P-HACKING FOR POPULARITY?

STATISTICAL SIGNIFICANCE AND MEDIA ATTENTION*

Brianna Funderburk, Sunmi Jung, Lester Lusher†

University of Pittsburgh

October 1, 2025

Abstract

If media outlets cover "sensational" research, and if studies producing these results have greater p-hacking, then one would find positive correlations between marginal statistical significance and media attention. We code 7,022 estimates from a random sample of 404 NBER working papers and precisely estimate little difference in the distribution of test statistics by media attention. Our estimates differ significantly from forecasts made by 215 economists, most of whom predicted positive correlations. We further find abstract views and downloads are uncorrelated with statistical significance. In exploratory analyses, we document patterns in statistical significance and ChatGPT-generated measures of interest, influence, and convincingness.

Keywords: p-hacking; selective reporting; media attention; Altmetric

JEL codes: A11, C13, C40

*We thank Nicholas Halliwell for providing excellent research assistance. All errors are our own.

†Corresponding author: Lester Lusher, Department of Economics, University of Pittsburgh and IZA, Email: lesterlusher@pitt.edu. Brianna Funderburk, Department of Economics, University of Pittsburgh, Email: brf94@pitt.edu. Sunmi Jung, Department of Economics, University of Pittsburgh, Email: suj49@pitt.edu.

1 Introduction

A long and growing literature has tackled issues related to publication biases and p-hacking in academia.[1] Publication biases (biases in the peer review process that favor statistical significance) and p-hacking (actions authors engage in to attain statistical significance) are typically identified by documenting statistical bunching at certain thresholds (e.g. 5% significant) in the distribution of published test statistics (e.g., Gerber and Malhotra, 2008a,b; Brodeur et al., 2016; Vivalt, 2019). Further research in economics has disentangled p-hacking from publication bias (Brodeur et al., 2023), and identified other factors that contribute to increased statistical bunching, including the paper's identification strategy (Brodeur et al., 2020), editorial statements on null findings (Blanco-Perez and Brodeur, 2020), data type and data sharing policies (Brodeur et al., 2024a), and pre-registration and pre-analysis plans (Brodeur et al., 2024b).

In this study, we explore the relationship between statistical significance and media attention in economics research. Understanding this relationship is important given the role public dissemination plays in determining how academia judges research quality (Fuoco, 2021). While all academics list their publications on their webpage and CV, many more include media coverage of their research and interviews they conduct with media outlets.

The media is also crucial for the diffusion of academic research, increasing the actionability of research by influencing decision-makers and policy prescriptions. If the distribution of test statistics among papers covered by the media displays greater heaping than non-covered papers, then the consequences of p-hacking and publication biases would likely be exacerbated since research with greater dissemination would on average contain more false positives. Despite the important connection between media and academia, to the best of our knowledge, no study yet has investigated the relationship between statistical significance and media coverage in economics research. Perhaps the most closely related study to ours comes from Brodeur et al. (2025), who utilize data from three health journals to document increased marginal statistical significance among media-covered studies.[2]

If the media simply covered a random or representative sample of all academic research, we would expect media-covered papers to display the same level of statistical bunching as seen among published papers. However, there may be additional channels through which the distribution of test statistics in papers covered by the media differs from those that do not receive media attention. While media decisions on

[1] A non-exhaustive list includes 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), Furukawa (2020), Havránek (2015), Ioannidis (2005), Ioannidis et al. (2017), Leamer (1983), Lybbert and Buccola (2021), McCloskey (1985), Miguel et al. (2014), Stanley (2005), and Stanley (2008).

[2] There are notable differences between health journals and economics journals, including the direct publishing of test statistics in the paper's abstract in health journals, and the overall higher viewership of health publications relative to economics publications.

which research to cite or cover has been studied extensively (see Fleerackers, 2023, for a review), the general concern is that media coverage may gravitate toward "sensational" results (Brown et al., 2018; Dempster et al., 2022). Papers with results that are more likely to attract media attention may then be more likely to contain false positives, particularly for more controversial or politicized results, and thus the distribution of findings covered by the media may contain even greater heaping at various statistical thresholds.

