IZA DP No. 18892

August 2026

When Samples Shape Significance

Sonia Ale
University of Pittsburgh

Md Shafiqul Islam
University of Pittsburgh

Lester Lusher
University of Pittsburgh
and IZA@LISER

Huda Osman
University of Pittsburgh

Jacob Stenstrom
University of Pittsburgh

IZA DP No. 18892 August 2026

When Samples Shape Significance*

Abstract

Random sampling yields unbiased estimates, yet sampling variation alone can produce incorrect conclusions. Using a setting where repeated draws from the same data-generating process are observable, we replicate findings from 24 papers in top economics journals and re-estimate each result using 40 independent resamples. Across 3,640 resampled results, t-statistics shrink by 31%, frequently leading to lost significance. Proxies for researcher degrees of freedom and sampling noise predict reduced significance. Our exercise provides novel evidence quantifying the effect of sampling variability on empirical research and highlights how statistical evidence can vary even when holding research design fixed.

JEL classification

C12, C18, C80

Keywords

sampling variation, statistical significance, reproducibility, publication bias, Google Trends

Corresponding author

Lester Lusher
lesterlusher@pitt.edu

* We thank Uma Sisira Akella, Sayantan Chakraborty, and Yasser El-Sayed for providing excellent research assistance. We thank Scott Carrell, Claire Duquennois, Osea Giuntella, Erzo F.P. Luttmer, Richard Murphy, Seth Stephens-Davidowitz, and seminar participants from the University of Pittsburgh, Case Western Reserve University, Southern Methodist University, University of Texas at Austin, and Loyola University Chicago for helpful comments. We especially thank the authors of the original manuscripts we replicated for providing code/data and/or clarifying their data collection process. This project would not have been possible without their generous correspondence, and by and large, the vast majority of authors whom we contacted responded positively to our replication efforts. The pre-analysis plan, produced on August 15, 2025, can be found at https://osf.io/ymrg4.

"*Insanity is doing the same thing over and over again, but expecting different results.*"
- Unknown origin, though often misattributed to Albert Einstein.

1 Introduction

From its earliest foundations, probability theory has highlighted how empirical inference is fundamentally driven by randomness. In *Ars Conjectandi*, Bernoulli (1713) formalized what would later become known as the Law of Large Numbers: random samples reveal population parameters only in expectation, and any particular realization may deviate (substantially) from the truth. This insight underlies modern statistical inference. Empirical estimates are random variables with sampling distributions that reflect uncertainty.

Modern empirical economics is built squarely on this logic. Under random sampling or assignment, estimators are unbiased and converge to their true parameters as sample size grows. Statistical inference formalizes this uncertainty through hypothesis testing, a framework developed by Fisher (1925) and later unified by Neyman and Pearson (1933), who distinguished between Type I errors (incorrectly rejecting a true null hypothesis) and Type II errors (failing to reject a false null hypothesis). Even under ideal identification and large samples, sampling variation implies that both errors occur with some non-zero probability.

In principle, these errors are well understood and properly accounted for: if one could repeatedly draw samples from the same data-generating process, the resulting distribution of estimates would be centered on the true population parameter. In practice, however, researchers almost always observe only a single realization. This limitation is typically viewed as innocuous, as long as estimators are unbiased and standard errors are correctly computed. The implicit assumption is that the estimates we observe (publish) are representative draws from the underlying sampling distribution.

A growing literature questions this assumption. Early work documented systematic publication bias, whereby statistically significant results are more likely to be published than null findings (Sterling, 1959; Franco et al., 2014). More recent evidence focusing on economics further tackles issues related to publication biases and p-hacking (e.g., Brodeur et al., 2016, 2023).[1] Related work highlights how researcher degrees of freedom (choices over specifications, samples, and transformations) can generate statistically significant results even in the absence of true effects (Leamer, 1983; Simmons et al., 2011; Huntington-Klein et al., 2025).

[1] Other studies related to publication biases and p-hacking include Andrews and Kasy (2019), Ashenfelter et al. (1999), Bruns et al. (2019), De Long and Lang (1992), Doucouliagos and Stanley (2013), Ferraro and Shukla (2020), Havránek (2015), Ioannidis (2005), Ioannidis et al. (2017), Lybbert and Buccola (2021), McCloskey (1985), Miguel et al. (2014), Stanley (2005), and Stanley (2008).

Together, this literature emphasizes that published estimates may differ sharply from the "true" distribution of estimates implied by random sampling. Importantly, this divergence need not arise from misconduct or invalid research designs. Even honest, well-identified studies may yield misleading inferences if sampling variation interacts with selective visibility of results. When only one realization from a stochastic process is observed, and when that realization is more likely to be reported if it is statistically significant, sampling noise alone can shape empirical conclusions.

In this paper, we isolate and quantify this mechanism. We exploit a distinctive, yet not well known, feature of Google Trends data: reported search intensity indices are constructed from random samples of all Google searches, and identical queries executed on different days yield different realizations of the underlying data-generating process.[2] As a result, Google Trends provides a unique opportunity to observe multiple independent draws from the exact same sampling process while holding research design fixed.

As documented in our pre-analysis plan, we first identified all papers that published in a top general interest economics journal[3] between 2015 and August 2025 that utilized Google Trends data, then selected papers based on the availability of replication packages. In total, 73 publications utilized Google Trends data, and of these, we include 24 papers in our study.[4] Google Trends data are typically used to measure public attention, information acquisition, and actual behavior, with applications spanning political behavior, media effects, labor markets, health, education, and finance. For the publications included in our study, we first computationally replicate the study's original findings with the authors' code and data.[5] Then, for each paper, we re-download the underlying Google Trends series on 40 different days and re-run the original analyses 40 separate times. This procedure generates 3,640 replicated estimates that differ only due to sampling variation. If published estimates were representative draws from the sampling distribution, and if selective reporting played no role, then our resampled estimates would, in expectation, mirror the original published estimates.

Overall, we find that this is not the case. In our resamples, test statistics shrink substantially, statistical significance is frequently lost, and sign reversals occur with non-trivial probability. On average, t-statistics shrink by 31%, and 55% of resamples are statistically significant at the 5% level (compared to 85% of original results). 70% of our resamples produced a smaller t-statistic, significantly larger than the 50%

[2]To our knowledge, the only study to explicitly identify and discuss this comes from Cebrián and Domenech (2023), who report Google Trends queries across six different days.

[3]These journals include the *Quarterly Journal of Economics*, *Econometrica*, *American Economic Review*, *Journal of Political Economy*, *Review of Economic Studies*, *Review of Economics and Statistics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Applied Economics*, and *American Economic Review: Insights*.

[4]Besides many publications lacking replication packages, others lacked essential components needed to resample (e.g., information on the exact Google Trends query).

[5]The majority of studies included in our sample used Google Trends data as an outcome variable, with a handful additionally using Google Trends data to measure "treatment."

one would expect under sampling noise centered around zero. 9% of resampled estimates were of the opposite sign as the original estimate. The shrinkage in t-statistics is driven by both smaller effect sizes (12% reduction) and inflated standard errors (40% increase).

We further document significant heterogeneity across papers. Though the majority of papers experience smaller t-statistics on average in our resamples, just under half of papers maintain statistical significance at the 5% level across all 40 resamples. Thus, we find that nearly half of papers remain "robust" in the sense that the entirety of their resamples maintained statistical significance. On the other hand, eight papers experience substantially reduced statistical significance across resamples relative to what was originally reported, and not a single paper improved in the share of estimates that were statistically significant. A handful of papers exhibit sharp drops in t-statistics and in the share of resamples that were statistically significant. For example, four papers produced t-statistics which on average shrunk by over 50% across resampled estimates.

Lastly, we document how these patterns are systematically related to paper characteristics associated with researcher discretion and sample variation. For instance, we find that for each researcher degree of freedom "exercised,"[6] our model predicts a nearly 14% decrease in t-statistics and a 12 percentage point decrease in the probability that the resample was statistically significant at the 5% level. Measures of sampling noise (e.g. the average cross-resample correlation)[7] are strongly associated with smaller resampled test statistics i.e. the papers that were the least robust to our resampling exercise were also those that utilized the noisiest sampling distributions.

This paper makes several important contributions. First, we contribute to the literature on publication bias and selective reporting by isolating a mechanism that operates even in the absence of intentional p-hacking or specification searching. A large body of work shows that statistically significant results are more likely to be written up and published, and that researcher degrees of freedom can amplify false positives. In contrast to this literature, our analysis holds the research design, specification, and estimation procedure fixed and varies only the random sample underlying the data itself. We show that sampling variation alone, when coupled with just a single observed realization, can substantially inflate statistical significance in publications. This mechanism does not require researcher misconduct, ex post specification choice, or strategic behavior; it arises simply from relying on a single realization of an inherently stochastic data

[6] Described in further detail later, these measures include an indicator for the paper utilizing multiple Google Trends queries, transforming the Google Trends data (e.g. log), taking a subsample of the Google Trends data, and dropping Google Trends indices that were equal to zero.

[7] The average cross-resample correlation is calculated as the average across all pairwise correlations across our resamples within a paper. Any queries that are especially popular on Google Trends will be estimated with stronger accuracy, and thus have less variation across resamples, generating a stronger correlation cross-resample correlation. Queries that generate pure noise will produce correlations closer to 0 across resamples.

source.

Second, our findings speak to the growing literature on reproducibility and replication in economics, but from a distinct perspective. Existing replication efforts typically assess whether published results can be recovered using the same data and code, or whether conclusions are robust to alternative samples, specifications, or research designs (Camerer et al., 2016; Young, 2019, 2022; Brodeur et al., 2026). By contrast, we examine what happens when the same empirical design is re-estimated using multiple independent realizations of the same data-generating process. This distinction is important: failure to replicate is often interpreted as evidence of flawed research practices or fragile identification, whereas our results demonstrate that substantial variation in statistical evidence can emerge even when replication is exact and identification is sound. In this sense, our paper highlights a previously underexplored dimension of reproducibility: sensitivity to latent sampling variation in widely used data sources.

