	744	776
Fuller, Netter, and Stegemoller (2002)	Dummy = 1 if Fifth or Higher Bid Table VII, Column “Private”	−0.019***	−0.018***	2060	1313
Golubov, Yawson, and Zhang (2015)	Ln (Acquirer Size) Table 1, Column “Full Sample”	−0.004***	−0.003***	12 491	14 863
Gorton, Kahl, and Rosen (2009)	Log 123–456 Size Ratio Table V, Column “Harford Waves (4)”	0.029***	0.041**	1334	1141
Harford, Humphery-Jenner, and Powell (2012)	Dictator Dummy Table 5, Column (1)	−0.524**	−0.594**	3934	3258
John, Knyazeva, and Knyazeva (2015)	Acquirer—Weak Labor Rights Table V, Column (3)	0.494***	0.267**	13 838	13 846

Continued

,

Table 1
Continued.

Paper	Key Economic Variable, Table, and Column	Coefficient		Observations	
		Published	Replicated	Published	Replicated
Li, Qiu, and Shen (2018)	OC *(Organizational Capital)* Table 2, Column 1	0.250***	0.306**	17 910	17 609
Ma, Whidbee, and Zhang (2019)	RPR *(Relative Price Ratio)* Table 3, Column "All"	−5.487***	−8.878***	19 119	16 014
Masulis, Wang, and Xie (2007)	GIM Index Table VI, Column (1)	−0.107**	−0.083**	3333	3380
Moeller, Schlingemann, and Stulz (2004)	Small *(Market Capitalization Acquirers)* Table 5, Column 1	0.016***	0.015***	9712	10 796
Nguyen and Phan (2017)	PU_Announcement *(Political Uncertainty)* Table 7, Column (3)	0.007**	0.010**	6376	6674
Roosenboom, Schlingemann, and Vasconcelos (2013)	Stock Liquidity Table 2, Column (3)	−0.038***	−0.006***	3815	4189
Wang and Xie (2009)	Shareholder-Rights Difference Table 5, Column "TCAR"	0.836***	0.594**	396	378

Table 2

Cumulative Abnormal Return Definitions in the M&A Literature.

The table presents the definitions of cumulative abnormal returns (CAR) used in the replicated papers. Panel A lists papers that use excess return models, where the daily abnormal return is the stock return less the market index. Panel B lists papers that use abnormal return models, where the daily abnormal return is the stock return less the return predicted by an empirical model. Event windows, *[start, end]*, indicate the start and end days use to sum abnormal returns. Both *start* and *end* are given in trading days relative to the announcement date. The market model include single index models, *SIM*, and the Fama-French 4 factor model, *FF4*. Estimation periods, *[start, end]*, list the start and end days used to estimate the model parameters. Both days are trading days relative to the announcement date.

#	References	Event Window	Market Index/Model	Model Estimation Period
Panel A: Excess Return Models				
1	Bates and Lemmon (2003), Gorton, Kahl, and Rosen (2009), Li, Qiu, and Shen (2018)	[−1, 1]	CRSP Value Weighted	N/A
2	Burch (2001)	[−1, 2]	CRSP Value Weighted	N/A
3	Fuller, Netter, and Stegemoller (2002)	[−2, 2]	CRSP Value Weighted	N/A
Panel B: Abnormal Return Models				
4	Becher, Griffin, and Nini (2021)	[−1, 1]	SIM CRSP Equal Weighted	[−271, −20]
5	Moeller, Schlingemann, and Stulz (2004)	[−1, 1]	SIM CRSP Equal Weighted	[−205, −6]
6	Nguyen and Phan (2017)	[−1, 1]	SIM CRSP Value Weighted	[−210, −11]
7	Harford, Humphery-Jenner, and Powell (2012), Masulis, Wang, and Xie (2007)	[−2, 2]	SIM CRSP Equal Weighted	[−210, −11]
8	Roosenboom, Schlingemann, and Vasconcelos (2013)	[−2, 2]	SIM CRSP Value Weighted	[−245, −46]
9	John, Knyazeva, and Knyazeva (2015)	[−2, 2]	SIM CRSP Value Weighted	[−210, −11]

Continued

Table 2
Continued.

#	References	Event Window	Market Index/Model	Model Estimation Period
10	Golubov, Yawson, and Zhang (2015)	$[-2, 2]$	SIM CRSP Value Weighted	$[-300, -91]$
11	Ma, Whidbee, and Zhang (2019)	$[-5, 1]$	SIM CRSP Equal Weighted	$[-370, -253]$
12	Wang and Xie (2009)	$[-5, 5]$	SIM CRSP Value Weighted	$[-210, -11]$

Table 3

Market-to-book ratio (Tobin's q) control variable alternative definitions used for stress testing.

The table presents the various definitions of market-to-book ratio used in the literature. This ratio is also referred to as Tobin's q in the literature. For each definition, the table lists the formula as `datasource.variable`. `Funda` indicates a variable sourced from Compustat Fundamentals Annual Data, and `Fundq` indicates a variable sourced from Compustat Fundamentals Quarterly data. `at` is the total book value of assets, `ceq` is the total book value of common/ordinary equity, `csho` is the total number of shares of common/ordinary equity as of the fiscal year end, `dlc` is the total book value of debt in current liabilities, `dltt` is the total book value of long-term debt, `lt` is the total book value of liabilities, `prccF` is the closed sharing price at the end of the fiscal year, `txdb` is the balance sheet value of deferred taxes, and `txditc` is the value of deferred taxes and investment tax credit. Missing values of `txdb` and `txditc` are treated as zeros. `preferred` is the liquidating value of preferred stock (`Funda. pstkl`) if it is available, otherwise it is the redemption value (`Funda. pstkrv`).

#	Formula References
1	`(Fundq.at - Fundq.ceq + Fundq.csho` × `Fundq.prccF) / Fundq.at` Becher, Griffin, and Nini (2021)
2	`(Funda.at - Funda.ceq + Funda.csho` × `Funda.prccF) / Funda.at` Burch (2001), Harford, Humphery-Jenner, and Powell (2012), Masulis, Wang, and Xie (2007), Moeller, Schlingemann, and Stulz (2004)
3	`(Funda.at + (Funda.prccF` × `Funda.csho) - Funda.ceq - Funda.txdb) / Funda.at` Golubov, Yawson, and Zhang (2015)
4	`Funda.at + (Funda.prccF` × `Funda.csho) - Funda.ceq - Funda.txdb / Funda.ceq` Gorton, Kahl, and Rosen (2009)
5	`Funda.csho` × `Funda.prccF / Funda.at` Li, Qiu, and Shen (2018)
6	`ln (Funda.csho` × `Funda.prccF / Funda.ceq)` Ma, Whidbee, and Zhang (2019)
7	`ln (Funda.csho` × `Funda.prccF / Funda.ceq)` Nguyen and Phan (2017)
8	`(Funda.lt - txditc + preferred + Funda.csho` × `Funda.prccF) / Funda.at` Roosenboom, Schlingemann, and Vasconcelos (2013)
9	`(Funda.at - Funda.ceq + Funda.csho` × `Funda.prccF) / Funda.at` Wang and Xie (2009)

Table 4

Leverage control variable alternative definitions used for stress testing.

The table presents the various definitions of leverage ratios used in the literature. For each definition, the table lists the formula as `datasource.variable`. `Funda` indicates a variable sourced from Compustat Fundamentals Annual Data, and `Fundq` indicates a variable sourced from Compustat Fundamentals Quarterly data. `Funda.at` (`Funda.atq`) is the annual (quarterly) total book value of assets, `Funda.ceq` is the total annual book value of common/ordinary equity, `Funda.csho` is the total annual number of shares of common/ordinary equity as of the fiscal year end, `Funda.dlc` (`Funda.dlcq`) is the total annual (quarterly) book value of debt in current liabilities, `Funda.dltt` (`Funda.dlttq`) is the total annual (quarterly) book value of long-term debt, `Funda.lt` is the total annual book value of liabilities, `Funda.prccF` is the closed sharing price at the end of the fiscal year, `Funda.txditc` is the annual value of deferred taxes and investment tax credit. Missing values of `Funda.txditc` are treated as zeros. `preferred` is the liquidating value of preferred stock (`Funda.pstkl`) if it is available, otherwise it is the redemption value (`Funda.pstkrv`).

#	Formula References
1	`(Fundq.dlttq + Fundq.dlcq) / Fundq.atq` Becher, Griffin, and Nini (2021)
2	`(Funda.dlc + Funda.dltt) / Funda.at` Burch (2001), Harford, Humphery-Jenner, and Powell (2012), Nguyen and Phan (2017)
3	`(Funda.dltt + Funda.dlc) / (Funda.lt - txditc + preferred + Funda.csho` $\times$ `Funda.prccF)` Golubov, Yawson, and Zhang (2015), Roosenboom, Schlingemann, and Vasconcelos (2013)
4	`(Funda.at - Funda.ceq) / Funda.at` Ma, Whidbee, and Zhang (2019)
5	`(Funda.dlc + Funda.dltt) / (Funda.at + Funda.csho` $\times$ `Funda.prccF)` Masulis, Wang, and Xie (2007)
6	`(Funda.dltt + Funda.dlc) / (Funda.at - Funda.ceq + Funda.csho` $\times$ `Funda.prccF)` Moeller, Schlingemann, and Stulz (2004)
7	`Funda.dlc + Funda.dltt / (Funda.dlc + Funda.dltt + Funda.csho` $\times$ `Funda.prccF)` Li, Qiu, and Shen (2018)
8	`(Funda.dlc + Funda.dltt) / Funda.at` Wang and Xie (2009)

Table 5

Sample selection criteria used in the literature.