Furthermore, this concern is arguably especially relevant in economics. Outside of economics, most disciplines require that authors do not disseminate their results until their paper has survived the peer review process. In economics, however, the majority of researchers release "working papers" prior to peer review and publication. Lusher et al. (2023) show that National Bureau of Economic Research working papers (NBER WPs) receive more views and downloads upon release than their published counterparts, suggesting that the media may be more likely to cover economics research that has not yet been vetted by the peer review process and thus possibly contain more type I errors. On one hand, if publication biases are pervasive in the peer review process (i.e. favors statistical significance), then media covering (a representative sample of) working papers instead of published work would smooth the distribution of test statistics; on the other hand, the media may be more likely to cover "faulty" working papers that fail to publish, and these studies may contain more false positives. Because Brodeur et al. (2023) suggests publication biases are not so prevalent in economics (i.e. observed bunching in the distribution of published test statistics can be more attributed to p-hacking), the greater concern perhaps is that media-covered research possesses additional statistical heaping beyond what's already produced in the profession overall.

To test this hypothesis, we construct a sample consisting of 7,022 hand coded estimates from a random sample of 404 NBER WPs.[3] We focus on the NBER WP series since they produce what is largely regarded as the most impactful research in economics and thus arguably more likely to be covered by the media. Measures for whether a paper received media attention and the degree and sources of media attention are provided by Altmetric, a company that tracks academic papers across news outlets (e.g. The New York Times) and social media (e.g. Twitter/X). Following the previous literature, we focus on papers with a clear identification strategy (difference-in-differences, experiment/randomized control trial, instrumental variable, regression discontinuity), and we collected estimates on the primary coefficient(s) of interest from main results tables. Each of our papers belongs to at least one of four NBER programs: Labor Studies, Economics of Education, Economics of Health, and Environment and Energy Economics. We oversampled from papers that received media attention in order to increase power in detecting differences between papers that received

[3]Described in further detail later, our original sample included 1,000 NBER WPs, 600 of which did not receive media attention, 400 of which did receive media attention. Papers were then dropped if they did not contain a clear identification strategy i.e. all 404 papers in our sample include some form of causal interpretation of results.

media attention vs. those that did not (213 without media attention, 191 with media attention).

In order to generate predictions over the expected relationship between media attention and statistical significance, we first conducted a prediction survey with 215 economists (i.e. "forecasters"). Forecasters were required to have or be pursuing a PhD in economics or finance, and recruitment was carried out by circulating the prediction survey among professional networks. Approximately 40% of forecasters were faculty members, another 40% were PhD students, and 20% held non-faculty positions (e.g. private sector). Forecasters were asked to make predictions for the point estimates from our baseline bi-variate regressions (described further later), and were rewarded based on the accuracy of their responses. By and large, forecasters predicted a positive relationship between marginal statistical significance and media attention: For instance, at the 5% threshold, 83.7% of forecasters predicted a positive relationship, 10.2% predicted a point estimate of zero, and 6.0% predicted a negative relationship.

Turning to the main exercise, we first document significant bunching in the distribution of test statistics among our sample of NBER WPs. The distribution is very similar to those identified in Brodeur et al. (2023) and Brodeur et al. (2020), particularly with extreme heaping at 5% significance, illustrating that the behaviors that drive bunching among initial journal submissions and final publications in economics apply to the NBER WP series as well.

We then compare the distribution of test statistics for NBER WPs that received some media attention against those that did not. In general, the distributions look very similar, with papers that received media attention having a slightly *smaller* peak at 5% significance. Econometric tests for differences in marginal significance (using Caliper tests with author and paper controls) detect no statistically significant differences in bunching. Our null results are precisely estimated as well: For instance, in our fully specified model, with 95% confidence we are able to rule out coefficients on a dummy for receiving media attention of larger than 3.6 percentage points in the probability of an estimate being marginally significant at the 5% level. In other words, we can confidently conclude that for every 100 reported point estimates, papers that received media attention have no more than 3.6 additional marginally significant estimates at the 5% level. Given the mean number of reported main estimates per paper is less than 20, our models are able to rule out effect sizes of even just one additional point estimate being statistically significant at the 5% level in papers that received media attention. We estimate similarly small differences at the 1% and 10% significance thresholds. Importantly, our results differ significantly from the distribution of predictions made by forecasters, who overestimated the relationship between marginal statistical significance and media attention.