Third, we provide new evidence on the interpretation of statistical significance in applied economics. Standard hypothesis testing frameworks, following Fisher (1925) and Neyman and Pearson (1933), explicitly acknowledge the possibility of Type I and Type II errors arising from sampling variability. However, applied work often implicitly treats statistically significant estimates as informative realizations, rather than as noisy draws from a distribution that is rarely observed in full. Our results underscore the gap between this theoretical understanding and empirical practice. When only a single realization of a stochastic process is observed, the conventional interpretation of p-values and t-statistics becomes problematic. Sampling uncertainty, rather than identification failure, can be a first-order driver of overstated evidence.

More broadly, our paper highlights a tradeoff at the core of modern empirical research. As economists increasingly rely on large, complex, and algorithmically generated data sources, the randomness embedded in data construction itself becomes an important object of inference. Google Trends is a particularly transparent example, but the underlying issue is more general: many experimental, survey, administrative, and proprietary datasets rely on sampling, aggregation, or filtering procedures that are rarely observed or repeated. Our findings suggest that treating such data as fixed inputs may lead researchers to overstate the precision and robustness of empirical results.

2 Data and Resampling Design

We start by describing the primary data source used in the analysis (Google Trends), and explain why the platform offers a unique setting for studying sampling variation. We then describe how our sample of papers is constructed, following the project's pre-analysis plan, and discuss how Google Trends data are

used in economics. We then outline the resampling design and provide summary statistics.

2.1 Google Trends

Google Trends is a publicly available tool that reports the relative frequency of Google search queries over time and across geographic units. For a given query—defined by a set of keywords, a time period, a geography, and a download type (time series or cross section)—Google Trends produces an index of search intensity normalized to range from 0 to 100. Higher values indicate greater search volume relative to the maximum observed within the query's scope.

Crucially, Google Trends indices are not constructed from the universe of Google searches. Instead, Google relies on random samples of searches to estimate query-level search intensity, a design choice that allows the platform to generate results quickly and at scale. As a consequence, identical Google Trends queries executed on different days produce different realizations of the search intensity series. Each download therefore represents an independent draw from the same underlying data-generating process, rather than a deterministic retrieval of a fixed dataset or a population parameter.

2.2 Sample Construction

The sample of papers analyzed in this study is constructed according to a pre-analysis plan. We begin by identifying all papers published from 2015 through August 2025 in leading economics journals that make use of Google Trends data. The journal set includes *Quarterly Journal of Economics*, *Econometrica*, *American Economic Review*, *Journal of Political Economy*, *Review of Economic Studies*, *Review of Economics and Statistics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Applied Economics*, and *American Economic Review: Insights*. To identify relevant papers, we review all publications in these journals during the sample period, search for references to "Google", and manually verify whether Google Trends data are used in the empirical analysis. In total, we identified 73 papers, each of which is listed and described in Appendix Table A1.

From this universe of 73 papers, we identify a subset of "candidate" papers that provide replication packages containing the data files and code required to reproduce results that rely on Google Trends. Because replication packages vary in completeness, some candidate papers include only transformed Google Trends variables rather than the underlying queries. In these cases, we reconstruct the original Google Trends queries using the descriptions in the paper text, the available code, and, when necessary, correspondence with the original authors. Our final sample consists of 24 papers, yielding 91 core empirical estimates that directly use Google Trends data. The majority of papers that failed to reach candidacy were due to lacking

replication packages, while another handful were due to being unable to recover the original Google Trends queries. The papers included in our study are marked with asterisks in Appendix Table A1.

Finally, as specified in the pre-analysis plan, we aimed to include all candidate papers. However, because of uncertainty in the number of potential papers that use Google Trends data and the time required to replicate individual studies, we pre-specified that paper inclusion would continue either until all candidate papers were exhausted, or until the end of 2025. We also pre-specified that we would expand our sample of candidate papers should the peer review process request it, whether by including the candidate papers that we failed to replicate prior to 2026, or by expanding the universe of papers to include more journals and/or years.

2.3 Use of Google Trends in Economics

As documented in Appendix Table A1, Google Trends data are widely used in applied economics as proxy measures of public attention, information demand, or salience related to economic and social phenomena. Applications span a broad range of fields, including political economy, media effects, labor markets, health economics, education, and finance. Google Trends variables are used flexibly, typically as outcome variables, but also as treatment variables, instrumental variables, or controls, and they often play a central role in either providing "first stage" evidence or as a primary outcome variable in the paper. Of the 73 papers we initially screened, 22% used Google Trends data in their main analyses, 36% used it for descriptive analyses, and 42% used it for "secondary" analyses, including identifying mechanisms and conducting robustness checks.

Despite the stochastic nature of Google Trends indices, nearly all published studies rely on a single downloaded realization of the data.[8] Empirical analyses typically treat the retrieved series as fixed inputs, and statistical inference proceeds without accounting for the sampling variation embedded in the construction of the Google Trends measure itself. This practice is standard across studies that utilize statistics from samples of data (e.g. unemployment rates) and is rarely discussed explicitly, but it implies that uncertainty arising from the data construction process is not reflected in reported estimates or standard errors.

2.4 Resampling Design

For each paper included in the sample, we first replicate the original results using the authors' reported specifications and estimation procedures. We then identify the full set of Google Trends queries used to

[8] We came across only two studies that explicitly downloaded multiple samples for the same query, and in both cases, the authors took the average of the Google Trends indices across samples.

construct the original dataset. A query is defined by a combination of keywords (between one and five), a time period, a geographic unit, and a download type (e.g. time series vs. cross section).

For each query, we conduct 40 resamples by executing the identical Google Trends query on 40 different days. We then reproduce the original results 40 times, each time replacing the original Google Trends values with a resampled realization, while holding all other aspects of the analysis fixed. This includes applying the same data transformations, sample restrictions, and estimation choices as in the original paper. As a result, differences between the original estimates and the resampled estimates arise solely from sampling variation in the Google Trends data.[9]

2.5 Summary Statistics

Table 1 presents summary statistics for the 24 papers. Papers use anywhere between one to 776 unique Google Trends queries in constructing their data, with the majority (58%) using just one Google Trends query. 67% of papers use the raw Google Trends indices themselves in their analyses, while the remaining 33% implement at least one transformation to their Google Trends data (e.g. log transformation). Nearly the entirety of papers in our sample use Google Trends to generate an outcome variable, while others also use Google Trends to generate treatment variables; note that multiple papers generate multiple variables, treatment(s) and outcome, from their Google Trends queries. Approximately half of papers opted to utilize only a subset of the data generated from their Google Trends queries; oftentimes, this took the form of downloading a longer time series than was actually utilized in the analysis.[10] 12% of papers opted to drop observations where the Google Trends index was equal to zero.[11] 8% utilized daily data, 29% weekly, and 46% monthly; the remainder were either units greater than a month, or a cross-section download with a queried window of greater than a month. We also calculated the average pairwise correlation across our 40 resamples for each paper; the average of this measure was 0.86, with a range of 0.37 and $>$0.99. 25% of papers had at least one query that produced a "missing" file i.e. a file filled entirely with missing observations; this occurs whenever there is inadequate search volume for the given query. This indicator

[9] One could alternatively think about our exercise as estimating the role of measurement error in Google Trends data. In our context, expected measurement error would be equal to zero (due to random sampling), but "shocks" to measurement deriving from natural sampling variation could lead to incorrect conclusions. So, the core story is about sampling variation, but the way sampling variation affects results is through the incorrect measurement of one or more variables in a model.

[10] We can only speculate as to why so many authors utilized only a subset of their data. The most skeptical interpretation would be an *ex post* rationalization after seeing the results for the subsample. A more innocuous explanation is that authors first opted to download "all" the Google Trends data prior to merging to their main analysis files. Google Trends indices are also values relative to other observations within the queried file, and so by focusing a "larger" query (e.g. more dates), the generated indices are arguably more stable.

[11] In most cases, zeroes were dropped in the analysis files with no explanation provided in the paper. One potential justification is that values of zero may reflect inadequate search volume for that observation (i.e. the "denominator" of total Google searches for that observation is low). The challenge however is that zeroes also reflect zero *actual* search volume for that observation.

was generated based on either the paper reporting it (e.g. authors describe having an imbalanced panel due to inadequate search volumes for certain space-time units), or in discovery when we replicated the original paper's queries during the resampling procedure.

Moving to the bottom panel of Table 1, we report summary statistics based on the original paper's findings. Papers produced anywhere between one to 16 results using Google Trends data, with an average of 3.8. Note that we only focus on estimates based on the main "treatment" variable(s) described in the paper i.e. we do not focus on coefficients for constants or control variables. The average t-statistic within each paper (across estimates) is quite high (8.25), and ranges between 1.58 and 61.43. The average p-value is 0.03, and the share of estimates that are significant at the 5% level is 0.85. This illustrates how the vast majority of papers report findings using Google Trends data that are considered strongly statistically significant.

3 Methods

The unit of observation in the "raw" data is at the paper-estimate-resample level, where each observation pertains to a result from a single resampling for a specific paper's estimate. Therefore, the total sample size is 3640, the number of included paper-estimates times 40 (the number of resamples for each query). As specified in our pre-analysis plan, for each resample, we coded the resampled point estimate, standard error, t-statistic, and p-value. Using these four resampled variables, we construct three primary metrics:[12]

1. Relative t-value: The resampled t-statistic divided by the original t-statistic
2. Statistically significant at 5%: An indicator for whether the resampled p-value is below 0.05
3. Shrunk t-value: An indicator for whether the resampled t-statistic is less than the original t-statistic

From these three metrics, we can test three different hypotheses. First, we test whether the relative t-value is different from 1, as this would suggest that our resampled t-statistics are on average different from the original t-statistics. If the relative t-value is systematically less than 1, then this would suggest that our resamples produce significantly weaker evidence in support of the original paper's hypotheses. The second metric tests whether statistical significance at 5% occurs less frequently in our resamples than in the original estimates; to the extent that researchers care (more) about "statistical significance," this test allows us to see how robust the original estimates are in relation to the 5% threshold. Finally, our third hypothesis tests

[12] Please see Dreber and Johannesson (2025) for a thorough discussion of reproducibility and replicability in economics, including a discussion of the metrics adopted in our manuscript.

whether the probability the resampled t-statistic is smaller than the original is different from 0.5. If published estimates using the Google Trends data are truly unbiased, then our resampled draws should produce results that center on the original estimate. If instead results are biased toward statistical significance (and larger t-statistics), then we should see that the probability of the t-stat shrinking exceed 0.5.