The table presents the sample selection criteria imposed by the replicated specifications reported in Table 1. Each row gives one criterion along with the specifications that impose it. `Sdc.dealValue` is the transaction value reported by SDC in millions of dollars. `Dsf` (`Msf`) indicates a variable sourced from the CRSP daily (monthly) stock file, and `Funda` indicates a variable sourced from Compustat Fundamentals Annual data. Subscripts give the measurement date: $t-n$ is n trading days before the announcement date, $y-1$ is the prior calendar year end, and $fy-1$ is the prior fiscal year end. Acquirer size measures are converted to millions of dollars. Nguyen and Phan (2017) specify the acquirer measurement date as four weeks before the announcement date, which is reported as $t-20$. Deal form and deal status categories are the values reported by SDC. The specifications impose additional conditions that are not varied in the sample selection stress tests and are not reported above, including the announcement period, the minimum transaction value, acquirer and target ownership thresholds, industry exclusions, target public status requirements, single-bid requirements, and data requirements for estimating abnormal returns.

#	Criterion References
	Panel A: Relative deal size
1	`Sdc.dealValue / (Dsf.prc` $\times$ `Dsf.shrout)`$_{t-6}$ $\geq 1\%$ Wang and Xie (2009).
2	`Sdc.dealValue / (Dsf.prc` $\times$ `Dsf.shrout)`$_{t-11}$ $\geq 1\%$ Becher, Griffin, and Nini (2021), Golubov, Yawson, and Zhang (2015), Harford, Humphery-Jenner, and Powell (2012), John, Knyazeva, and Knyazeva (2015), Masulis, Wang, and Xie (2007).
3	`Sdc.dealValue / (Dsf.prc` $\times$ `Dsf.shrout)`$_{t-20}$ $\geq 1\%$ Nguyen and Phan (2017).
4	`Sdc.dealValue / (Msf.prc` $\times$ `Msf.shrout)`$_{y-1}$ $\geq 5\%$ and $\leq 150\%$ Gorton, Kahl, and Rosen (2009).
5	`Sdc.dealValue / (Msf.prc` $\times$ `Msf.shrout)`$_{y-1}$ $\geq 5\%$ and $\leq 200\%$ Ma, Whidbee, and Zhang (2019).
6	`Sdc.dealValue / Funda.at`$_{fy-1}$ $\geq 1\%$ Li, Qiu, and Shen (2018), Roosenboom, Schlingemann, and Vasconcelos (2013).
7	`Sdc.dealValue / (Funda.csho` $\times$ `Funda.prccF + Funda.dltt + Funda.dlc + Funda.pstkl)`$_{fy-1}$ $\geq 1\%$ Moeller, Schlingemann, and Stulz (2004).
8	`No restriction/Not specified` Bates and Lemmon (2003), Burch (2001), Fuller, Netter, and Stegemoller (2002).

Continued

,)

Table 5
Continued.

#	Criterion References
Panel B: Deal form	
1	`Merger` Burch (2001).
2	`Merger, Acquisition` Bates and Lemmon (2003).
3	`Merger, Acq. Maj. Int., Acq. of Assets` Li, Qiu, and Shen (2018).
4	`Merger, Acquisition, Acq. Maj. Int., Acq. of Assets` Becher, Griffin, and Nini (2021).
5	All except `Buyback`, `Exchange Offer`, and `Recapitalization` John, Knyazeva, and Knyazeva (2015), Roosenboom, Schlingemann, and Vasconcelos (2013).
6	`No restriction/Not specified` Fuller, Netter, and Stegemoller (2002), Gorton, Kahl, and Rosen (2009), Golubov, Yawson, and Zhang (2015), Harford, Humphery-Jenner, and Powell (2012), Moeller, Schlingemann, and Stulz (2004), Masulis, Wang, and Xie (2007), Ma, Whidbee, and Zhang (2019), Nguyen and Phan (2017), Wang and Xie (2009).
Panel C: Deal status	
1	`Completed` Fuller, Netter, and Stegemoller (2002), Gorton, Kahl, and Rosen (2009), Golubov, Yawson, and Zhang (2015), Harford, Humphery-Jenner, and Powell (2012), Li, Qiu, and Shen (2018), Moeller, Schlingemann, and Stulz (2004), Masulis, Wang, and Xie (2007), Wang and Xie (2009).
2	`Completed, Withdrawn` Bates and Lemmon (2003), Burch (2001).
3	`No restriction/Not specified` Becher, Griffin, and Nini (2021), John, Knyazeva, and Knyazeva (2015), Ma, Whidbee, and Zhang (2019), Nguyen and Phan (2017), Roosenboom, Schlingemann, and Vasconcelos (2013).

,

Table 6

CAR Stress Tests

This table reports summary statistics on stress tests on the significance of the coefficient on the variable of interest and control variables by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Stress tests allow for within-literature variation on (1) the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ([−205, −6], [−210, −11], [−245, −45], [−252, −20], [−272, −20], [−300, −91], and [−370, −253]). Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
Variables of Interest	4543	0.774	1.052	1.257	7.802
Control Variables	52 234	0.336	1.390	1.347	16.592

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
Variables of Interest	−3.711	−1.475	−0.560	−0.228	0.087	0.674	1.261	1.888	2.512	4.720	6.634
Control Variables	−14.243	−3.397	−1.430	−0.948	−0.291	0.222	0.893	1.675	2.352	4.946	15.480

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
Variables of Interest	80.176	55.775	36.897	2.838
Control Variables	61.418	16.202	11.496	12.327

Table 7

Comparing the Distributions of Variables of Interest and Control Variables in CAR Stress Tests

This table reports a two-sample Kolmogorov-Smirnov test for the equality of distributions between the distribution of *tstat_diff* on variables of interest and the distribution of *tstat_diff* on control variables (corresponding to those described in Table 6).

Group	D	p-value
Controls	0.197	$<$0.001
Variables of Interest	$-$0.004	0.816
Combined K-S	0.197	$<$0.001

Table 8

Individual Paper Mean *tstat_diff* Comparisons

This table reports the mean *tstat_diff* of variables of interest, control variables, their differences, and the related p-value individually for each randomized replicated paper (corresponding to those presented in Figure 2, Panel A). Difference marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

Paper	Variable of Interest Mean *tstat_diff*	Control Variable Mean *tstat_diff*	Difference	p-value
1	1.914	0.466	1.447***	<0.001
2	0.229	0.253	−0.024	0.727
3	−0.081	0.970	−1.052***	<0.001
4	0.861	0.573	0.288***	<0.001
5	0.907	0.034	0.872***	<0.001
6	0.681	0.162	0.518***	<0.001
7	0.040	0.471	−0.431***	<0.001
8	0.576	0.318	0.258***	<0.001
9	0.904	0.503	0.401***	<0.001
10	0.896	0.331	0.564***	<0.001
11	1.720	0.092	1.628***	<0.001
12	1.575	0.565	1.009***	<0.001
13	0.707	−0.000	0.708***	<0.001
14	0.542	0.045	0.497***	<0.001
15	0.221	0.250	−0.028	0.737

Table 9

CAR Stress Tests: Replicated t-Statistics < 3

This table reports summary statistics on stress tests on the significance of coefficients on variables of interest with replicated t-statistics < 3 by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Stress tests allow for within-literature variation on (1) the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ([−205, −6], [−210, −11], [−245, −45], [−252, −20], [−272, −20], [−300, −91], and [−370, −253]). Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
Variables of Interest	3633	0.597	0.762	0.089	3.156

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
Variables of Interest	−1.966	−1.327	−0.564	−0.265	0.040	0.560	1.154	1.527	1.830	2.486	3.220

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
Variables of Interest	77.744	65.612	46.052	3.548

Table 10

Do Higher Hurdle Rates Help?

This table reports the potential flexibility of specification choice available to authors using the distribution of *tstat_diff* from the Cumulative Abnormal Return (CAR) stress tests reported in Table 6. Specifically, the table reports the proportion of stress tests that would result in an t-statistic being significant just above a threshold to just below another significance threshold. Panel A reports standard significance thresholds, Panel B reports the 95% significance threshold motivated from Andrews and Kasy (2019) (A-K (2019)) by assuming variable of interest estimates that are significant at the 95% are 100× more likely to be published than estimates < 95%, and Panel C reports the 99% significance threshold motivated from A-K (2019) assuming variable of interest estimates that are significant at the 95% are 100× more likely to be published than estimates that are < 95% significant.