We then consider alternate proxies for popularity to test for whether papers that garnered more attention also had greater heaping at significance thresholds. These include Altmetric's overall composite score for

how much media attention the NBER WP received, as well as abstract views and downloads within the first month of the NBER WP's release. Across these additional measures, we again fail to find any evidence of positive selection between a paper's popularity and its propensity to have statistically significant estimates. Occasionally, downloads are negatively related to significance at the 10% and 1% levels, though these findings are sensitive to controls and bandwidth.

Lastly, we utilize a large language model (ChatGPT o3) to generate "scores" for each NBER WP along the dimensions of how much media attention the paper received, and how "interesting," "influential," and "convincing" each paper is. As a validation exercise, we first see that ChatGPT's media attention score is very strongly correlated with whether the paper received any media attention in the Altmetric data. Papers that were rated as more "interesting" by ChatGPT were also more likely to have received media attention. Regarding marginal statistical significance, we find only weak evidence of some potential correlations. For example, papers rated as more "interesting" had less statistical significance at the 10% level and more statistical significance at the 5% level, though this result is sensitive to bandwidth choice. "Influential" papers are less likely to possess marginally statistically significant estimates at the 5% level, which perhaps suggests that papers that tend to p-hack at the 5% level are doing analyses that are less impactful on the profession overall. Finally, we find no correlations between a paper's "convincingness" and marginal significance.

Altogether, our results highlight little concern about the media differential covering certain economics papers based on whether they contain more marginally significant estimates, a proxy for authors engaging in p-hacking. This stands in contrast to the opinions of economists, who forecasted a positive relationship between media attention and marginal statistical significance. At a minimum, our results should assuage concerns that economics research covered by the media is more likely to be "false" than the overall set of working papers produced (at least among NBER WP). Our results further suggest that economics papers which may be particularly sensational are either not receiving special attention from media outlets or are not more likely to be p-hacked. The lack of a correlation further implies that media outlets are unlikely directly responding to a paper's statistical significance—this is perhaps unsurprising, as media outlets likely care more about a paper's headline versus the statistical precision of the paper's main result.

2 Data Sources and Background

Our primary data come from the National Bureau of Economic Research (NBER), a leading network of over 1,800 economists with academic positions at North American institutions. Economists are accepted into the NBER through a rigorous selection process. 47 current or former NBER affiliates have won a Nobel

prize, and 13 have chaired the President's Council of Economic Advisers. As such, the NBER is often regarded as the preeminent network of economists in academia.

The main goal of the NBER is to share recent, high-quality academic research produced by economists. The primary way the NBER disseminates this research is through its working paper (WP) series. Every week, NBER affiliates submit their working papers for release the following Monday. Annually, more than 1,200 working papers are distributed to over 900 subscribing organizations and numerous individual subscribers. Each WP is submitted as part of at least one of 19 research programs (e.g. Labor Studies). NBER WPs are not yet peer-reviewed at the time of submission to the series. More information on the NBER and its WP series can be found at https://www.nber.org/about-nber.

We then utilize data from Altmetric. Altmetric is a platform that tracks online attention to academic research papers, including NBER WPs, providing a comprehensive view of how research is being shared and discussed across various channels on the internet. The platform aggregates data from a wide range of sources to provide insights into the impact of scholarly articles, papers, and other academic work. Their sources include social media platforms (such as Twitter, Facebook, and LinkedIn), mainstream media outlets (such as CNN, Fox News, and the New York Times), policy documents, and other digital spaces.

Altmetric's primary measure is their "Altmetric Attention Score." This score is calculated based on the volume, source, type of attention, and engagement level a paper receives. For example, mentions in major news outlets or policy documents contribute more to the score than a simple tweet. Volume captures how many times a piece of research has been mentioned, shared, or discussed across different platforms; source reflects how some high-profile outlets, such as major news organizations or influential policy documents, tend to carry more weight than mentions in smaller blogs or social media posts; type of attention reflects how mentions in policy reports or a citation in a widely read newspaper article contributes more significantly than a tweet or Facebook post; engagement level captures shares, comments, likes, and other forms of interaction of the paper. Note that Altmetric acts independently from citations in academic journals i.e. the Altmetric Attention Score focuses solely on online and social media engagement.

An Altmetric Attention Score of zero means the paper did not receive any media attention across Altmetric's monitored platforms. We use this measure as our primary outcome variable to identify whether a NBER WP received any media attention. Utilizing data on the population of NBER WP releases from 2004-2019 matched with their Altmetric statistics, Lusher et al. (2023) find that approximately 15% of NBER WPs receive some media attention. For comparison, measuring media mentions directly through major media outlets, Ziegler et al. (2021) estimates one in 11 NBER WPs receives media attention upon its first month of release.