We also construct the following secondary metrics:

1. Statistically significant at 10%: An indicator for whether the resampled p-value is below 0.10
2. Relative effect size: The resampled point estimate divided by the original point estimate
3. Relative standard error: The resampled standard error divided by the original standard error

The first metric mirrors the primary metric of statistical significance, but instead considers the 10% level. The relative effect size metric reflects whether the magnitudes of the point estimates changed in our resamples, and is perhaps the metric most closely related to "economic significance." If effect sizes shrink in our resamples, then this metric would be systematically smaller than one. The relative standard error metric reveals whether the standard errors in resampled results are systematically different from the original standard errors. If results from resamples are substantially noisier, then this metric would be greater than one. Note that these latter two constitute the numerator and denominator for the relative t-value metric, and thus explain why resampled t-statistics could differ from the original (different effect sizes and/or different standard errors).

We conduct analyses using both the "raw" data and data collapsed to the paper level. When collapsing the data, we simply calculate the average of the above metrics across observations (estimate-resamples) so that each estimate within a paper gets equal weight.[13]

4 Results

We start with Figure 1 by plotting the distribution of t-statistics separately for the original paper's findings (in red) and for our resamples (in blue). Original t-statistics are normalized to be positive, while resampled t-statistics are normalized to be positive if the result was in the same direction as the original study, and negative if in the opposite direction. We first note that the majority of original results are statistically significant at the 5% level (greater than 1.96). The sizes of t-statistics are rather large too: the median t-statistic is 3.48, and the 25th percentile is 2.36. Hence, the majority of original results are not only

[13] We pre-specified that we would additionally collapse to the paper-exhibit level, where exhibit is a table or figure. However, these results are virtually identical to results at the paper level, since the majority of Google Trends results are contained within a single exhibit. Hence we opt to not include these results in the manuscript.

statistically significant, but what many would consider to be "highly" statistically significant (p-values less than 0.001).

Our resampled t-statistics lead to a notable leftward shift in the distribution i.e. t-statistics shrink in response to our resampling exercise. The peak in the distribution of resampled t-statistics is well below 1.96, though still greater than zero. Results "flip" direction in nearly 9% of resamples i.e. produce a coefficient in the opposite direction as the original finding. Moreover, over half of these "flipped" findings still produce statistically significant t-statistics (i.e. t-statistics less than -1.96). In other words, a non-trivial share of our resamples produced statistically significant evidence in the opposite direction of the original paper's findings, while resamples in the same direction as the original findings still lead to substantially smaller t-statistics and lost statistical significance.

To characterize these results more formally, in Table 2 we report results from our six pre-registered metrics, and provide 95% confidence intervals on the values for these metrics. We first find that the average relative t-value (the resampled t-statistic divided by the original t-statistic) is 0.69 i.e. t-statistics shrunk on average by 31%. We are able to reject the hypothesis of a relative t-value equal to one (p-value$<$0.0001), suggesting that t-statistics did in fact shrink in our resamples. Resampled estimates were statistically significant in 55% of cases, substantially less than the 85% reported in original studies. And resampled t-statistics shrunk in 70% of cases, substantially more than the 50% one would expect in the presence of random sampling variation alone.

Moving to our secondary metrics, we find that statistical significance at the 10% level also shrunk substantially. The relative effect sizes were on average 0.88 i.e. coefficients shrunk by about 12% in our resamples, while the average relative standard error was 1.40 i.e. standard errors increased by a magnitude of 40%, on average.[14] Both of these metrics differ substantially from one. Thus, the shrinking in t-statistics can be attributed to both smaller magnitudes (i.e. reduced "economic significance") and larger standard errors (i.e. reduced confidence).

4.1 Heterogeneity Across Papers

We next present results for our three primary metrics for each paper. A metric for a single paper is calculated by averaging across estimate-resamples within each paper. We start with Figure 2, which plots the average relative t-value for each paper.[15] Immediately, one can observe substantial heterogeneity in

[14]For the relative effect size and relative standard error metrics, we removed one outlier paper from the sample: this paper had substantially more variation in relative effect sizes (between -10.25 and 42.32) and relative standard errors (between 0.12 and 1134.69). Dropping this paper from the other four metrics do not substantially change our results.

[15]We opt to not identify specific papers in our figures, beyond the journal they published in. The primary purpose of our study is to measure the overall role of sampling variation, and not to directly juxtapose resampled results paper by paper. Still, the list

resampled results by paper. Starting from the top, there are 8 papers that produced relative t-values greater than 1 i.e. for a significant share of papers, our resampling exercise increased the statistical precision of the results.[16] On the other end of the figure, there are 4 papers which produce resamples with substantially smaller t-statistics (relative t-values below 0.5), while another significant share of papers still produce relative t-values that are significantly less than one. Overall, 63% of papers produce relative t-values less than 1.

Next, in Figure 3, we plot the share of original estimates (in red) and resampled estimates (in blue) that were statistically significant at the 5% level. We first note that there are 11 papers where the original results were always statistically significant and 100% of our resamples produced statistically significant estimates. Thus, to the extent that one may only care about t-statistics crossing the 1.96 threshold, our resampling exercise has no influence on the conclusions one would draw for nearly half of the papers in our sample. For eight papers, the share of resamples that are statistically significant is substantially smaller than the share of original results that are statistically significant, highlighting how the main results in terms of lost statistical significance can be attributed to a minority share of papers. There are zero cases where our resamples produced greater statistical significance (at the 5% level).

Finally, in Figure 4, we plot the share of resamples that produce t-statistics that were strictly less than the original t-statistic. 14 papers produce shrunk t-statistics on average, while the remaining 10 produce the same or larger t-statistics. For several papers, resampled t-statistics are identical to the original t-statistics due to a variable being defined by a threshold relative to the Google Trends index (e.g. an indicator that the Google Trends index exceeded a value of 10, where all resamples produced indices greater than 10).

Still, comparing across the three paper-level figures, one can see a positive correspondence between smaller t-statistics, lost statistical significance, and increases in the probability of t-statistics shrinking: for example, "JPE #1" and "JPE #2" have average relative t-values below 0.5, achieve statistical significance in less than 40% of resamples, and have shrunken t-statistics in over 90% of resamples. Other examples are not so straightforward, however: for example, 100% of our resamples have smaller t-statistics in "AEJ: Policy #3", yet these resamples still retain statistical significance in 100% of cases, while the relative t-value is not as small as the handful of cases below 0.5 (though still significantly less than 1). In this latter case, if one cared more about statistical significance (versus the size of the t-statistic), then one may conclude our resampling exercise to support the original hypothesis of the paper.

of included studies can be found in Appendix Table A1 (marked with asterisks), and our replication packages include the relevant code and data from the original manuscripts, which could thus be used to link individual papers to results.

[16] One additional paper has a relative t-value of exactly 1: its estimate is defined by a threshold indicator that did not flip in any of the 40 resamples, so the resampled and original t-statistics are numerically identical.

For completeness, as specified in our pre-analysis plan, we also report results for our three secondary metrics by paper in Appendix Figure A1, Figure A2, and Figure A3. The results for statistical significance at the 10% level largely reflect those at the 5% level: statistical significance is substantially less likely in our resamples than in the original estimates for nine papers, while the majority of papers retain statistical significance. Results from the relative effect sizes and relative standard errors reveal that both shrunk point estimates and increased standard errors can explain the overall reduction in t-statistics.

4.2 Researcher Degrees of Freedom

Our main results establish that t-statistics shrink (with point estimates decreasing and standard errors increasing), and statistical significance is frequently lost, simply by resampling from the same data-generating process as original publications (i.e. holding research design fixed). If published results reflected random draws from its corresponding sampling distribution, then our resampling exercise should have produced a distribution of results that centered around the original published results. Instead, our results suggest that in conjunction with the sample generated utilizing Google Trends, authors must have made decisions (unconsciously or otherwise) that resulted in inflated t-statistics and increased statistical significance.

While we cannot directly observe the actual decisions authors engaged in, we can observe whether the "final" decision for how the Google Trends data was utilized differs from how it was originally generated. In this sense, we can observe the authors' "exercised" degrees of freedom, and see whether these are correlated with our primary metrics. As proxies for researcher degrees of freedom, we consider four variables: 1) whether the data utilized more than one Google Trends query, as multiple queries suggests the authors had greater flexibility in choosing the exact queries for their data, primarily by including multiple search terms,[17] 2) whether the authors transformed the Google Trends data in any fashion (e.g. taking logs), 3) whether only a subset of the downloaded Google Trends data was utilized, and 4) whether the authors decided to drop observations where the Google Trends value was equal to 0. Each of these four dimensions were identified in our pre-analysis plan.

The results from this exercise are presented in Table 3. Each column pertains to a single regression where our three primary metrics are utilized as outcome variables. In odd columns, we consider a measure for the count of exercised degrees of freedom (0-4), while even columns consider a model which separately identifies each of the four degrees of freedom (estimated as dummy variables). Despite the small

[17] For example, in one paper with multiple queries (not included in our study), the authors study how cashless payment technology impacts consumer behavior, and include searches for various supermarket chains. In another example with only one query, the authors utilize a weekly time series of searches for "1040" to illustrate a jump in searches around tax filing season. In the former example, authors have arguably more degrees of freedom in generating their dataset (e.g. choosing specific supermarket chains). Of course, authors who utilized only one query may have considered other queries that went unreported.

sample size of 24 papers, we see that exercised researcher degrees of freedom is strongly associated with our three primary metrics. Relative t-values and statistical significance are significantly reduced for papers that exercised more degrees of freedom, while the probability the resampled t-statistic shrinks increases in researcher degrees of freedom. When we break this down by each of the four components, we first find that the biggest contributor is the indicator for multiple Google Trends queries. Transforming the Google Trends data further leads to reduced relative t-values and lost statistical significance; the coefficient on the significance indicator is statistically significant at the 5% level, while the coefficient on the relative t-value itself is not. Taking subsamples of the Google Trends data points in the same direction across all three outcomes but is not statistically significant in any of them. Lastly, the decision to drop Google Trends indices equal to zero is actually positively associated with relative t-values and statistically significant resamples (though these coefficients are statistically insignificant), but still also positively associated with the probability of the t-statistic shrinking (p-value = 0.076).