Panel A: Standard hurdles	Proportion of Stress Tests
99% (t = 2.58) to < 99% (t = 2.58)	80.29%
99% (t = 2.58) to < 95% (t = 1.96)	52.15%
99% (t = 2.58) to < 90% (t = 1.65)	39.87%
Panel B: A-K (2019) 95% hurdle to standard hurdles	Proportion of Stress Tests
A-K 95% (t = 3.02) to < 99% (t = 2.58)	60.00%
A-K 95% (t = 3.02) to < 95% (t = 1.96)	34.50%
A-K 95% (t = 3.02) to < 90% (t = 1.65)	20.95%
Panel C: A-K (2019) 99% hurdle to standard hurdles	Proportion of Stress Tests
A-K 99% (t = 3.48) to < 99% (t = 2.58)	40.86%
A-K 99% (t = 3.48) to < 95% (t = 1.96)	15.80%
A-K 99% (t = 3.48) to < 90% (t = 1.65)	10.69%

Table 11

Control Variable Definition Stress Tests: Variables of Interest

This table reports summary statistics on stress tests on the significance of the coefficient on the variable of interest by varying the definition of two independent control variables common across the majority of replicated specifications: market-to-book (Tobin's Q) and leverage. These are performed on all specifications that contain these variables (12) from replicated papers. Stress tests allow for within-literature variation based on the definitions presented in Table 3 and Table 4. Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
All specifications	1136	0.132	0.533	2.336	11.693
<3	848	0.106	0.499	2.068	9.994

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
All specifications	−1.251	−1.047	−0.501	−0.338	−0.029	0.032	0.207	0.546	1.446	2.320	3.473
<3	−1.251	−1.088	−0.506	−0.251	−0.029	0.025	0.135	0.323	1.502	2.145	2.391

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
All specifications	62.676	19.894	10.651	2.288
<3	61.320	24.764	14.268	3.301

Table 12

Control Variable Definition Stress Tests: Control Variables

This table reports summary statistics on stress tests on the significance of coefficients on control variables by varying the definition of two independent control variables common across the majority of replicated specifications: market-to-book (Tobin's Q) and leverage. These are performed on all specifications that contain these variables (12) from replicated papers. Stress tests allow for within-literature variation based on the definitions presented in Table 3 and Table 4. Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
All specifications	12 784	0.044	0.836	1.021	31.564

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
All specifications	−9.492	−2.381	−0.696	−0.386	−0.111	−0.001	0.086	0.484	1.097	3.618	11.730

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
All specifications	46.127	5.006	2.933	5.178

Table 13

Comparing the Distributions of Variables of Interest and Control Variables in Control Variable Stress Tests

This table reports a two-sample Kolmogorov-Smirnov test for the equality of distributions between the distribution of *tstat_diff* on variables of interest (corresponding to those described in Table 11) and the distribution of *tstat_diff* on control variables (corresponding to those described in Table 12).

Group	D	p-value
Controls	0.190	$<$0.001
Variables of Interest	$-$0.014	0.661
Combined K-S	0.190	$<$0.001

,

Table 14

Sample Selection Stress Tests: Variables of Interest

This table reports summary statistics on stress tests on the significance of the coefficient on the variable of interest by varying the sample selection criteria across (1) minimum required relative deal size, (2) maximum allowed relative deal size, (3) deal form, and (4) deal status. Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
All specifications	4386	0.816	1.387	0.871	5.555
<3	3508	0.577	1.045	−0.661	4.178

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
All specifications	−3.408	−2.503	−1.046	−0.398	0.001	0.604	1.450	2.315	3.342	5.461	6.589
<3	−3.408	−2.503	−1.257	−0.536	−0.003	0.588	1.324	1.827	2.094	2.769	3.031

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
All specifications	75.000	52.800	38.957	6.944
<3	72.879	63.062	48.690	8.679

Table 15

Sample Selection Stress Tests: Control Variables

This table reports summary statistics on stress tests on the significance of coefficients on control variables by varying the the sample selection criteria across (1) minimum required relative deal size, (2) maximum allowed relative deal size, (3) deal form, and (4) deal status. Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N			Mean		Std. Dev.		Skewness		Kurtosis	
Panel A: Summary statistics											
All specifications	82 042			0.425		1.664		0.909		9.344	
	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
All specifications	−8.098	−4.142	−1.835	−1.081	−0.280	0.202	1.009	2.193	3.364	5.821	12.100
	Increase in tstat_diff			Drop in Significance		Loss of Significance		Increase in Significance			
Panel C: Changes in statistical significance thresholds (%)											
All specifications	61.358			16.421		11.285		12.378			

Table 16

Comparing the Distributions of Variables of Interest and Control Variables in Sample Selection Stress Tests

This table reports a two-sample Kolmogorov-Smirnov test for the equality of distributions between the distribution of *tstat_diff* on variables of interest (corresponding to those described in Table 14) and the distribution of *tstat_diff* on control variables (corresponding to those described in Table 15).

Group	D	p-value
Controls	0.170	$<$0.001
Variables of Interest	$-$0.006	0.691
Combined K-S	0.170	$<$0.001

,

Stress Testing an Economic Literature

Vanitha Ragunathan
University of Queensland

Jared Stanfield
University of Oklahoma

Robert Tumarkin[§§§§]
UNSW Sydney

November 18, 2026

Internet Appendix

* Corresponding Author. UNSW Business School, UNSW Sydney, Sydney NSW 2052, Australia. Telephone: +(61) 02 9065 1656. E-mail: r.tumarkin@unsw.edu.au.

A Technical Implementation Details

Traditional econometric code embodies a form of dramatic irony; the researcher has information about the data that the programming language does not. For example, a researcher will know that share prices are represented as floating-point numbers, while ticker symbols are represented as text. However, a statement adding the two (i.e., `price + ticker`) is syntactically valid in SQL, SAS, and other data manipulation languages. The language will only recognize the error when the code is executed against the data. This inability of traditional data manipulation languages to reason about the data before execution significantly inhibits their ability to write expressive, succinct econometric specifications.

Declarative econometrics bridges the knowledge gap between the researcher and the computer by enabling the language to reason about a specification before execution, that is, as the researcher develops the specification. The *Foghorn* source code defines all functions and operations at two levels. The "type level" describes rules that the variables involved in a function must satisfy. These rules can embed logic using cross-sectional indices, time-series indices, data types, and other variable-specific information. The "term level" implements the actual function or operation. For instance, *Foghorn* implements algebraic addition similarly to most programming languages at the term level. However, its type-level rule stipulates that the operands must have identical cross-sectional and/or time-series indices and be of numeric data types (i.e., information that can be added). Thus, adding a numeric share price to a textual ticker symbol is forbidden.

It is the ability to encode logic about the data that makes declarative econometrics powerful, expressive, and succinct. Returning to Example 1, verbose data manipulation steps to convert variables from one set of indices to another are noticeably absent. However, while some programming languages achieve syntactical succinctness by ignoring information about the data, this declarative econometric code achieves simplicity for the opposite reason: it can reason about the data. *Foghorn* would not let a researcher combine CRSP data with Compustat data to compute Tobin's q without explicit reindexing. Moreover, the `reindex` function does not require detailed instructions, as the language can determine and validate the necessary conversions.

Declarative econometrics offers a general framework applicable across all social science disciplines. At its core, declarative econometrics elevates econometric information from the data to the language. However, the approach does not achieve this by restricting the datasets with which it can work. *Foghorn*'s core is database-agnostic, defining general rules that validate relationships among specifications and variables. Dataset-specific information enters the language through topic modules. For example, the corporate finance module, *Coficat*, provides information about many

finance datasets.[23] Users may write modules targeting specific research areas and extend the language's capabilities.

Foghorn will be publicly available for use by econometric researchers. However, we recognize that other researchers may write new modules for *Foghorn* or create other declarative econometric languages. Therefore, when describing how *Foghorn* operates, we emphasize those features that make a declarative econometric approach viable. A complete technical discussion on the design and implementation of the language is beyond the scope of this paper. Details are provided in the companion paper (Tumarkin 2026).

A.1 Embedded domain-specific language

Instead of writing a declarative language from scratch, we believe it is best to leverage existing technology with an embedded domain-specific language (EDSL). An EDSL operates within a host language, serving as a dialect that facilitates specific tasks. The EDSL allows a programmer to revert to the host language when necessary. Therefore, a declarative econometric EDSL can offer data manipulation and estimation tools based on a precise top-down specification while preserving the full capabilities of the host general-purpose programming language.

We implement *Foghorn* as an EDSL within Haskell, a language that has strong support for declarative programming. *Foghorn* is designed for social science researchers to use without requiring knowledge of Haskell. As exemplified by the sample code in Example 1, most specifications can be written in a syntax similar to existing econometric packages with the addition of algebraic constructs to define variables.[24]

Haskell's many unique features make it a desirable host for a declarative econometric EDSL. These include, but are not limited to, lazy evaluation, general algebraic data types, recursive data types, higher-order functions, abstract type classes, and dependent types. However, a technical discussion of the advantages these features provide for a declarative econometric language is deferred to the companion paper (Tumarkin 2026).

Most notably, for this discussion, Haskell code exists at both the term level and the type level. This separation is a key aspect of Haskell's powerful and expressive type system. Most people are familiar with what Haskell considers the term-level, which, in traditional languages, is where the steps of a function are performed. The type-level in Haskell is a unique environment that can be

[23] *Coficat* is tortured wordplay for a *co*-rporate *fi*-nance research copy-*cat* tool.

[24] *Foghorn* is an open-source programming language that accepts contributions that identify issues, fix bugs, and expand its scope. While *Foghorn* is intended to be used for econometric research without learning Haskell, understanding the source code of the core language requires intermediate to advanced Haskell. This is not uncommon in programming languages where the language's source code is often far more complicated than programs written in the language. For example, many Python users would have trouble working through the source to Python's most popular implementation, CPython, which is not written in Python but C.

extended to embed logic. Working at the type-level, one can describe intricate relationships among the inputs to and the output from functions. *Foghorn* uses this to analyze, validate, and infer critical aspects of econometric specifications.