Furthermore, we link each NBER WP to its corresponding webpage on RePEc (Research Papers in Economics). RePEc is often considered the central platform for housing and disseminating all types of economic research, including both working and published papers. The site hosts over 4.4 million research pieces from over 4,000 journals and 5,600 working paper series, with over 71,000 registered authors. RePEc further provides rankings of authors and institutions based on research productivity and citations. We use RePEc data to measure abstract views and downloads for each NBER WP.

Finally, for each NBER WP in our sample, we utilize a large language model (ChatGPT o3) to generate four additional measures. These were generated from four prompts: "On a scale from 0-100, where 100 means the most media attention, how much media attention did the NBER working paper [paper title] receive upon its release?" and "On a scale from 0-100, where 100 means the most [interesting/influential/convincing], how [interesting/influential/convincing] was the NBER working paper [paper title] receive upon its release?"[4] We then rescaled the ChatGPT scores by dividing by 10.

Our sample includes NBER WPs released between 2004 and 2020 in at least one of four research programs: Labor Studies, Economics of Education, Economics of Health, and Environment and Energy Economics. From this sample of 6,647 NBER WPs (1,362 of which received media attention, 5,285 did not), we randomly selected 1,000 papers to "read" for potential coding. Following Brodeur et al. (2020) and Brodeur et al. (2023), upon reading the manuscript, we removed manuscripts which did not contain a clear experiment or identification strategy, including difference in differences, instrumental variables, or regression discontinuity designs, and/or a randomized control trial or experiment.[5] From 6,647 WPs (1,362 with media attention, 5,285 without), we randomly selected 1,000 papers to read. To ensure statistical power, we oversampled from papers that received media attention (600 without, 400 with).After excluding papers lacking a clear identification strategy (e.g., DiD, IV, RDD, RCT), our final analytic sample contains 404 papers, of which 213 did not receive media attention and 191 did.

For each of these 404 papers, we coded coefficients and their standard errors from the main "treatment" variable(s) in any main results tables. Estimates from summary statistics, appendices, robustness checks, placebo tests, and figures were not collected. Within main tables, we omitted any coefficients not corresponding to the primary treatment variable (e.g. controls). Any cases of ambiguity were flagged and removed in our primary analyses, while robustness analyses check for the sensitivity to the inclusion of ambiguous cases. This process, including a focus on main results tables, closely follows that of Brodeur

[4] We additionally instructed the model to "Just give the one number in your response, and do not output anything other than that one number. Always give a response."

[5] Examples of omitted papers include theoretical papers, literature reviews, methodology papers, descriptive exercises, structural estimations, and other identification strategies such as propensity score matching.

et al. (2016), Blanco-Perez and Brodeur (2020), Brodeur et al. (2020), and Brodeur et al. (2023). Our final sample includes 7,022 test statistics.[6]

2.1 Summary Statistics

Appendix Table A1 presents summary statistics for our analytic sample at the paper level, split by whether the paper received media attention or not. Papers with media attention tend to have fewer main estimates (15.7) compared to those that did not (18.9).[7] While there does not appear to be any seasonality to which NBER WP's get covered by the media (in terms of month of release), more recent NBER WPs were much more likely to receive media attention (average year of 2016 vs. 2013). We also include the number of WPs released in the same week, a potentially important control since Lusher et al. (2023) show that heavier release weeks reduce attention to individual papers (conditional on week of year fixed effects).

Next, we see that over half of NBER WPs in our sample belong to the "Labor Studies" program, with the second most common program being "Economics of Health."[8] Author characteristics also appear to be roughly balanced across media attention, including the number of authors on the paper and the number of prior NBER WPs written by the authors on the paper. The media also does not cover papers differentially by identification strategy.

The final rows of Appendix Table A1 summarize additional outcomes: Altmetric Attention Score, abstract views and downloads (within the first month of release of the NBER WP), and the four ChatGPT scores (media attention, interest, influence, convincingness). Appendix Table A2 formally tests for correlations between media attention and these measures. We first find that downloads are positively associated (though only weakly) with media attention. Then, we see that the ChatGPT score for "media attention" is very strongly related to whether the paper actually received media attention, lending credence to the large language model's ability to classify papers. Finally, we find a strong, positive correlation between ChatGPT's "interesting" score and whether the paper received media attention. The remaining outcomes (abstract views, influence, and convincingness) remain uncorrelated with media attention.