4.3 Sampling Noise

We next consider two proxy measures for the sampling variation underlying the original Google Trends queries. We may expect to see different results for our metrics by sampling noise, since, all else equal, greater sampling variation implies the possibility of more extreme results. If extreme results are more likely to be reported, we'd expect to find a correspondence between sampling noise and our robustness metrics. Our first measure of sampling noise comes from the time series unit of the Google Trends query: daily, weekly, monthly, or greater than monthly. We assume that all else equal, queries with shorter time series units will inherently carry more noise, since weeks are aggregations of days, months are aggregations of weeks, etc.

Our second measure of sampling noise comes from estimating the average pairwise correlation across resamples for each query within a paper, then averaging across queries (for papers with more than one query).[18] This "cross-resample" correlation captures how consistent resamples are with each other, with stronger correlations implying less statistical variation. We find substantial variation in cross-resample correlation across papers: The majority of papers have strong cross-resample correlations (above 0.95), seven papers are very strong (above 0.99), while others are substantially lower (e.g., 0.37 and 0.44).[19] Note that we did not pre-register testing for heterogeneity in our results by this second measure, and thus results should be interpreted cautiously.

[18] With 40 resamples for each query, we calculate $\binom{40}{2} = 780$ pairwise correlations for each query.

[19] Unreported, some queries within papers produced pairwise correlations close to zero, suggesting the query produced indices that are as informative as random noise.

These results are presented in Table 4. In odd columns, we consider dummies for the time series utilized (daily, weekly, monthly), with the omitted category being greater than monthly, while the even columns consider the cross-resample correlation and an indicator for whether any queries reported by the original authors or in any of our resamples produced "missing" files. In column (1), we estimate a substantial drop in the relative t-value in papers that utilized daily data. The coefficients on weekly and monthly data are small in magnitude and statistically insignificant. From column (3), we see that these drops correspond to a significant drop in the share of resamples that were statistically significant. The magnitude of these effects are substantial as well: daily data produce relative t-values that are 57% smaller on average.

We additionally find that the cross-resample correlation is significantly associated with our primary metrics. The coefficient in column (2) is approximately equal to 0.95, suggesting that a 10% increase in the cross-resample correlation (i.e. less noise) is associated with a 10% increase in the relative t-value. In column (4), we see that papers with stronger cross-resample correlations were also more likely to produce statistically significant results in resamples, and in column (6), we see that stronger cross-resample correlations are associated with a smaller likelihood of a shrunk t-statistic. Lastly, the relative t-values and share of significant estimates were not impacted in papers which utilized queries that produced missing values, but we do find weak evidence that papers with missing queries were more likely to produce resamples that shrunk the t-statistic.

5 Discussion and Conclusion

A foundational assumption in empirical work is that estimates researchers observe and report are representative draws from the underlying sampling distribution. When this assumption holds, published results are informative about population parameters, and the process of hypothesis testing behaves as intended. Yet this assumption can fail quietly: a single noisy realization, selectively reported, may cause the published record to overstate the strength of empirical evidence, even when the underlying research design is executed flawlessly and no misconduct of any kind has occurred.

In this paper, we study this mechanism directly. We exploit a distinctive feature of Google Trends data: because its indices are constructed from random samples of searches, identical queries executed on different days yield different realizations of the underlying data-generating process. This allows us to observe multiple independent draws from the same sampling distribution while holding the research design, specification, and estimation procedure fixed. Replicating 24 papers published in top economics journals and re-estimating each result using 40 independent resamples, we find that t-statistics decline by 31% on aver-

age, statistical significance is frequently lost, and sign reversals occur with non-trivial probability. These patterns are systematically related to proxies for researcher degrees of freedom and to measures of sampling noise in the underlying Google Trends queries.

A few key features of our results warrant discussion. First, the mechanism we isolate does not require intentional p-hacking, strategic specification searching, or any form of research misconduct. It arises simply from the combination of two ordinary features of empirical research: relying on a single realization of an inherently stochastic data source, and the well-documented tendency for statistically significant results to be more likely to reach publication. When these two forces interact, sampling variation alone can substantially inflate the share of published results that cross conventional significance thresholds.

Second, the heterogeneity we document across papers is especially informative. For a majority of papers in our sample, our resampling exercise leaves statistical significance intact, suggesting that their results are robust to the particular realization of the Google Trends data that they originally downloaded. For a handful of papers, however, the results are highly sensitive to sampling variation, and it is these papers that drive the aggregate patterns we document. The correlation between sensitivity and proxies for researcher degrees of freedom further suggests that the fragile results are not simply unlucky draws, but are systematically related to choices made during data construction and analysis.

Third, our results have implications beyond Google Trends. Many widely used data sources in economics, including survey microdata, administrative records, and other algorithmically generated datasets, rely on sampling, aggregation, or filtering procedures whose randomness is rarely acknowledged or revisited. In these settings, the sampling variation embedded in data construction is not reflected in reported standard errors, and researchers typically have no opportunity to observe alternative realizations.[20] Google Trends is unusual precisely because it *can* be resampled, making the latent uncertainty visible. The broader implication is that the robustness of empirical results to data construction choices arguably deserves more scrutiny than it currently receives.

Finally, we note that our quantitative findings are strikingly similar to those of related literatures that measure robustness through different mechanisms. Campbell et al. (2026) systematically evaluate the robustness of 17 non-experimental papers published in the *American Economic Review* by subjecting each to a series of alternative, expert-validated specifications using the same data. Across their robustness checks, they find that approximately 50% remain statistically significant and that the relative t-value of the robustness tests is about 70% (with relative effect sizes averaging 74% and relative standard errors averaging

[20] Another context where it may be plausible to realize alternate observations from the same DGP is when replicating studies that use LLMs to generate measures.

134%). Brodeur et al. (2026) similarly find robustness reproducibility of 70% on the significance indicator and 77% on the relative t-value across a broader sample of 110 papers in economics and political science. Our corresponding metrics (a 55% significance rate and average relative t-values of 0.69) are of a comparable magnitude to both studies.

What is notable about this comparison is what differs across these exercises. Campbell et al. (2026) vary the specification: different control variables, clustering choices, sample restrictions, and functional forms, all tested on the same fixed dataset; Brodeur et al. (2026) similarly vary analytical decisions within the same data. Experimental replication efforts draw a new sample from a potentially different population, varying both the data-generating process and the context; Camerer et al. (2016) evaluate the replicability of laboratory experiments in economics, while Camerer et al. (2018) extend this approach to social science experiments published in *Nature* and *Science*, both finding replication rates well below 100%. Related work from Young (2019) also studies experimental contexts but instead conducts randomization inferences and juxtaposes his results against reported test statistics from regressions. Our exercise does none of the above: we hold the research design, specification, and data-generating process entirely fixed, varying only the random sample of searches that Google Trends happens to draw on a given day. The fact that our metrics are broadly similar to those from robustness exercises and experimental replications, even though we introduce no variation in research design, suggests that sampling variation in the data itself is a quantitatively important contributor to the fragility of empirical results.

We close by returning to the epigraph: Doing the same thing over and over and expecting different results may be a form of insanity, but in empirical research, doing the same thing over and over *does* produce different results. Our study documents how this variation, interacting with selective reporting, leads to inflated test statistics.

References

ANDREWS, I. AND M. KASY (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109, 2766–94.

ASHENFELTER, O., C. HARMON, AND H. OOSTERBEEK (1999): "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias," *Labour Economics*, 6, 453–470.

BERNOULLI, J. (1713): *Ars coniectandi*, Impensis Thurnisiorum, fratrum.

BRODEUR, A., S. CARRELL, D. FIGLIO, AND L. LUSHER (2023): "Unpacking p-hacking and publication bias," *American economic review*, 113, 2974–3002.

BRODEUR, A., M. LÉ, M. SANGNIER, AND Y. ZYLBERBERG (2016): "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 8, 1–32.

BRODEUR, A., D. MIKOLA, N. COOK, L. FIALA, ET AL. (2026): "Computational Reproducibility and Robustness of Empirical Economics and Political Science Research Between 2022 and 2023," *Nature*, forthcoming.

BRUNS, S. B., I. ASANOV, R. BODE, M. DUNGER, ET AL. (2019): "Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research," *Research Policy*, 48, 103796.

CAMERER, C. F., A. DREBER, E. FORSELL, T.-H. HO, J. HUBER, M. JOHANNESSON, M. KIRCHLER, J. ALMENBERG, A. ALTMEJD, T. CHAN, ET AL. (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436.

CAMERER, C. F., A. DREBER, F. HOLZMEISTER, ET AL. (2018): "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644.

CAMPBELL, D., A. BRODEUR, A. DREBER, M. JOHANNESSON, J. KOPECKY, L. LUSHER, AND N. TSOY (2026): "How Robust are Empirical Papers from the AER?" Working paper, Institute for Replication.

CEBRIÁN, E. AND J. DOMENECH (2023): "Is Google Trends a quality data source?" *Applied Economics Letters*, 30, 811–815.

DE LONG, J. B. AND K. LANG (1992): "Are all Economic Hypotheses False?" *Journal of Political Economy*, 100, 1257–1257.

DOUCOULIAGOS, C. AND T. D. STANLEY (2013): "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity," *Journal of Economic Surveys*, 27, 316–339.

DREBER, A. AND M. JOHANNESSON (2025): "A framework for evaluating reproducibility and replicability in economics," *Economic Inquiry*, 63, 338–356.

FERRARO, P. J. AND P. SHUKLA (2020): "Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy*, 14, 339–351.

FISHER, R. A. (1925): "Theory of statistical estimation," in *Mathematical proceedings of the Cambridge philosophical society*, Cambridge University Press, vol. 22, 700–725.

FRANCO, A., N. MALHOTRA, AND G. SIMONOVITS (2014): "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science*, 345, 1502–1505.

HAVRÁNEK, T. (2015): "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting," *Journal of the European Economic Association*, 13, 1180–1204.

HUNTINGTON-KLEIN, N., C. C. PORTNER, I. MCCARTHY, ET AL. (2025): "The sources of researcher variation in economics," Tech. rep., National Bureau of Economic Research.

IOANNIDIS, J. P. (2005): "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124.