A.2 Encoding econometric logic (Database agnostic reasoning in the core language)

Foghorn is designed to be usable across the social sciences. Thus, it contains a "core" language that emphasizes general econometric logic, lacking information about a specific data set or research area. *Foghorn*'s core provides the building blocks of econometric analysis, including specification design, algebraic calculations, data aggregation, data merging, and other common operations. By tracking index and data type information, the core can achieve a level of concision that is not possible when a language lacks such information.

A declarative econometric language like *Foghorn* needs to reason about the data. In programming language design, this ability is most powerfully expressed through type-level logic. In *Foghorn*, each variable is declared with type-level information about its indexing variables i (i.e., cross-sectional and/or time-series indices) and data type d (e.g., integer, floating-point number, text). We notate a variable as `Var i d`, where, for example, CRSP stock share price would be declared at the type-level as `Var (Permno, TradingDate) Float`. Variables carry this type-level information around in the code, with the language tracking and reasoning about types throughout the specification.

All operations and functions in *Foghorn* contain "type signatures" that define rules about the relationships among input variables and output at the type level. Type signatures are declared as

```
Constraints => Input 1 -> Input 2 -> … -> Input N -> Output.
```

Functions may include any number of constraints and inputs, with only the output being mandatory. For example, the type signature for addition, subtraction, and multiplication is written as:

$$\underbrace{\texttt{Numeric d}}_{\text{Constraint}} \texttt{ => } \underbrace{\texttt{Var i d}}_{\text{Input 1}} \texttt{ -> } \underbrace{\texttt{Var i d}}_{\text{Input 2}} \texttt{ -> } \underbrace{\texttt{Var i d}}_{\text{Output}}.$$

The type constraint, `Numeric d`, is left of the double-arrow and indicates that the data type `d` must be numerical. On the right-hand side, there are two input variables and an output variable, each of type `Var i d`. *Foghorn* can reason that these operations are valid only when the input share indexing identifiers `i` and a numeric data type `d` are used, and the result is a new variable of the same indexing and data type.

Rules in *Foghorn* are often written using type-level variables (a language feature called parametric polymorphism) instead of specific indices or data types.[25] The above example has two type-level variables, `i` and `d`. Type-level variables ensure that rules are as general as possible, and *Foghorn* can

[25]Type-level variables are distinct from term-level variables. Term-level variables express values, permitting the language to manipulate values. Type-level variables express types, permitting the language to reason about type information.

apply them to any dataset. *Foghorn* has a rich, extendable list of data types that enable it to precisely reason about econometric specifications. For example, many kinds of economic data are stored as integers. These include standard industry classification (SIC) codes, fiscal years, calendar years, many identifiers (e.g., CIK, permno, and gvkey), and actual integer numbers. *Foghorn* separates these into separate types. The function to compute Fama-French industry classifications (Fama and French 1992) will only work with SIC codes. Arithmetic operations are only supported for types where the operations make sense (e.g., addition of SIC codes is not permitted).

Type constraints may be of arbitrary complexity due to Haskell's support for advanced programming at the type level. Yet, *Foghorn* does not require that users worry about type-level programming. *Foghorn* inherits Haskell's ability to infer types. For example, consider the reindexing function which has the type signature:

```
Reindexable i i' => Var i d -> Var i' d
```

The type constraint `Reindexable i i'` defines a type class. A source index variable set `i` is paired with a target one `i'` in `Reindexable` only when a variable indexed by observation identifier `i` can be transformed into one indexed by identifier `i'`. Type inference means that *Foghorn* can contextually determine the target indexing variables from reindexing. For example, assume we have variables `a` and `b` indexed by `i` and `i'`, respectively. Type-inference ensures that the expression `reindex a + b` is fully defined. *Foghorn* knows the source indexing for `a` (each variable carries information about index type), and it can infer the target indexing from `b` on the other side of the addition operand. This expression is validated provided that the index type pair `i` and `i'` are an instance of `Reindexable` (and that the underlying data types permit addition).

A.3 Modularity (Database specific information through topic modules)

Foghorn uses a plug-in module system to target different branches of econometric literature. In *Foghorn*, modules must define the data sets available for a literature and the conversions among variable indices.[26] This approach ensures that a declarative econometric language is not tied down to any specific econometric methodology. For example, one module may link firm financial data to stock return data based on what data was publicly known to the market (i.e., the firm's financial data relevant to a specific market observation is from the most recent data published by the company). Another module may use contemporaneous data (i.e., the firm financial data is linked to any return data that occurs during the corresponding fiscal year). A third module may provide both options, requiring the econometrician to specify the linking method explicitly. A fourth module may not allow such index conversions to occur directly.

[26]Common data sets used in finance are already in *Foghorn*. Additional types of data sets, both within Finance and in other areas of the social sciences, will be added to the core distribution over time. However, users are not restricted to the core datasets and may implement their own.

A module consists of two pieces: a standardized database and its *Foghorn* code. The database is topic-specific, providing a standard against which empirical specifications may be executed. It contains the datasets commonly used in the research area and tables, or other rules, to link those datasets. A module's database serves as a foundation; it does not contain processed data, leaving calculations to the transpiled code generated by *Foghorn*.

The *Foghorn* code serves as the second piece of the module. The code provides information about the database, such as the data sources, variables, index information, and permissible index transformations. This information is used by the core language when analyzing specifications and transpiling. Modules can provide topic specific functionality. *Foghorn* provides a clean way to compose base-level functions into succinct module-level ones. Haskell is a functional programming language, meaning that functions, not objects, are the primary units of programs. Functions are composed simply by using a "dot" operator (i.e., the mathematical definition of composition $h = f \circ g$ is translated literally yielding `h = f . g`. Thus, it is simple to write module-level functions that address common problems in a specific economic literature by composing the building blocks provided by the core.

A.4 Transpiling

Foghorn translates top-down specifications into standard database code, a process known as transpiling. While new to econometrics, transpilation is an established technique in programming languages to add type-safety, enhance abstraction, and add concision to non-statically-typed languages. For example, there are many web-development languages for type-safe, high-level web development that transpile to Javascript, thereby preventing errors and making it easier to write modular code. *Foghorn* similarly employs transpilation, eliminating many types of mistakes in econometric coding.

Transpilation ensures that a declarative econometric language does not constitute a closed system. Having another tool for bottom-up data manipulation would not achieve our objectives for literature-level stress-testing and bi-directional transparency. There are numerous established methods for manipulating data and estimating econometric models. *Foghorn* uses a plug-in architecture for transpiling, allowing it to target various data manipulation languages. Users can pick the transpiler needed for their preferred target language. *Foghorn* currently has SQL and Stata transpilers, with a SAS transpiler under development.

Foghorn uses a simple approach to transpilation. *Foghorn* analyzes a specification and determines the basis set of variables underlying the required estimation panel. This basis set includes final variables used in estimation, intermediate variables needed for calculations, and source variables. It then categorizes the variables into two groups based on whether the variable has been

computed. Source variables are placed into the computed group; intermediate variables and final variables are placed into the uncomputed group. For each uncomputed variable, it checks if all its basis variables are in the computed set. If so, that variable is calculated and placed into the computed group. On iteration, all variables are eventually moved from the uncomputed group to the computed group. In summary, *foghorn* analyzes the specification, determines the order in which to perform algebraic, aggregating, reindexing, time-shifting, and other calculations, and generates all the necessary database code for the researcher. In other words, transpilation converts functions in *Foghorn* into their data manipulation equivalents.

Transpilation provides several benefits. The most clear advantage of transpilation is leveraging existing tools. Thus, a declarative approach can incorporate any data process or econometric estimator from any software package. One only needs to define a declarative syntax for a new process or estimator and then write a corresponding transpiler. Declarative econometrics is unique in that it defines the complete estimator at the specification level. By creating a direct link between specification and implementation, and then allowing researchers to manipulate specifications, new types of economic research become possible.

Transpilation greatly improves econometric transparency. Data manipulation languages are generally verbose, requiring hundreds if not thousands of lines of code to implement a standard economic paper. A person reviewing such code may suffer from information overload, failing to identify critical assumptions. A declarative specification, on the other hand, is concise and emphasizes the logic of any empirical study. Key assumptions form the basis of the declarative code, with implementation details provided in the transpiled, step-by-step data manipulation code. For example, *Foghorn* has an SQL transpiler that exports a specification into the step-by-step data generating-process.

A.5 Whole-specification coding (Benefits and disadvantages)

Declarative econometrics promotes a whole-specification approach to econometrics. As the language can reason about a specification through type-level logic, the estimator definition is intertwined with the variable definitions. There are several benefits of an approach where an empirical researcher can focus on specification design instead of implementation. Declarative econometrics makes key assumptions explicit, improving empirical clarity and transparency.

Declarative econometrics is not intended to, nor should it, replace standard techniques. Datasets are idiosyncratic and intimate knowledge is required for convincing empirical research. Thus, there are significant advantages to researchers working directly with data from the bottom-up. However, declarative econometrics makes new types of literature-level analysis possible, improving our critical understanding of standard practices through stress testing and replication.