[6] The entirety of papers in our sample reported coefficients and standard errors (as opposed to solely reporting p-values or t-statistics). We construct z-statistics as the ratio of the coefficient and standard error and thus assume they follow an asymptotically standard normal distribution under the null hypothesis.

[7] Our regression analyses will use the inverse of the number of tests presented in the same article to weight observations in order to account for the imbalance in the number of test statistics across paper type.

[8] Note that NBER WPs can belong to more than one program, and thus the sum of shares across programs exceeds one.

3 Methods

We employ two methods for identifying differences in statistical significance by media attention. The first is a visual inspection of the distribution of z-statistics separately for papers that did vs. did not receive media attention. We only include estimates on the interval $z \in [0, 10]$ with bins of width 0.1. If papers that received media attention possess greater p-hacking, then we'd expect to see sharper peaks in z-statistics just above 1.65, 1.96, and/or 2.58.

Our econometric approach borrows from several studies (e.g. Gerber and Malhotra, 2008a; Brodeur et al., 2020, 2023) to conduct Caliper tests. The Caliper test quantifies and conducts inference on the extent of the statistical bunching. Moreover, the Caliper test can directly compare two different distributions of test statistics (e.g., papers with vs. without media attention). In essence, the Caliper test compares the number of test statistics in a narrow range above and below a specific statistical significance threshold (e.g. 5% significant).

We start with the equation:

$$Pr(Significant_{pe} = 1) = \Phi(\alpha + \beta \text{Media}_p + \delta X_p) \tag{1}$$

where $Significant_{pe}$ is an indicator variable for whether estimate e from paper p is statistically significant. We estimate equation (1) via logit models, with standard errors clustered at the paper level. For each statistical threshold of interest (1.65, 1.96, 2.58), we estimate equation (1) using the windows $z \pm 0.50$ and $z \pm 0.30$. The primary variable of interest Media_p is an indicator for whether the paper received media attention, while X_p contains the control variables presented in Appendix Table A1. We report marginal effects from the logit regressions such that β can be interpreted as the predicted change in percentage points in the probability that a specific estimate is statistical significant if the estimate came from a paper that received media attention. Additional analyses replace Media_p with other variables of interest: Altmetric Attention Score, abstract views and downloads, and ChatGPT-generated scores.

3.1 Prediction survey

In order to generate point estimates and distributions of expected values for β, we conducted a prediction survey with 215 economists (henceforth "forecasters"). Forecasters were given the above information regarding our random sampling of NBER WPs, each of which would come from one of four NBER programs, contain a valid identification strategy, and be matched to Altmetric data to determine media attention. Forecasters were asked to generate their best guess for β at each statistical threshold (1.65, 1.96, 2.58) with

the bandwidth $z \pm 0.50$ from equation (1) while omitting control variables X_p. This way, forecasters had to simply predict the bi-variate correlation between media attention and statistical significance. Examples were given over the interpretation of β, and were always presented symmetrically (positive and negative examples) so as not to prime participants into a specific direction.[9] Forecasters were also told that mechanically β could not be less than -1 or greater than 1. The full survey provided to forecasters can be found in the appendix.

In order to incentivize accurate predictions, forecasters were given the following payment scheme:

$$\$15 - (\overline{Sq.Error} \times 1500) \tag{2}$$

where the average of the squared errors $\overline{Sq.Error}$ is calculated across the forecaster's predictions for the three different β's (at 1.65, 1.96, and 2.58). Each squared error is the squared difference between the forecaster's prediction for β and the actual β estimated from our data. Note that forecasters were simply making predictions over point estimates, not the statistical precision of estimates (which would also be a function of the number of papers and estimates collected). Forecasters were also given a baseline of $5 for participating, and thus could earn up to $20 in total via an Amazon gift card.

We recruited forecasters via our own professional networks (i.e. emailing faculty and student colleagues and asking them to share with others), posting on the Economic Science Association Google Groups page, and conducting the survey via the Social Science Prediction Platform.[10] Forecasters needed to either have a PhD in economics or finance, or be currently enrolled in a PhD program in economics or finance. Our final sample consisted of 215 forecasters: 32.1% identified as female, 51.3% were located in North America, 32.1% were located in Europe, 40.9% were faculty members, 40.5% were PhD students, and 40.5% identified Labor/Public/Health/Environment/Education as their field of expertise.