IOANNIDIS, J. P., T. D. STANLEY, AND H. DOUCOULIAGOS (2017): "The Power of Bias in Economics Research," *Economic Journal*, 127, F236–F265.

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

LYBBERT, T. J. AND S. T. BUCCOLA (2021): "The Evolving Ethics of Analysis, Publication, and Transparency in Applied Economics," *Applied Economic Perspectives and Policy*, 43, 1330–1351.

MCCLOSKEY, D. N. (1985): "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests," *American Economic Review: Papers and Proceedings*, 75, 201–205.

MIGUEL, E., C. CAMERER, K. CASEY, J. COHEN, K. M. ESTERLING, A. GERBER, R. GLENNERSTER, D. GREEN, M. HUMPHREYS, G. IMBENS, ET AL. (2014): "Promoting Transparency in Social Science Research," *Science*, 343, 30–31.

NEYMAN, J. AND E. S. PEARSON (1933): "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.

SIMMONS, J. P., L. D. NELSON, AND U. SIMONSOHN (2011): "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.

STANLEY, T. D. (2005): "Beyond Publication Bias," *Journal of Economic Surveys*, 19, 309–345.

——— (2008): "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection," *Oxford Bulletin of Economics and Statistics*, 70, 103–127.

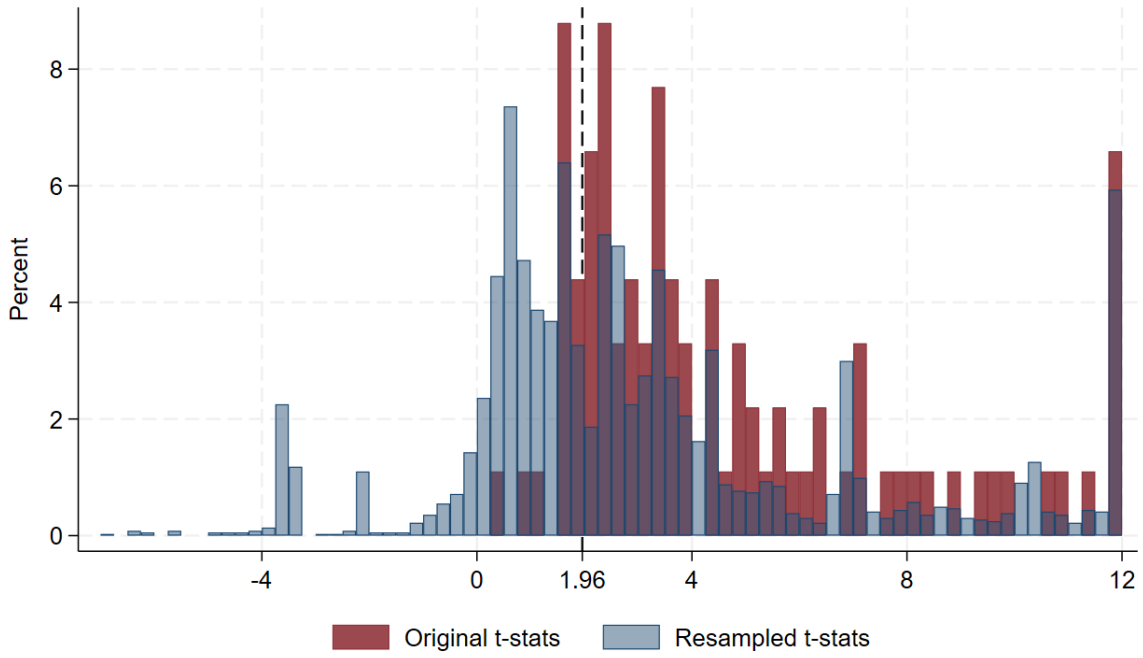
STERLING, T. D. (1959): "Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa," *Journal of the American statistical association*, 54, 30–34.

YOUNG, A. (2019): "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *The quarterly journal of economics*, 134, 557–598.

——— (2022): "Consistency Without Inference: Instrumental Variables in Practical Application," *European Economic Review*, 147, 104112.

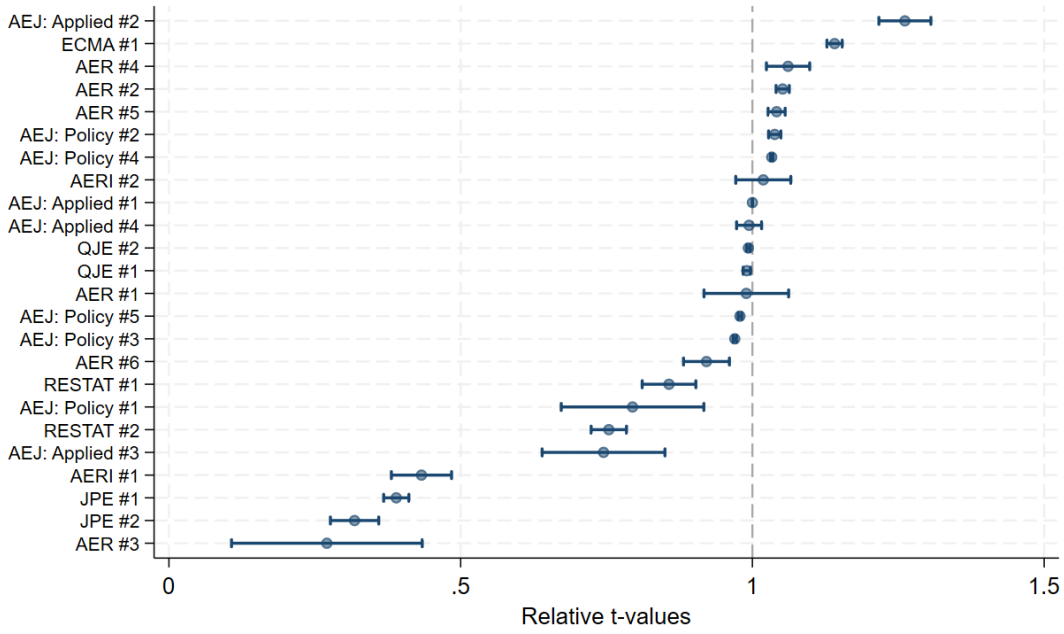
6 Figures and Tables

Figure 1: Distribution of original and resampled t-statistics



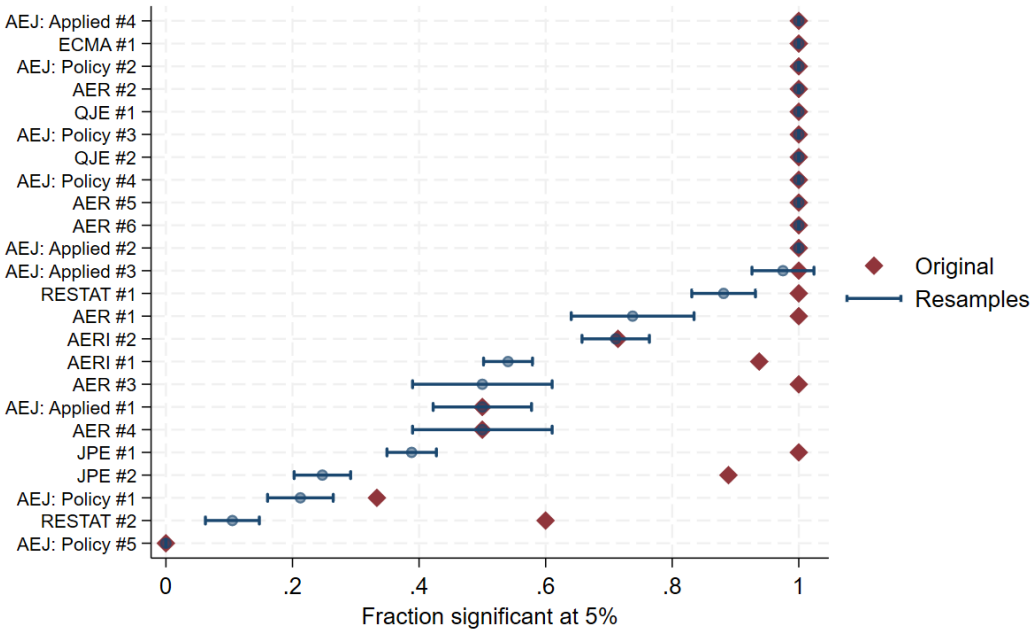
Notes: We plot the distributions of the t-statistics across the 24 papers separately by the original results (in red) and the resampled results (in blue). The t-statistics have been normalized so that all values from the original papers are positive. Resampled results are normalized so that t-statistics in the same direction as the original study are positive, and those in the opposite direction are negative. t-statistics over 12 are censored and included as 12. Total sample size is 3640 paper-estimate-resamples, with each paper-estimate receiving 40 resamples.

Figure 2: Relative t-values by paper



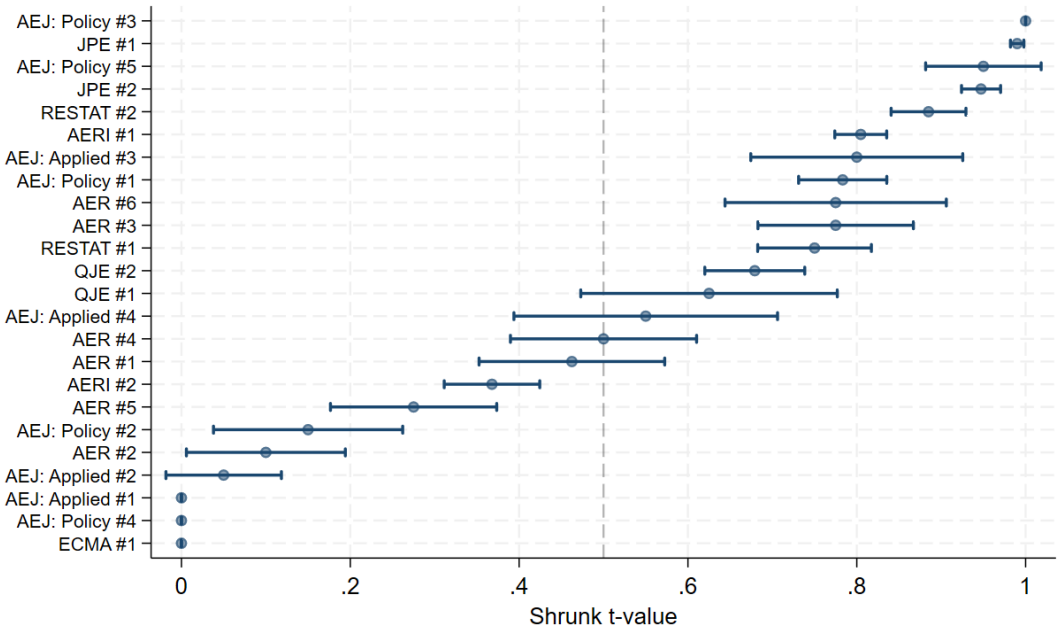
Notes: The relative t-value is calculated by dividing the resampled t-statistic by the original paper-estimate t-statistic. In this figure, we then collapse the data to the paper level, averaging over estimate-resamples, and report the mean relative t-value and its 95% confidence interval. Relative t-values greater (less) than 1 indicate larger (smaller) t-statistics from our resamples.