Finally, we note that declarative econometrics has some disadvantages. *Foghorn* requires modules for specific topics. Creating a module requires building a standard database for the topic and encoding the logic into a *Foghorn* library. Thus, declarative econometrics is best suited to established topics where the upfront cost of writing a module can be amortized across paper replication, stress testing, and new research questions. For a unique empirical question, standard bottom-up techniques are better suited.

Moreover, declarative econometrics is not intrinsically designed to be efficient. *Foghorn* applies sample selection as a last step. As a result, the transpiled code will perform computations over an entire source dataset as it implements a study. It will compute variables that do not ultimately meet sample selection criteria. This inefficiency is by design. Computing variables over an entire dataset guarantees that the transpiled code will execute without crashing and generate the correct final sample for estimation.

Foghorn is just one possible implementation of a declarative econometric language. Other notation systems and implementations are possible. However, we believe the features highlighted above are critical to enabling new types of research, particularly in the areas of stress testing. Further details about the notation system and implementation are in the technical companion paper (Tumarkin 2026).

B Glossary

This glossary provides definitions of programming and econometric terms as used throughout the paper.

B.1 Programming Terms

Declarative language A programming language in which the programmer specifies what the program should accomplish, rather than detailing how to perform each step. Programs written in declarative languages emphasize logic and relationships, often leading to more concise code, leaving the system to work out the control flow and implementation details needed to accomplish the task. In *foghorn*, researchers write declarative econometric specifications, with the transpiler determining the steps necessary for implementation.

Embedded Domain-Specific Language (EDSL) A specialized programming language built within a general-purpose language, designed to express solutions in a specific domain (e.g., econometrics) more naturally and concisely. EDSLs leverage the host language's syntax and features while providing domain-specific tools and constructs. Developers commonly use EDSLs to express logic specific to a domain, benefiting from enhanced type safety and improved tooling support. Haskell is well-suited for EDSLs because of its strong abstractions and type safety. *foghorn* is an EDSL of Haskell, providing concise, type-safe econometric logic.

Function composition The process of combining two or more functions to create a new function, where the output of one serves as the input of the next. Function composition allows programmers to build complex operations from simple, reusable parts in a concise, declarative manner. In *foghorn*, the composition of econometric functions improves modularity, readability, and code reuse.

Higher-order function A function that takes other functions as arguments, returns a function as its result, or both. Higher-order functions are central to expressing abstract computation succinctly and avoiding repetitive code. *foghorn* employs higher-order functions to improve modularity and reduce boilerplate.

Imperative language A programming language in which programs are written as sequences of explicit instructions that specify how a computer should perform tasks. These languages focus on describing the control flow and individual steps needed to manipulate program state, often using variables, loops, and conditional statements. Imperative programming contrasts with declarative styles, where the focus is on desired outcomes rather than step-by-step procedures. Existing econometric languages are either imperative or used in an imperative style (e.g., SAS, SQL).

Module A self-contained unit of code that groups related functions, types, and definitions under a common namespace. Modules promote code organization, reuse, and abstraction by allowing programmers to separate and isolate components within a program. *foghorn* uses modules for its core language and for working in specific areas of empirical research (e.g., corporate finance).

Static typing A system in which the types of variables and functions are evaluated at compile time rather than at runtime. Static typing allows many errors — such as type mismatches or invalid operations — to be caught early, improving reliability and program safety. *foghorn* variables carry type information, including the cross-sectional and/or time-series indexing dimensions, allowing the language to validate that operations make sense automatically. *foghorn* uses static typing alongside type-level programming and type inference to encode and enforce econometric logic in specifications.

Term-level The "usual" layer of programming and code execution where values are defined, manipulated, and passed to functions. Unlike type-level programming, term-level programming focuses on concrete computations and data transformations. Most everyday programming tasks — such as arithmetic, data processing, and control flow — occur at the term level. In *foghorn*, the term-level defines the actual operations.

Transpiling The process of converting source code written in one programming language into source code of another language, while preserving program behavior. Unlike compiling, which often targets low-level machine code, transpiling typically produces human-readable code in another high-level language. *foghorn* allows econometricians to write specifications with enhanced features and stronger guarantees — such as improved type safety — while producing output in established data and econometric software.

Type A type is an abstract classification that specifies the kind of values a variable, expression, or function can take and the operations that may be performed on them. Types provide a framework for reasoning about program behavior, ensuring consistency and preventing invalid operations. *foghorn* variables carry type information on their index variables (i.e., the cross-sectional and/or time-series dimensions) and information type (e.g., numeric, SIC code, text). This, combined with static typing, type inference, and type-level programming lets the language reason about econometric specifications.

Type checking Type checking is the process by which a compiler or interpreter verifies that program constructs are used consistently with their declared or inferred types. It ensures, for example, that operations are applied to compatible data types and that functions receive valid arguments. *foghorn* uses type checking to ensure that econometric specifications, and the variables within them, are consistent.

Type inference Type inference is the process by which a compiler automatically determines the types of expressions without requiring explicit type annotations from the programmer. This enables concise code while retaining the benefits of static typing, since the compiler can still catch type errors at compile time. *foghorn* leverages Haskell's type inference to achieve strong econometric type safety with reduced syntactic overhead, as an econometrician does not need to label a variable's index or informational types.

Type signature A formal declaration of the inputs and output of a function, along with any constraints. *foghorn* uses type signatures to define econometric logic.

Type-level A layer above the term-level where logic about types is encoded. In strongly typed functional languages like Haskell, type-level programming allows developers to encode invariants, perform compile-time checks, and enforce constraints through types themselves. Type-level programming is the practice of writing programs that use a language's type system to encode logic, perform computations, and write expressions and abstractions that occur within a language's type system rather than at the term-level. This approach can enforce complex invariants, guarantee program properties at compile time, and eliminate certain classes of runtime errors. In *foghorn*, type-level programming defines econometric operations, and specifications.

B.2 Econometric & Research Terms

Bi-directional transparency Transparency in both directions: top-down (explicit specifications and assumptions) and bottom-up (generated data-manipulation code). Ensures hidden assumptions are revealed.

Empirical specification A formal, high-level description of the econometric model being estimated, consisting of: the dependent variable, explanatory variables, the functional form of the relationship, and the chosen estimator. In this paper, the specification is written at a high level and does not include step-by-step data manipulation.

Empirical procedure The complete process of estimating a model from raw data. It consists of the complete data manipulation process that transforms raw data into the final dataset and the econometric estimation procedure.

Indexing variables Identifiers that uniquely label dataset observations, which may be cross-sectional or time-series.

Reindexing Reindexing is the process of aligning data across datasets with different indexing systems.

Replication Re-implementing a paper's empirical specification to see if results can be reproduced with comparable magnitude, significance, and sample size. In this paper, replication is the process of independently coding an empirical specification as described within a paper. The replication is conducted from raw data, limiting author-provided information to proprietary data. Replication in this context is judged successful when: (i) coefficient estimates and significance are similar, (ii) sample sizes align, and (iii) control variables behave similarly.

Reproduction Obtaining exact results, or results with minor variations, to a published paper's empirical specification using data and code provided by the authors.

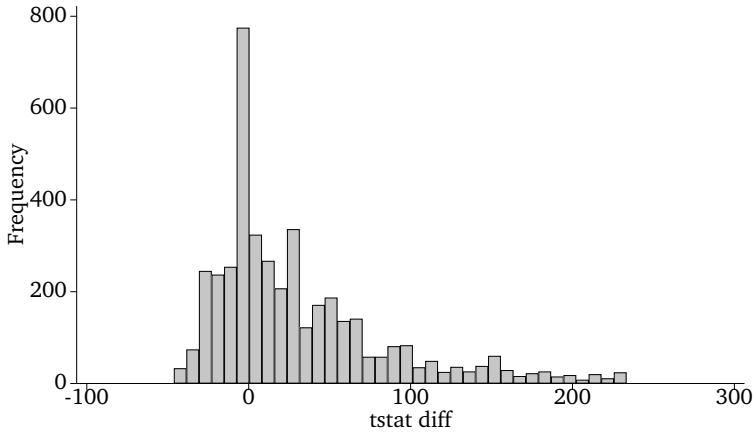
Robustness test A paper-level sensitivity check in which the authors vary aspects of their own methodology to show that results are not overly sensitive to small changes. These tests are typically idiosyncratic, tied to one paper's assumptions

Stress test A literature-level test that systematically varies common assumptions or practices across multiple papers.