4 Main Results

4.1 Media attention

We start with Appendix Figure A1, which first plots the distribution of z-statistics for our full sample of NBER WPs. The distribution displays a clear hump in test statistics that starts around 1.5, and peaks around 2.0. In the next two panels of Appendix Figure A1, we present the distribution of test statistics from

[9] For example, we included the text "β = 0.05 means that an estimate from a paper that received media attention is 5 percentage points more likely to be statistically significant. β = -0.05 means that an estimate from a paper that received media attention is 5 percentage points less likely to be statistically significant."

[10] More on the Social Science Prediction Platform can be found at https://socialscienceprediction.org/.

initial journal submissions to the *Journal of Human Resources* from Brodeur et al. (2023) (panel b), and the distribution of published test statistics among 25 leading economics journals from Brodeur et al. (2020) (panel c). Overall, the distribution from NBER WPs closely reflects the distribution from these two studies, though with slightly more bunching just after the 10 percent level threshold in Brodeur et al. (2023). Thus, it appears that any p-hacking behavior among NBER WPs is not particularly different from p-hacking among papers submitted to the *Journal of Human Resources* nor from the distribution of published statistics from top economics journals.

Next, we present our main visual results in Figure 1, where we split the sample of z-statistics by whether the NBER WP received media attention (panel a) or not (panel b). We first note that both distributions of test statistics display a hump that peaks around 2.0. Otherwise, no noticeable differences between the distributions is apparent, other than perhaps a slightly higher peak at 2.0 for papers *without* media attention. Potential concerns that papers that receive media attention are more likely to be p-hacked are certainly not apparent when juxtaposing the two panels.

To econometrically test for differences in marginal significance across the two distributions, we turn to our Caliper tests. We start with Table 1, where the covariate of interest is an indicator for whether the paper received media attention. We report standard errors adjusted for clustering by article in parentheses, and we weight observations by the inverse of the number of tests presented in the same article. Column (1) considers a simple bi-variate regression, and across all three statistical thresholds (10%, 5%, 1%), we estimate a negative relationship between media attention and statistical significance. As the coefficients reflect marginal effects from logit regressions, the coefficients can be interpreted as percentage point changes. For instance, at the 5% level, an estimate that received media attention is 2.9 percentage points less likely to be marginally statistically significant. As we move to the second and third columns, we include additional control variables, and in columns (4) through (6), we consider a narrower bandwidth ($z \pm 0.30$). Across all specifications considered, the point estimates remain negative and remarkably stable.

To interpret the economic magnitude of our effects, we consider our fully specified model at the 5% significance threshold in column (3). Our point estimate suggests that an estimate from a NBER WP that received media attention is 4.5 percentage points less likely to be statistically significant, or for every 100 reported test statistics, we'd expect less than five to lose statistical significance. Given the mean number of reported test statistics is 18.9 for papers without media attention and 15.7 for those with media attention, the -0.045 point estimate suggests that on average, less than one reported test statistic loses statistical significance among NBER WPs that received media attention. With a standard error of 0.041, we can similarly approximate a 95% confidence interval of less than two changed estimates in either direction. For example,

we can rule out point estimates larger than 3.6 (-0.045 + 1.96 $\times$ 0.041), and thus our model rules out effect sizes of even just one additional point estimate being statistically significant at the 5% level in papers that received media attention. In total, we interpret the results from Table 1 as precise "null" effects.

For robustness, in Appendix Table A3, we reproduce Table 1 while keeping any estimates we flagged as ambiguous (i.e. uncertain whether the estimate constituted a main result), and the results remain largely the same. In Appendix Table A4, we consider a continuous measure of media attention (Altmetric Attention Score), and again find precisely estimated null effects, though some coefficients flip and become positive. Still, in these cases, the estimates are especially tiny; for example, in column (3) for 1% significance, doubling the amount of media attention a NBER WP received leads to a 0.6 percentage point increase in the probability an estimate is marginally significant.