Figure 3: Share of results statistically significant at 5% by paper



Notes: We plot the share of original estimates (in red) and the share of resamples (in blue) that were statistically significant at the 5% level by paper.

Figure 4: Share of resamples with t-statistics smaller than original, by paper



Notes: We plot the share of resamples that produced t-statistics that were strictly smaller than the original paper-estimate's t-statistic, by paper.

Table 1: Summary statistics at the paper level

	Mean	SD	Min	Max
Characteristics of paper's data:				
- # of Google Trends queries	101.08	215.70	1	776
- More than 1 Google Trends query	0.42	0.50	0	1
- Transform Google Trends data (e.g. log)	0.33	0.48	0	1
- Google Trends used as outcome variable	0.92	0.28	0	1
- Google Trends used as treatment variable	0.12	0.34	0	1
- Take subsample of Google Trends data	0.46	0.51	0	1
- Drop Google Trends indices = 0	0.12	0.34	0	1
- Daily data	0.08	0.28	0	1
- Weekly data	0.29	0.46	0	1
- Monthly data	0.46	0.51	0	1
- Average correlation across resamples	0.86	0.19	0.37	>0.99
- Queries produce missing values	0.25	0.44	0	1
Characteristics of paper's findings:				
- # of original estimates	3.79	4.28	1	16
- Average t-statistic	8.25	12.10	1.58	61.43
- Average p-value	0.03	0.06	<0.001	0.20
- Share of estimates significant at 5%	0.85	0.27	0	1
Number of papers	24			

Notes: Sample includes papers published from 2015 through August 2025 in top economics journals (*Quarterly Journal of Economics*, *Econometrica*, *American Economic Review*, *Journal of Political Economy*, *Review of Economic Studies*, *Review of Economics and Statistics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Applied Economics*, and *American Economic Review: Insights*) that utilized Google Trends data and provided replicable data and code.

Table 2: Main results at the paper-estimate-resample level

Primary metrics	Range	Mean	95% CI
- Relative t-value: $\frac{\text{resample t-value}}{\text{original t-value}}$	-1.17 - 6.63	0.69	[0.67, 0.70]
- Resample statistically significant at 5%	0 - 1	0.55	[0.54, 0.57]
- Shrunk t-value: $\mathbb{1}$(resample t $<$ original t)	0 - 1	0.70	[0.69, 0.72]
Secondary metrics			
- Resample statistically significant at 10%	0 - 1	0.61	[0.59, 0.62]
- Relative effect size: $\frac{\text{resample effect size}}{\text{original effect size}}$	-2.58 - 12.67	0.88	[0.84, 0.91]
- Relative standard error: $\frac{\text{resample s.e.}}{\text{original s.e.}}$	0.15 - 10.11	1.40	[1.36, 1.43]

Notes: Sample includes 40 resamples for each original paper-estimate (i.e. 40 times 91 = 3640 observations). For the relative effect size and relative standard error metrics, we removed one outlier paper from the sample: this paper had substantially more variation in relative effect sizes (between -10.25 and 42.32) and relative standard errors (between 0.12 and 1134.69).

Table 3: Primary metrics as a function of exercised researcher degrees of freedom

	Relative t-value		% sig. at 5%		% shrunk t-stat	
	(1)	(2)	(3)	(4)	(5)	(6)
Researcher degrees of freedom	-0.136		-0.121		0.152	
	(0.047)		(0.043)		(0.052)	
	[0.008]		[0.011]		[0.008]	
More than 1 Google Trends query		-0.317		-0.289		0.326
		(0.099)		(0.070)		(0.124)
		[0.005]		[0.001]		[0.016]
Transform Google Trends data (e.g. log)		-0.099		-0.162		0.150
		(0.087)		(0.065)		(0.151)
		[0.269]		[0.022]		[0.333]
Take subsample of Google Trends data		-0.079		-0.026		-0.016
		(0.081)		(0.057)		(0.138)
		[0.343]		[0.656]		[0.912]
Drop Google Trends indices = 0		0.017		0.085		0.237
		(0.090)		(0.070)		(0.126)
		[0.849]		[0.241]		[0.076]
Observations	24	24	24	24	24	24
Mean of Y	0.88	0.88	0.72	0.72	0.55	0.55

Notes: Each column corresponds to a single regression. Relative t-value is calculated as the resampled t-statistic divided by the original t-statistic. Statistical significance is an indicator for whether the resampled t-statistic was significant at the 5% level. "Shrunk t-stat" is an indicator for whether the resampled t-statistic was smaller than the original t-statistic. These three metrics are then averaged across all estimate-resamples for each paper. Researcher degrees of freedom ranges between zero and four, and counts the number of indicators that switch on from the even columns. Standard errors in parentheses, p-values in brackets.

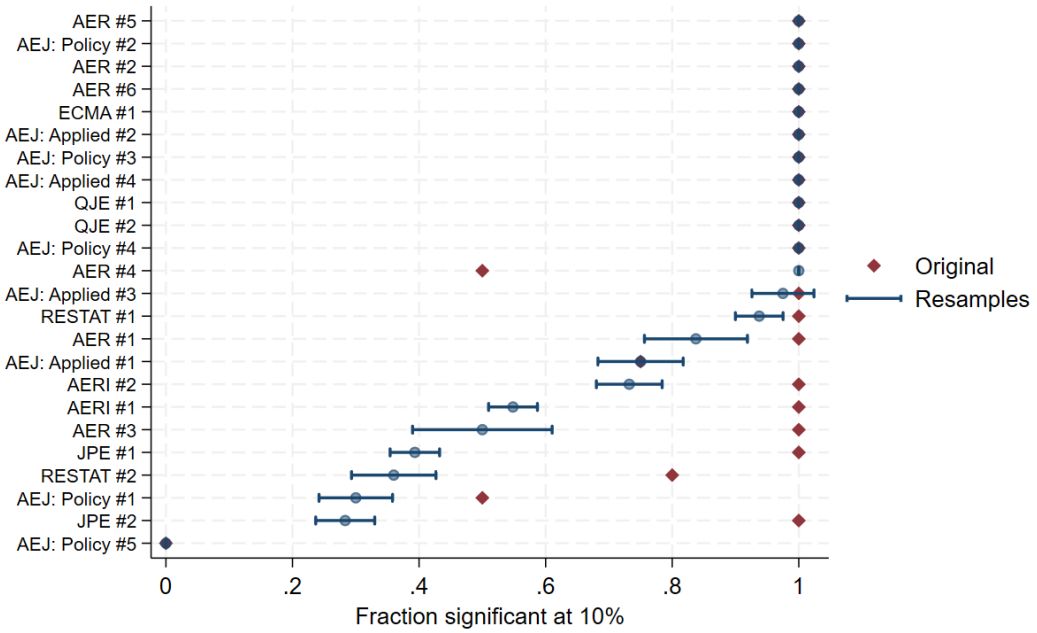
Table 4: Primary metrics as a function of sampling noise

	Relative t-value		% sig. at 5%		% shrunk t-stat	
	(1)	(2)	(3)	(4)	(5)	(6)
Daily data	-0.573		-0.438		0.369	
	(0.066)		(0.114)		(0.134)	
	[0.000]		[0.001]		[0.012]	
Weekly data	-0.043		0.116		-0.215	
	(0.127)		(0.141)		(0.182)	
	[0.738]		[0.423]		[0.251]	
Monthly data	0.023		0.133		-0.036	
	(0.089)		(0.123)		(0.178)	
	[0.798]		[0.295]		[0.840]	
Cross-resample correlation		0.951		0.672		-0.613
		(0.199)		(0.141)		(0.212)
		[0.000]		[0.000]		[0.009]
Queries produce missing values		0.034		-0.015		0.197
		(0.113)		(0.106)		(0.091)
		[0.770]		[0.892]		[0.042]
Observations	24	24	24	24	24	24
Mean of Y	0.88	0.88	0.72	0.72	0.55	0.55

Notes: Each column corresponds to a single regression. Relative t-value is calculated as the resampled t-statistic divided by the original t-statistic. Statistical significance is an indicator for whether the resampled t-statistic was significant at the 5% level. "Shrunk t-stat" is an indicator for whether the resampled t-statistic was smaller than the original t-statistic. These three metrics are then averaged across all estimate-resamples for each paper. "Cross-resample correlation" is the average pairwise correlation across resamples for each query within a paper. Standard errors in parentheses, p-values in brackets.