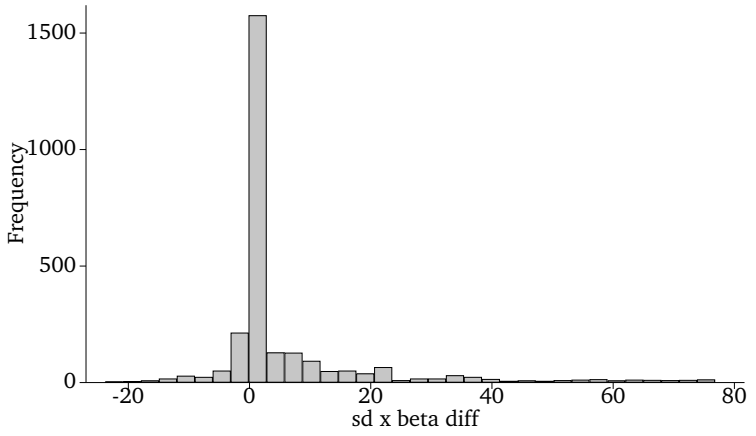
Figure IA1

CAR Stress Tests Histograms: Variation in Coefficient Estimates

This histogram plots the changes in the magnitude of the coefficient on the variable of interest by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. In Panel A, *Beta Magnitude % Change* is defined as the difference between the replicated coefficient estimate and the stress test coefficient estimate divided by the replicated coefficient if the replicated coefficient is > 0 and the stress test coefficient minus the replicated coefficient over the absolute value of the replicated coefficient estimate if the replicated coefficient is < 0. In Panel B, *Beta Magnitude Diff × SD* is defined as the difference between the replicated coefficient estimate and the stress test coefficient estimate when the replicated coefficient is > 0 and the stress test coefficient minus the replicated coefficient when the replicated coefficient is < 0. These differences are then multiplied by the within-replicated-sample standard deviation of the variable of interest (or one, if the variable of interest is an indicator variable). Stress tests allow for within-literature variation on (1) the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ([−205, −6], [−210, −11], [−245, −45], [−252, −20], [−272, −20], [−300, −91], and [−370, −253]). *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Both panels plot the 5th through 95th percentiles of all stress tests.



Panel A: Beta Magnitude % Change



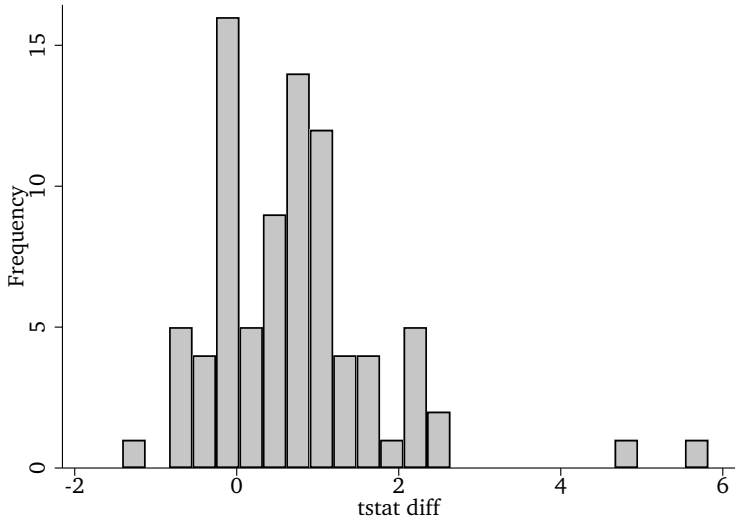
Panel B: Beta Magnitude Diff × SD

,

Figure IA2

CAR Stress Test Histograms: CAR Window Variation Only

This histogram plots *tstat_diff* for stress tests on the significance of the coefficient on the variable of interest and control variables by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Unlike the baseline analysis reported in Figure 1, the stress tests only allow for within-literature variation on the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]). The abnormal return calculation and the expected return estimation period are held fixed. *tstat_diff* defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). The figure plots the 1st through 99th percentiles of all stress tests for variables of interest.



,

Table IA1

Replication of Bates and Lemmon (2003), Table 8, Column (1).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and target and bidder termination fees. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Bates and Lemmon (2003). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (1)	
	Published	Replicated
Target termination fee indicator	0.010 (0.960)	0.014* (1.703)
Bidder termination fee indicator	-0.030** **(-2.340)**	-0.022** **(-2.009)**
Deal includes a lockup of target shares	0.002 (0.820)	0.002 (0.286)
Deal status (1=completed 0=withdrawn)	0.039*** (3.440)	0.018** (1.978)
Prior bidding indicator	-0.063*** (-6.120)	-0.057*** (-5.959)
Stock offer	-0.025*** (-2.540)	-0.034*** (-4.145)
Tender offer	0.091*** (7.530)	0.080*** (7.366)
Bidder toehold	-0.063*** (-4.320)	-0.001** (-2.434)
Deal attitude (1=hostile 0=friendly or unsolicited)	0.040** (2.260)	0.032** (2.206)
Log marketvalue of equity	-0.012*** (-4.600)	<0.001*** (-3.521)
Number of observations Adjusted-R^2	3037 0.067	3203 0.063

Table IA2

Replication of Becher, Griffin, and Nini (2021), Table 6, Column (1).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and financial covenant violations. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Becher, Griffin, and Nini (2021). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. Standard errors are reported in cornered parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (1)	
	Published	Replicated
Financial covenant violation	1.860***	1.093**
	〔 **0.687** 〕	〔 **0.537** 〕
Size	-0.057***	-0.627***
	〔 0.007 〕	〔 0.059 〕
Stock price runup	-0.041	-0.003
	〔 0.241 〕	〔 0.002 〕
Market-to-book ratio	-0.277**	-0.278**
	〔 0.108 〕	〔 0.118 〕
Operating cash flow / assets	-0.711	9.852**
	〔 0.933 〕	〔 3.952 〕
Leverage ratio	0.870	0.847
	〔 0.556 〕	〔 0.574 〕
Observations	7191	7299
Adjusted R-squared	0.018	0.030

Table IA3

Replication of Burch (2001), Table 6, Column (6).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and target lockup options. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 4-day trading window beginning 1 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Burch (2001). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *p*-values are reported in brackets. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (6)	
	Published	Replicated
Lockup (0/1)	-0.012* **[0.093]**	-0.017*** **[0.002]**
Toehold	0.001** [0.039]	<0.001 [0.994]
Completed	<0.001 [0.945]	0.002 [0.820]
Hostile	0.001 [0.880]	0.015 [0.209]
Free cash flow	-0.002 [0.910]	-0.030 [0.285]
Instown	-0.023 [0.106]	-0.012 [0.370]
Litigation	-0.006 [0.427]	-0.002 [0.804]
Market-to-book	-0.007*** [0.002]	-0.009*** [<0.001]
Size	-0.002 [0.289]	-0.003* [0.086]
Leverage	0.003 [0.833]	0.023* [0.060]
Observations	744	776
Adjusted R-squared	0.018	0.043

Table IA4

Replication of Fuller, Netter, and Stegemoller (2002), Table VII, Column "Private".

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and bidding firm acquisition frequency. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Fuller, Netter, and Stegemoller (2002). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *p*-values are reported in brackets. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column "Private"	
	Published	Replicated
Dummy = 1 if target is acquired with common stock	0.043*** [0.007]	0.115** [0.025]
Dummy = 1 if target is acquired with combo	0.009 [0.460]	0.032 [0.580]
Dummy = 1 if first bid	-0.003 [0.685]	-0.010 [0.305]
Dummy = 1 if fifth or higher bid	-0.019*** **[<0.001]**	-0.018*** **[0.005]**
Dummy = 1 if target is foreign	-0.012* [0.062]	-0.006 [0.474]
Dummy = 1 if bidder or target is a tech firm	-0.004 [0.431]	-0.002 [0.722]
Dummy = 1 if target and bidder are in same industry	0.004 [0.358]	-0.013** [0.035]
Log of relative size	0.007*** [0.010]	0.001 [0.662]
Log of target size	0.001 [0.442]	0.002 [0.390]
Interaction variable = relative size × stock	0.011** [0.012]	0.011** [0.020]
Interaction variable = relative size × combo	0.003 [0.513]	0.002 [0.739]
N	2060	1313
F-statistic	5.140	6.310
Adjusted R2	0.035	0.042

Table IA5

Replication of Golubov, Yawson, and Zhang (2015), Table 1, Column "Full Sample".

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and acquirer size. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Golubov, Yawson, and Zhang (2015). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column "Full Sample"	
	Published	Replicated
Intercept	0.032***	0.051***
	(2.941)	(3.226)
Ln (acquirer size)	-0.004***	-0.003***
	(5.496)	**(-2.856)**
Tobin's Q	-0.002***	-0.001*
	(2.966)	(-1.912)
Run-up	-0.013***	-0.005***
	(4.512)	(-3.484)
Free cash flow	-0.012	-0.014*
	(1.331)	(-1.711)
Leverage	0.017**	0.001
	(2.523)	(0.087)
Sigma	0.350**	0.075
	(2.306)	(0.694)
Relative size	0.002	29.263**
	(1.549)	(2.191)
Relatedness	<-0.001	0.001
	(0.160)	(0.683)
Tender offer	0.002	-0.001
	(0.392)	(-0.131)
Hostile	0.007	<0.001
	(0.592)	(-0.010)
Public × All-cash	-0.003	-0.006
	(0.755)	(-1.490)
Public × Stock	-0.032***	-0.042***
	(12.268)	(-9.530)
Private × All-cash	-0.004	-0.002
	(1.520)	(-0.933)
Private × Stock	-0.001	-0.002
	(0.259)	(-0.822)
Subsidiary × All-cash	0.007***	0.005**
	(2.606)	(2.066)
N	12491	14863
R2	0.057	0.082
Adj. R2	0.055	0.080

Table IA6

Replication of Gorton, Kahl, and Rosen (2009), Table V, Column "Harford Waves (4)".