4.2 Prediction survey

We next juxtapose our estimates against those made by forecasters in our prediction survey. Forecasters were asked to make three predictions (at the 10%, 5% and 1% levels) from our bi-variate model with $z \pm 0.50$ i.e. the coefficients reported in column (1) of Table 1. In Figure 2, we plot the distribution of forecasts made at the 10% and 5% significance levels (results for 1% can be found in Appendix Figure A2). For reference, we plot the actual point estimate with a red line. Overall, forecasters predicted a positive relationship between marginal statistical significance and media attention. At the 10% threshold, 79.5% of forecasters predicted a positive relationship, 14.0% predicted a point estimate of zero, and 6.5% predicted a negative relationship; the corresponding percentages at the 5% threshold were 83.7%, 10.2%, and 6.01%. The mean forecast was 0.161 (with a 95% confidence interval of [0.129, 0.193]) at the 10% level, and 0.165 (95% confidence interval of [0.133, 0.197]) at the 5% level i.e. on average, forecasters predicted a significant positive relationship between media attention and statistical significance. This stands in stark contrast to the estimates we find, and highlights how our null results differ from overall expectations in the profession.[11]

4.3 Abstract views and downloads

While our main analysis focuses on media attention (as approximated by Altmetric), other measures of viewership could arguably be equally as important. Next, in Table 2, we consider the number of abstract views and downloads within the first month of the NBER WP's release. For ease of interpretation, we take

[11] In unreported analyses, we find some evidence that professors had more accurate forecasts than non-professors at the 5% level (p=0.058). However, they were not more accurate at the 1% and 10% levels, and their forecasts at the 5% level were still significantly positive (mean forecast of 0.148 and 95% confidence interval of [0.107, 0.189]).

logs of these two variables such that coefficients can be interpreted as the change in probability of an estimate being statistically significant in response to doubling the amount of abstract views or downloads a paper received. Once again, we mostly find statistically insignificant relationships between viewership and statistical significance. The only exception comes from columns 1-3 for significance at the 1% level, where we estimate a significant (at the 5% level) negative relationship between downloads and statistical significance; this result however is sensitive to bandwidth choice, as we no longer retain statistical significance when we narrow the bandwidth to 2.58 ± 0.30 in columns 4-6.[12]

Moreover, the economic magnitudes of these effects are small. Focusing on the point estimate from column (3) for 1% significant, doubling the number of downloads a NBER WP received is associated with less than one additional reported estimate losing statistical significance. At the 5% level, we can similarly rule out meaningful positive effects: For example, using our fully specified model in column (3), we can rule out point estimates larger than 0.046 (0.016 + 1.96 $\times$ 0.015), which again corresponds to less than one additional estimate gaining statistical significance (in response to doubling the amount of downloads a NBER WP received). Thus, the pattern of results for abstract views and downloads mimic the results from media attention: There is little difference in the propensity for marginal statistical significance by viewership of the article.

4.4 Large Language Model (LLM) measures

Lastly, as an exploratory analysis, Table 3 reports results using four additional LLM-generated measures (ChatGPT o3). Appendix Table A7-Table A10 present results when each measure is estimated separately. For "Media score," the model predicts no statistically significant effects at the 10% and 5% levels, but a negative correlation at the 1% level; however, Appendix Table A7 shows no relationship across all three thresholds. These results largely confirm the primary finding of a null relationship between media attention and statistical significance.

Turning to the remaining three scores, papers with higher "interesting scores" show less statistical significance at the 10% level but more at the 5% level. This suggests that papers that select into topics or results that the LLM finds interesting also tend to p-hack by moving estimates from the 10% threshold to the 5% threshold. "Influential score" is negatively correlated with significance at the 5% level i.e. NBER WPs that the LLM predicted to be influential possess *less* p-hacking at the 5% level. This possibly suggests that papers that do have a greater propensity to p-hack have less influence on the profession overall. Finally,

[12] As abstract views and downloads are highly correlated with each other, in Appendix Table A5 and Table A6, we separately report results by abstract views and downloads. The pattern of results remain largely the same, though we detect some weak evidence of downloads being negatively associated with statistical significance at the 10% level (columns 4-6).

we estimate no relationship between statistical significance and "convincing score". This last result is potentially interesting because it suggests the LLM did not assume a paper is more convincing just because it possessed more statistical significance. Of course, for all three measures, it is likely that the LLM generated scores not only based on the content within the NBER WP itself, but also based on any reaction to the paper made available online (including whether the paper received media attention), and thus these results should be interpreted cautiously.