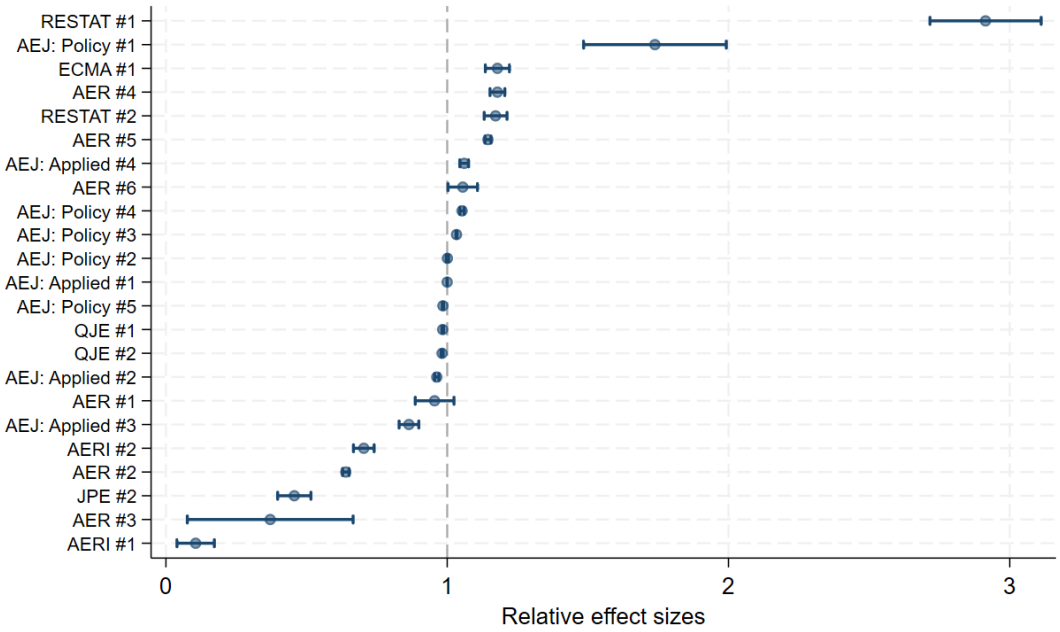
A Appendix Figures and Tables

Figure A1: Share of results statistically significant at 10% by paper



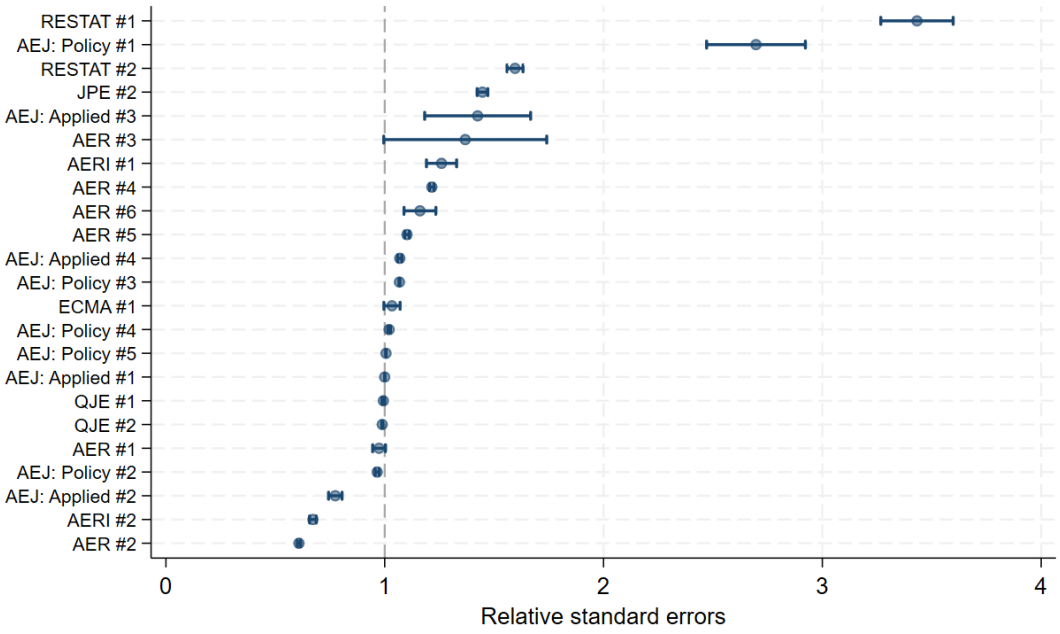
Notes: We plot the share of original estimates (in red) and the share of resamples (in blue) that were statistically significant at the 10% level.

Figure A2: Relative effect sizes by paper



Notes: The relative effect size is calculated by dividing the resampled effect size by the original paper-estimate effect size. In this figure, we then collapse the data to the paper level, averaging over estimate-resamples, and report the mean relative effect size and its 95% confidence interval. Relative effect sizes greater (less) than 1 indicate larger (smaller) effect sizes from our resamples. One outlier paper is dropped from the figure.

Figure A3: Relative standard errors by paper



Notes: The relative standard error is calculated by dividing the resampled standard error by the original paper-estimate standard error. In this figure, we then collapse the data to the paper level, averaging over estimate-resamples, and report the mean relative standard error and its 95% confidence interval. Relative standard errors greater (less) than 1 indicate larger (smaller) standard errors from our resamples. One outlier paper is dropped from the figure.

Table A1: List of publications from top economics journals using Google Trends (GT) data

Paper	Journal (Year)	Description	Role of GT
- The Effect of Terrorism on Employment and Consumer Sentiment	AEJ: Applied (2018)	Studies how terrorist attacks affect local employment and consumer sentiment, showing that successful attacks generate higher search volume on Google Trends while failed attacks have smaller effects.	Outcome
- Social Media and Corruption*	AEJ: Applied (2018)	Examines whether social media can discipline corruption in nondemocratic settings, finding that anti-corruption blog posts caused negative stock returns in targeted Russian state-controlled companies, with effects on corporate governance amplified when blogger search intensity was higher.	Treatment interaction
- Minority Salience and Political Extremism	AEJ: Applied (2021)	Tests whether heightened minority salience increases political extremism, finding higher extremist vote shares in mosque municipalities after Ramadan, with Muslim-related search intensity confirming the salience shock.	Outcome
- The Power of Example: Corruption Spurs Corruption*	AEJ: Applied (2021)	Investigates whether corruption scandals induce further dishonest behavior; uses Google Trends data on audit-related searches descriptively to confirm public salience of corruption revelations.	Outcome
-The Political Boundaries of Ethnic Divisions	AEJ: Applied (2021)	Studies how redrawing political boundaries affects ethnic divisions and intergroup conflict, controlling for Google Trends searches for district names to account for media attention bias in violence reporting after splitting.	Control
- Unemployment Insurance as a Worker Indiscipline Device? Evidence from Scanner Data*	AEJ: Applied (2022)	Investigates whether more generous unemployment insurance reduces on-the-job worker effort, using Google Trends searches for unemployment-related terms to verify worker awareness of the benefit extensions.	Outcome
- From Hashtag to Hate Crime: Twitter and Antiminority Sentiment	AEJ: Applied (2023)	Examines whether social media exposure and influential anti-Muslim tweets translate into offline hate crimes, controlling for contemporaneous general salience of Muslim-related topics using search intensity.	Control
- Contagious Dishonesty: Corruption Scandals and Supermarket Theft	AEJ: Applied (2023)	Studies spillovers from corruption scandals to retail theft, using scandal-related search intensity to validate that scandals generate sufficient public awareness to plausibly affect behavior.	Outcome
- Energy Saving May Kill: Evidence from the Fukushima Nuclear Accident	AEJ: Applied (2023)	Investigates behavioral responses to energy-saving campaigns following the Fukushima accident, using search intensity for 'energy-saving' to capture households' active information-seeking about electricity conservation.	Outcome
- Media Coverage of Immigration and the Polarization of Attitudes*	AEJ: Applied (2025)	Tests whether media attention to immigration polarizes attitudes, using immigration-related search activity descriptively to validate that spikes in TV coverage reflected genuine increases in public attention.	Outcome
- Taxpayer Search for Information: Implications for Rational Attention	AEJ: Policy (2015)	Studies how taxpayers seek information about tax policy, using search intensity to measure information demand and attention to both tax filing/compliance and tax planning decisions.	Outcome
- Telecracy: Testing for Channels of Persuasion	AEJ: Policy (2015)	Examines how television exposure affects political persuasion, using search behavior to verify that voters did not substitute toward online political information after losing access to analog TV.	Outcome
- Consumers' Response to State Energy Efficient Appliance Rebate Programs	AEJ: Policy (2017)	Analyzes consumer responses to energy efficiency rebates; uses GT search volume for the rebate program as descriptive evidence of when consumers became aware of the program.	Outcome

Paper	Journal (Year)	Description	Role of GT
- Direct and Spillover Effects of Middle School Vaccination Requirements*	AEJ: Policy (2019)	Studies the effects of vaccination mandates, using search activity to capture parental attention to vaccination-related information.	Outcome
- How Taxing Is Tax Filing? Using Revealed Preferences to Estimate Compliance Costs*	AEJ: Policy (2020)	Estimates compliance costs of tax filing, using search intensity for tax-related terms to provide evidence of taxpayer procrastination around the filing deadline.	Outcome
- Criminal Deterrence when There Are Offsetting Risks: Traffic Cameras, Vehicular Accidents, and Public Safety	AEJ: Policy (2020)	Examines the effect of red-light camera programs on traffic accidents and public safety, using search intensity for camera-related terms to verify public awareness of the referendum that exogenously ended the program.	Outcome
- Advertising and Environmental Stewardship: Evidence from the BP Oil Spill	AEJ: Policy (2020)	Studies environmental disasters and firm behavior, using search intensity to capture public attention to environmental issues.	Outcome
- Raising the Bar: Minimum Wages and Employers' Hiring Standards	AEJ: Policy (2022)	Analyzes minimum wage effects on hiring standards, using search intensity for 'Mindestlohn'(minimum wage) to validate the timing of employer anticipation of the policy.	Outcome
- Removing Welfare Traps: Employment Responses in the Finnish Basic Income Experiment	AEJ: Policy (2022)	Studies labor supply responses to basic income, using search intensity for 'basic income' to verify that recipients were aware of receiving the unconditional payment.	Outcome
- Digitization and the Market for Physical Works: Evidence from the Google Books Project	AEJ: Policy (2023)	Examines digitization of books, using Google search measures as controls for consumer interest in physical works.	Control
- More than Words: Leaders' Speech and Risky Behavior during a Pandemic	AEJ: Policy (2023)	Studies how leaders' anti-science speech affects social distancing behavior, using search intensity for protest and pronouncement terms to validate the salience of key events.	Outcome
- Economic Effects of Environmental Crises: Evidence from Flint, Michigan*	AEJ: Policy (2023)	Examines the effects of water contamination in Flint, using search behavior to capture public attention and concern.	Outcome
- Public Pensions and Private Savings*	AEJ: Policy (2024)	Studies how pension policy affects household savings, using search intensity to verify reform awareness.	Outcome
- Understanding the Link between Temperature and Crime	AEJ: Policy (2024)	Studies how temperature affects crime through time use and alcohol consumption, using search intensity as a supplementary outcome to verify the temperature-alcohol relationship.	Outcome
- Who Watches the Watchmen? Local News and Police Behavior	AEJ: Policy (2025)	Studies local news coverage and police behavior, using search intensity to measure the salience of crime in public opinion.	Outcome
- Pay Transparency and Gender Equality*	AEJ: Policy (2025)	Examines how pay transparency affects gender outcomes, using search intensity to capture public attention to firms' gender pay disclosures.	Outcome
- Pandering in the Shadows: How Natural Disasters Affect Special Interest Politics	AEJ: Policy (2025)	Studies disaster-driven political behavior, using search intensity alongside TV news coverage to measure crowd-out of voter attention to politics.	Outcome
- Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing*	AER (2015)	Examines media effects on teen childbearing, using search intensity as both an outcome and a treatment proxy.	Outcome/Treatment
- Consumers as Tax Auditors*	AER (2019)	Studies whether consumer rewards for requesting receipts improve firm tax compliance, using Google search volume for the program name as a validation check that consumers pay attention to lottery result announcements.	Outcome
- Sources of Inaction in Household Finance: Evidence from the Danish Mortgage Market*	AER (2020)	Examines mortgage refinancing inertia, using search behavior to proxy for information acquisition.	Outcome