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and the distribution of firm sizes in the acquired firm's industry. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Gorton, Kahl, and Rosen (2009). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *p*-values are reported in brackets. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column "Harford Waves (4)"	
	Published	Replicated
Log 123–456 size ratio	0.029*** **[0.007]**	0.041** **[0.023]**
Cash	0.001 [0.881]	0.004 [0.520]
Ratio	0.003 [0.694]	0.007 [0.497]
Log value	-0.006*** [<0.001]	-0.004** [0.029]
Tar priv	0.015*** [0.001]	0.024*** [<0.001]
Tar sub	0.022*** [<0.001]	0.020*** [0.010]
Cross industry	0.007 [0.104]	0.009 [0.405]
Competing bid	0.006 [0.720]	-0.036** [0.014]
Tender offer	0.017* [0.056]	0.019* [0.086]
Stock market return	-0.001 [0.861]	0.003 [0.558]
Market/book	0.002* [0.080]	0.001 [0.114]
Industry Herfindahl	0.036 [0.720]	0.109** [0.032]
Observations R2	1334 0.113	1141 0.075

Table IA7

Replication of Harford, Humphery-Jenner, and Powell (2012), Table 5, Column (1).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and managerial entrenchment. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Harford, Humphery-Jenner, and Powell (2012). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. Standard errors are reported in cornered parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (1)	
	Published	Replicated
Dictator dummy	-0.524**	-0.594**
	〔**-0.210**〕	〔**0.243**〕
Subsidiary	2.059***	2.749***
	〔-0.327〕	〔0.385〕
Private	1.495***	2.325***
	〔-0.278〕	〔0.386〕
All cash	0.313	0.286
	〔-0.310〕	〔0.254〕
All stock	-0.813**	-0.498
	〔-0.345〕	〔0.426〕
Log firm age	-0.016	-0.072
	〔-0.152〕	〔0.167〕
Stock run-up	0.977***	-2.037***
	〔-0.340〕	〔0.351〕
PRIV	-0.085***	0.051
	〔-0.019〕	〔0.037〕
Log market value	-0.318***	<0.001
	〔-0.083〕	〔<0.001〕
Tobin's q	0.292**	0.069
	〔-0.117〕	〔0.164〕
Free cash flow	6.625*	-1.796
	〔-3.892〕	〔2.206〕
Leverage	3.187***	-1.182
	〔-1.041〕	〔0.793〕
Industry M&A	-0.156	-2.785**
	〔-6.130〕	〔1.375〕
Relative size	0.146	1.707***
	〔-0.793〕	〔0.614〕

Continued

Table IA7
Continued

	Column (1)	
	Published	Replicated
Tech	0.314	-0.294
	[-0.248]	[0.340]
Conglomerate	0.038	0.422*
	[-0.227]	[0.238]
Competed	-0.948	-1.073
	[-0.701]	[0.653]
Volume	0.124	-0.332*
	[-0.098]	[0.182]
Cross-border	2.964**	-0.186
	[-1.160]	[0.294]
Friendly	-2.786***	-2.640***
	[-0.870]	[0.872]
Serial_3	0.092	-0.154
	[-0.274]	[0.312]
Number of observations	3934	3258
F-statistic	8.310***	8.804
Adjusted R2	0.073	0.054

Table IA8

Replication of John, Knyazeva, and Knyazeva (2015), Table V, Column (3).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and state-level labor rights. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in John, Knyazeva, and Knyazeva (2015). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (3)	
Acquirer—weak labor rights	0.494***	0.267**
	(5.440)	**(2.122)**
Acquirer size	-0.262***	-0.325***
	(-5.250)	(-6.621)
Relative deal size	-0.882***	2.024***
	(-3.530)	(10.551)
Diversifying acquisition	-0.081	-0.447**
	(-0.930)	(-2.115)
Tech indicator	0.276	-0.939***
	(0.590)	(-3.617)
Number of observations	13838	13846
R2	0.010	0.034

Table IA9

Replication of Li, Qiu, and Shen (2018), Table 2, Column 1.

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and organizational capital. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Li, Qiu, and Shen (2018). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. Standard errors are reported in cornered parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column 1	
OC	0.250***	0.306**
	〔**0.084**〕	〔**0.125**〕
ROA	-0.007	-0.004
	〔0.008〕	〔0.007〕
M/B	-0.032	-0.083**
	〔0.021〕	〔0.041〕
LEVERAGE	2.921***	3.452***
	〔0.396〕	〔1.061〕
PAST_RETURN	0.002*	0.008***
	〔0.001〕	〔0.002〕
TOP5_INSTITUTIONS	-1.905***	-1.131
	〔0.563〕	〔1.128〕
FIRM_SIZE	-0.405***	-0.513***
	〔0.040〕	〔0.082〕
ALL_CASH	0.536***	0.371**
	〔0.127〕	〔0.157〕
ALL_STOCK	0.023	-0.015
	〔0.196〕	〔0.313〕
DIVERSIFYING	-0.037	-0.069
	〔0.132〕	〔0.179〕
TENDER_OFFER	1.138***	1.272***
	〔0.311〕	〔0.357〕
RELATIVE_SIZE	0.950***	0.175
	〔0.168〕	〔0.127〕
PRIVATE_TARGET	2.270***	2.026***
	〔0.186〕	〔0.251〕
SUBSIDIARY_TARGET	2.758***	2.815***
	〔0.198〕	〔0.279〕
No. of obs.	17910	17609
Adj. R2	0.053	0.035

Table IA10

Replication of Ma, Whidbee, and Zhang (2019), Table 3, Column "All".

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and acquirer stock price relative to its 52-week high. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 7-day trading window beginning 5 trading days prior to and ending 1 trading days following the merger announcement. Reference price ratio (RPR) is the ratio of the acquirer's preannouncement stock price to its 52-week high price. All variables are as defined in Ma, Whidbee, and Zhang (2019). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column "All"	
	Published	Replicated
RPR	-5.487*** **(-8.560)**	-8.878*** **(-12.977)**
Ln(M/B)	-0.285** (-2.140)	-1.169*** (-8.175)
Stock	-1.208*** (-4.060)	-1.459*** (-4.768)
Cash	0.145 (0.910)	0.453*** (2.833)
Private	2.321*** (10.600)	1.322*** (6.228)
Stock × Private	2.125*** (5.040)	2.299*** (5.178)
Rel. size	2.216*** (8.800)	1.803*** (6.574)
Size	-0.398*** (-7.690)	-0.171*** (-3.107)
Leverage	-0.081 (-0.210)	0.424 (0.964)
Dormant > 1 yr	0.628*** (2.880)	0.695** (2.349)
Same industry	0.107 (0.670)	0.237 (1.437)
Tender offer	1.058** (2.220)	0.139 (0.380)
Hostile	-0.759* (-1.890)	-0.826 (-1.597)
Toehold	0.047 (0.100)	-0.565** (-2.431)
Cross border	-0.187 (-0.770)	-0.028 (-0.113)
Past return	0.777*** (4.700)	2.943*** (15.511)
N Adj. R2	19119 0.064	16014 0.083

Table IA11

Replication of Masulis, Wang, and Xie (2007), Table VI, Column (1).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and corporate governance. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Masulis, Wang, and Xie (2007). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (1)	
	Published	Replicated
Antitakeover Provisions:		
GIM index	-0.107** **(-2.490)**	-0.083** **(-2.072)**
Bidder Characteristics:		
Log(total assets)	-0.301*** (-3.590)	-0.328*** (-3.634)
Tobin's q	-0.085 (-0.680)	-0.033 (-0.227)
Free cash flow	1.902 (0.860)	-0.755 (-0.335)
Leverage	0.678 (0.640)	-0.284 (-0.239)
Stock price runup	-0.906** (-2.540)	-1.526*** (-4.156)
Deal Characteristics:		
Industry M&A	-1.096 (-0.770)	-1.637 (-1.110)
Relative deal size	0.209 (0.360)	1.486** (2.253)
High tech	0.420 (0.920)	-0.197 (-0.506)
High tech × relative deal size	-6.078*** (-3.150)	-0.824 (-0.355)
Diversifying acquisition	-0.269 (-0.880)	0.165 (0.685)
Public target × stock deal	-3.902*** (-7.290)	-2.758*** (-7.314)
Public target × all-cash deal	-2.082*** (-3.340)	-0.942* (-1.891)
Private target × all-cash deal	-1.969*** (-3.530)	-0.685 (-1.502)
Private target × stock deal	-1.689*** (-3.100)	-0.408 (-1.280)
Subsidiary target × all-cash deal	-1.472*** (-2.900)	-0.069 (-0.190)
Number of obs. Adjusted-R^2	3333 0.062	3380 0.055

Table IA12

Replication of Moeller, Schlingemann, and Stulz (2004), Table 5, Column (1)

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and firm size. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Moeller, Schlingemann, and Stulz (2004). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *p*-values are reported in brackets. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (1)	
	Published	Replicated
Intercept	0.015*** [<0.001]	0.004 [0.196]
Private	-0.004* [0.085]	-0.007*** [0.004]
Public	-0.032*** [<0.001]	-0.028*** [<0.001]
Small	0.016*** **[<0.001]**	0.015*** **[<0.001]**
Conglomerate	-0.004* [0.051]	0.001 [0.736]
Tender offer	0.015*** [0.001]	0.014*** [0.003]
Hostile	-0.012 [0.195]	0.019* [0.065]
Competed	-0.007 [0.299]	-0.010 [0.234]
All equity	-0.003 [0.341]	0.001 [0.829]
All cash	-0.004** [0.047]	0.002 [0.315]
Relative size	0.012*** [0.001]	<0.001 [0.102]
Tobin's q	-0.001* [0.064]	<0.001 [0.573]
Debt/assets(mkt.)	0.001 [0.876]	-0.012* [0.057]
Liquidity index	-0.009*** [0.003]	-0.001 [0.121]
CF/assets (mkt.)	0.001 [0.811]	0.031*** [0.002]
n Adjusted-R2	9712 0.052	10796 0.018