Paper	Journal (Year)	Description	Role of GT
- Long-Run Growth of Financial Data Technology*	AER (2020)	Studies how technological progress shifts financial analysis from fundamental research toward demand-based trading strategies, using search intensity for 'order flow' versus 'fundamental analysis' as suggestive evidence of this transition.	Outcome
- The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment*	AER (2020)	Examines whether income transparency widens the well-being gap between richer and poorer individuals, using search activity for tax records to capture public attention to income information.	Outcome
- The Welfare Effects of Peer Entry: The Case of Airbnb and the Accommodation Industry	AER (2022)	Studies Airbnb entry, using search intensity as a control or instrument for local demand.	Control/Instrument
- Can You Move to Opportunity? Evidence from the Great Migration	AER (2022)	Examines the long-run effects of the Great Migration on upward mobility in northern destination cities, using search behavior as a proxy for persistent racial animus in destination commuting zones.	Outcome
- Conflict and Intergroup Trade: Evidence from the 2014 Russia-Ukraine Crisis*	AER (2023)	Studies trade disruptions during conflict, using search intensity to measure the intensity of consumer boycott activity as one mechanism driving the trade decline.	Outcome
- Financial Technology Adoption: Network Externalities of Cashless Payments in Mexico	AER (2024)	Examines adoption of cashless payments, using search activity to measure consumer shopping behavior.	Outcome
- Enforcing Wealth Taxes in the Developing World: Quasi-experimental Evidence from Colombia	AER: Insights (2021)	Studies wealth tax enforcement, using search volume to illustrate public attention to the Panama Papers leak.	Outcome
- The COVID-19 Pandemic Disrupted Both School Bullying and Cyberbullying*	AER: Insights (2022)	Examines changes in bullying during COVID-19 using Google search intensity as a real-time proxy for actual bullying behavior, validated against pre-pandemic survey data.	Outcome
- Partisan Fertility and Presidential Elections*	AER: Insights (2022)	Studies partisan fertility responses to a surprise presidential election outcome, using GT search data to corroborate fertility intent across partisan and ethnic groups.	Outcome
- Tell Me Something I Don't Already Know: Learning in Low- and High-Inflation Settings*	Econometrica (2025)	Examines learning under inflation, using search intensity to measure information demand.	Outcome
- Something to Talk About: Social Spillovers in Movie Consumption*	JPE (2016)	Studies social spillovers in movie consumption, using search behavior as both an outcome and treatment proxy to show momentum is locally rather than nationally driven.	Outcome/Treatment
- Attack When the World Is Not Watching? US News and the Israeli-Palestinian Conflict*	JPE (2018)	Examines conflict timing and media attention, using search intensity to measure public attention.	Outcome
- Auctions versus Posted Prices in Online Markets	JPE (2018)	Compares market mechanisms online, using search behavior to proxy for consumer interest.	Outcome
- The Tail That Wags the Economy: Beliefs and Persistent Stagnation	JPE (2020)	Studies belief-driven stagnation, using search intensity as evidence that crisis-related economic beliefs remained persistently elevated.	Outcome
- The Effects of Fluoride in Drinking Water	JPE (2021)	Examines health and labor market effects of fluoride, using search intensity for fluoride to measure public concern about fluoride exposure.	Outcome
- Excess Sensitivity of High-Income Consumers*	QJE (2018)	Tests whether consumption exhibits excess sensitivity to predictable income payments, using Google Trends search intensity to measure household attention and awareness.	Outcome

Paper	Journal (Year)	Description	Role of GT
- Kinship, Cooperation, and the Evolution of Moral Systems	QJE (2019)	Examines how kinship structure shapes moral systems, using search behavior to measure the relative salience of shame versus guilt across societies.	Outcome
- Race and Economic Opportunity in the United States: an Intergenerational Perspective	QJE (2020)	Studies intergenerational mobility, using Google search intensity for racial epithets as a proxy for explicit racial bias among whites across local areas.	Treatment
- The Assessment Gap: Racial Inequalities in Property Taxation	QJE (2022)	Examines racial inequality in property taxation, using Google search-based measures of racial animus to test whether the assessment gap in property tax burdens faced by Black and Hispanic homeowners is larger in areas with greater racial prejudice.	Control
- What We Teach About Race and Gender: Representation in Images and Text of Children's Books	QJE (2023)	Studies representation of race, gender, skin color, and age in a century of award-winning children's books, using Google search intensity to establish the outsized public reach and influence of mainstream award-winning books.	Outcome
- Evolution vs. Creationism in the Classroom: The Lasting Effects of Science Education	QJE (2024)	Examines how science education standards affect students' knowledge, beliefs, and career choices, using search behavior to verify public attention to evolution and creationism.	Outcome
- A Monetary-Fiscal Theory of Sudden Inflations*	QJE (2025)	Studies sudden inflation dynamics, using search intensity for fiscal-inflation terms to proxy for household attention to government finances.	Outcome
- The Impact of Unemployment Insurance on Job Search	RESTAT (2016)	Examines UI effects on job search, using Google search data to measure job search effort.	Outcome
- Geography, Ties, and Knowledge Flows	RESTAT (2019)	Studies citation networks, using search behavior to measure knowledge flows.	Outcome
- Do People Avoid Morally Relevant Information?	RESTAT (2021)	Examines information avoidance, using search intensity as both an outcome and control.	Outcome/Control
- The Welfare Consequences of Centralization: Evidence from a Quasi-Natural Experiment in Switzerland	RESTAT (2021)	Studies centralization effects on individual welfare, using Google search data to measure Swiss residents' awareness of and information about the reforms.	Outcome
- Coronavirus Perceptions and Economic Anxiety*	RESTAT (2021)	Studies how pandemic perceptions drive economic anxiety, using search intensity as both treatment and outcome.	Outcome
- Collective Reputation in Trade: Evidence from the Chinese Dairy Industry	RESTAT (2022)	Studies reputation effects, using Google search ratios to measure the accuracy of consumer information about the parties directly involved in the scandal across destination countries.	Outcome
- How Do Taxpayers Respond to Public Disclosure and Social Recognition Programs? Evidence from Pakistan	RESTAT (2022)	Examines tax compliance responses to public disclosure and social recognition programs, using search intensity to capture public attention.	Outcome
- Stuck in the Seventies: Gas Prices and Consumer Sentiment	RESTAT (2022)	Examines how gas prices affect consumer sentiment, using Google search volume to measure consumers' attention.	Outcome
- The Impact of Social Networks on EITC Claiming Behavior	RESTAT (2022)	Examines EITC claiming behavior, using search intensity to measure information diffusion.	Outcome
- Inflation and Attention Thresholds	RESTAT (2023)	Studies attention to inflation, using search behavior to measure attention thresholds.	Outcome
- Impulse Purchases, Gun Ownership, and Homicides: Evidence from a Firearm Demand Shock	RESTAT (2023)	Examines firearm purchase delay laws during a demand shock, using search intensity to separate purchase intentions from actual sales.	Outcome
- Mums Go Online: Is the Internet Changing the Demand for Health Care?	RESTAT (2023)	Studies internet use and healthcare demand, using search behavior to measure information seeking.	Outcome

Paper	Journal (Year)	Description	Role of GT
- Fear and the Safety Net: Evidence from Secure Communities*	RESTAT (2024)	Examines fear and welfare use, using search intensity to capture perceived risk.	Outcome
- Economic Distress and Electoral Consequences: Evidence from Appalachia	RESTAT (2024)	Studies how poverty labels affect voting, using search behavior to measure public awareness of the ARC (Appalachian Regional Commission).	Outcome
- The Impact of Fear on Police Behavior and Public Safety	RESTAT (2024)	Examines how fear of workplace risk drives police behavior, using search intensity to measure public awareness.	Outcome
- Jam-Barrel Politics	RESTAT (2024)	Studies executive-legislative exchange of targeted road contracts for legislative support in Colombia, using search behavior to document public awareness of the practice.	Outcome
- Financial Constraints and Propagation of Shocks in Production Networks	RESTAT (2024)	Examines shock propagation through production networks, using Google search data to confirm the tax shock was unanticipated.	Outcome
- Public Scrutiny, Police Behavior, and Crime Consequences: Evidence from High-Profile Police Killings	RESTAT (2025)	Studies police behavior and crime after high-profile police killings, using search behavior to identify the onset of public scrutiny.	Treatment
- Tax Policy Transmission and Household Expenditures	RESTAT (2025)	Examines household spending responses to India's 2017 GST reform, using Google search intensity and media coverage to verify consumer awareness of the policy.	Outcome
- Migrants, Ancestors, and Foreign Investments	REStud (2019)	Studies migration networks, using search behavior to measure information flows.	Outcome
- The Effectiveness of Hiring Credits	REStud (2019)	Examines hiring subsidies' employment effects, using Google search activity to measure public awareness/knowledge of the policy.	Outcome