Table IA13

Replication of Nguyen and Phan (2017), Table 7, Column (3).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and government policy uncertainty. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 3-day trading window beginning 1 trading days prior to and ending 1 trading days following the merger announcement. All variables are as defined in Nguyen and Phan (2017). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (3)	
	Published	Replicated
PU_ANNOUNCEMENT	0.007** **(2.040)**	0.010** **(2.265)**
SIZE	-0.007*** (8.400)	-0.005** (-2.451)
MARKET-TO-BOOK_RATIO	-0.003** (2.570)	-0.001 (-1.108)
PAST_12_MONTH_RETURNS	-0.003 (1.060)	0.005*** (2.863)
AVERAGE_SALES_GROWTH	-0.006 (1.040)	<0.001*** (5.078)
BOOK_LEVERAGE	0.022*** (3.940)	0.005 (0.678)
NONCASH_WORKING_CAPITAL	0.016 (1.600)	-0.005 (-0.403)
FIRM_AGE	0.006** (2.720)	<0.001** (2.227)
EXCESS_CASH	0.001*** (4.930)	<0.001 (-0.437)
DEAL_RATIO	0.001 (1.440)	0.024* (1.829)
STOCK_DUMMY	-0.002 (0.660)	-0.006* (-1.744)
CASH_DUMMY	0.008*** (4.360)	0.013*** (5.105)
HIGH_TECH_DUMMY	-0.007** (2.210)	-0.004 (-1.484)
DIVERSIFYING_DUMMY	-0.001 (0.440)	-0.003 (-1.200)

Continued

Table IA13
Continued

	Column (3)	
	Published	Replicated
HOSTILE_DUMMY	-0.020 (1.390)	-0.021 (-1.366)
PUBLIC_TARGET_DUMMY	-0.017*** (4.720)	-0.027*** (-6.401)
CHALLENGE_DUMMY	0.015** (2.300)	-0.009 (-1.345)
TARGET_INDUSTRY_M&A_INTENSITY	0.001 (0.570)	<0.001 (-1.421)
Intercept	0.012 (0.740)	-0.005 (-0.169)
No. of obs. Adj. R2	6376 0.030	6674 0.113

Table IA14

Replication of Roosenboom, Schlingemann, and Vasconcelos (2013), Table 2, Column (3).

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and acquiring firm stock liquidity. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 5-day trading window beginning 2 trading days prior to and ending 2 trading days following the merger announcement. All variables are as defined in Roosenboom, Schlingemann, and Vasconcelos (2013). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *p*-values are reported in brackets. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column (3)	
	Published	Replicated
Stock Liquidity	-0.038*** **[0.006]**	-0.006*** **[<0.001]**
Payment = Equity	-0.008*** [0.005]	-0.016*** [<0.001]
Target = Public	-0.024*** [<0.001]	-0.026*** [<0.001]
Target = Subsidiary	0.010*** [0.001]	0.006* [0.061]
Relative Size	-0.002 [0.800]	0.005 [0.418]
Total Assets	<0.001 [0.996]	0.001 [0.560]
Leverage	<0.001 [0.997]	0.022* [0.067]
Market to Book	0.001 [0.514]	-0.015 [0.108]
Cash Flow	0.018 [0.401]	-0.003 [0.840]
Observations	3815	4189
Adjusted-R^2	0.057	0.074

Table IA15

Replication of Wang and Xie (2009), Table 5, Column "TCAR".

The table reports estimation results from linear regression models examining the relation between the market reaction to acquisition announcements and shareholder rights. The dependent variable is the acquiring firm's cumulative abnormal return, which is aggregated over the 11-day trading window beginning 5 trading days prior to and ending 5 trading days following the merger announcement. All variables are as defined in Wang and Xie (2009). Replication code written in the foghorn declarative econometric language and the transpiled SQL data manipulation code are available on Github. *t*-statistics are reported in parentheses. Coefficients displayed in bold represent the key economic variable of interest in the paper. Coefficients marked with ***, **, and * are significant at the 1%, 5% and 10% level, respectively.

	Column "TCAR"	
	Published	Replicated
Shareholder-rights difference (Target index - bidder index)	0.836*** **(3.480)**	0.594** **(2.170)**
Bidder Characteristics:		
Log(market cap)	2.084 (1.250)	0.582 (0.207)
Tobin's Q	-0.543 (-0.570)	-0.589 (-0.539)
Leverage	-10.512* (-1.710)	-13.799*** (-2.590)
Return on assets (ROA)	35.827*** (2.660)	38.425*** (2.866)
Target Characteristics:		
Log(market cap)	-3.985** (-2.480)	-6.677** (-2.194)
Tobin's Q	-0.114 (-0.100)	-0.624 (-0.649)
Leverage	-1.101 (-0.170)	4.273 (0.876)
Return on assets (ROA)	-12.927 (-1.200)	-9.049 (-0.854)
Deal Characteristics:		
Market cap ratio	-5.770 (-1.590)	-8.000*** (-2.623)
Tender offer	9.859** (2.550)	5.007* (1.833)
Diversifying acquisition	3.873 (1.500)	3.232 (1.579)
All cash deal	2.022 (0.440)	2.059 (0.782)
Merger of equals	-9.446*** (-3.990)	-0.385 (-0.100)
High-tech combination	0.042 (0.010)	-1.284 (-0.574)
Number of Obs.	396	378
Adjusted R2	0.216	0.190

Table IA16

CAR Stress Tests: Changes in Coefficient Estimates

This table reports summary statistics on stress tests on the changes in the magnitude of the coefficient on the variable of interest by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. *Beta Magnitude % Change* is defined as the difference between the replicated coefficient estimate and the stress test coefficient estimate divided by the replicated coefficient if the replicated coefficient is > 0 and the stress test coefficient minus the replicated coefficient divided by the absolute value of the replicated coefficient estimate if the replicated coefficient is < 0. *Beta Magnitude Diff* $\times$ *SD* is defined as the difference between the replicated coefficient estimate and the stress test coefficient estimate when the replicated coefficient is > 0 and the stress test coefficient minus the replicated coefficient when the replicated coefficient is < 0. These differences are then multiplied by the within-replicated-sample standard deviation of the variable of interest (or one, if the variable of interest is an indicator variable). Stress tests allow for within-literature variation on (1) the CAR announcement window ($[-1, +1]$, $[-2, +2]$, $[-3, +3]$, $[-5, +5]$, $[-1, +5]$, $[-5, +1]$, $[0, +5]$, and $[-5, 0]$), (2) abnormal return calculation (equally and value weighted single-index model (SIM), a Fama-French 3-factor (FF3) and 4-factor (FF4) model, and the equally weighted (OEW) and value weighted (OVW) market return). (3) the expected return estimation period for tests estimating the market return using the SIM, FF3, and FF4 ($[-205, -6]$, $[-210, -11]$, $[-245, -45]$, $[-252, -20]$, $[-272, -20]$, $[-300, -91]$, and $[-370, -253]$). Panel A reports summary statistics. Panel B reports distribution statistics.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
Beta Magnitude % Change	4545	42.341	225.838	−0.997	115.807
Beta Magnitude Diff × SD	4545	8.206	21.564	1.558	6.551

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
Beta Magnitude % Change	−4200	−472	−28.13	−21.05	−2.22	15.47	60.74	149.99	233.33	730	4000
Beta Magnitude Diff × SD	−68	−40	−12.80	−3.49	0.00	0.30	10.55	37.60	60.50	79	101

Table IA17

CAR Stress Tests: Variation in CAR Window Only

This table reports summary statistics on stress tests on the significance of the coefficient on the variable of interest by varying the definition of Cumulative Abnormal Returns (CARs) performed on 15 specifications from replicated papers. Unlike the baseline analysis found in Table 6, the stress tests only allow for within-literature variation on the CAR announcement window ([−1, +1], [−2, +2], [−3, +3], [−5, +5], [−1, +5], [−5, +1], [0, +5], and [−5, 0]). Abnormal return calculations and the expected return estimation period are held fixed. Panel A reports summary statistics for *tstat_diff*. *tstat_diff* is defined as the reported/replicated t-statistic less the stress test t-statistic (for reported/replicated t-statistics > 0) and the stress test t-statistic less the reported/replicated t-statistic (for reported/replicated t-statistics < 0). Panel B reports distribution statistics for *tstat_diff*. Panel C reports summary statistics on the proportion of stress tests resulting in lower significance (*tstat_diff* > 0), a reduction in the level of significance (i.e. moving from 1% significance to 5% significance), a loss of significance, and an increase in the level of significance.

	N	Mean	Std. Dev.	Skewness	Kurtosis
Panel A: Summary statistics					
CAR Window Variation Only	84	0.706	1.089	1.804	8.937

	Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Panel B: Distribution											
CAR Window Variation Only	−1.414	−1.414	−0.574	−0.323	−0.073	0.657	1.110	2.066	2.291	5.821	5.821

	Increase in tstat_diff	Drop in Significance	Loss of Significance	Increase in Significance
Panel C: Changes in statistical significance thresholds (%)				
CAR Window Variation Only	80.176	52.651	36.897	4.708