5 Conclusion

This study provides the first systematic evidence on the relationship between statistical significance and media attention in economics research. Using more than 7,000 hand-coded estimates from a random sample of 404 NBER working papers, we show that papers receiving media coverage do not display systematically greater bunching at conventional significance thresholds than those that do not. These results stand in stark contrast to the expectations of professional economists, who overwhelmingly forecast a positive association between marginal significance and media coverage.

Across multiple measures of visibility—including Altmetric scores, abstract views, and downloads—we consistently find precise null effects. At most, media-covered papers differ by less than a single marginally significant estimate, and in some cases, the relationship is weakly negative. Exploratory analyses using Large Language Model-generated measures of "interest," "influence," and "convincingness" further suggest that the media and broader attention dynamics in economics are not tightly linked to the statistical significance of reported results.

Taken together, these findings imply that concerns about the media amplifying p-hacked or spurious results in economics may be overstated—at least within the domain of NBER working papers. While sensational results may drive journalistic interest in other disciplines, economics research covered by the media appears no more prone to marginal significance than the broader pool of work. More generally, our results highlight a disconnect between economists' perceptions of what drives media coverage and the empirical reality. Understanding what factors actually shape public visibility of economics research—and how they interact with research credibility—remains a promising direction for future work.

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.

BLANCO-PEREZ, C. AND A. BRODEUR (2020): "Publication Bias and Editorial Statement on Negative Findings," *Economic Journal*, 130, 1226–1247.

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

BRODEUR, A., N. COOK, AND A. HEYES (2020): "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110.

BRODEUR, A., N. COOK, AND C. NEISSER (2024a): "P-hacking, data type and data-sharing policy," *The Economic Journal*, 134, 985–1018.

BRODEUR, A., N. M. COOK, J. S. HARTLEY, AND A. HEYES (2024b): "Do preregistration and preanalysis plans reduce p-hacking and publication bias? evidence from 15,992 test statistics and suggestions for improvement," *Journal of Political Economy Microeconomics*, 2, 527–561.

BRODEUR, A., N. M. COOK, A. HEYES, AND T. WRIGHT (2025): "Media Stars: Statistical Significance and Research Impact," Tech. rep., I4R Discussion Paper Series.

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

BROWN, D. K., S. HARLOW, V. GARCÍA-PERDOMO, AND R. SALAVERRÍA (2018): "A new sensation? An international exploration of sensationalism and social media recommendations in online news publications," *Journalism*, 19, 1497–1516.

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.

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

DEMPSTER, G., G. SUTHERLAND, AND L. KEOGH (2022): "Scientific research in news media: a case study of misrepresentation, sensationalism and harmful recommendations," *Journal of Science Communication*, 21, A06.

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

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.

FLEERACKERS, A. (2023): "Unreviewed science in the news: Why and how journalists cover preprint research," .

FUOCO, R. (2021): "How to get media coverage and boost your science's impact." *Nature*.

FURUKAWA, C. (2020): "Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method," MIT mimeo.

GERBER, A. AND N. MALHOTRA (2008a): "Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 3, 313–326.

GERBER, A. S. AND N. MALHOTRA (2008b): "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research*, 37, 3–30.

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

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

LUSHER, L., W. YANG, AND S. E. CARRELL (2023): "Congestion on the information superhighway: Inefficiencies in economics working papers," *Journal of Public Economics*, 225, 104978.

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.

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.

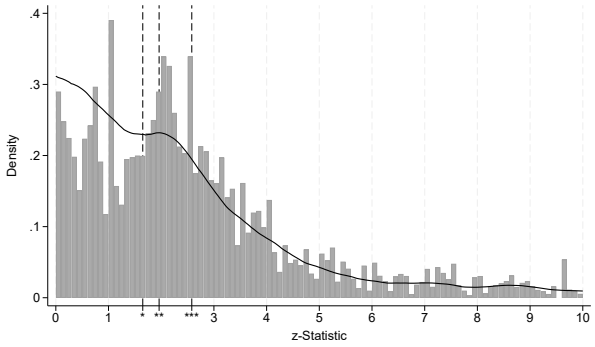
VIVALT, E. (2019): "Specification Searching and Significance Inflation Across Time, Methods and Disciplines," *Oxford Bulletin of Economics and Statistics*, 81, 797–816.

ZIEGLER, L. ET AL. (2021): *What is the Media Impact of Research in Economics?*, Universität Wien, Department of Economics.

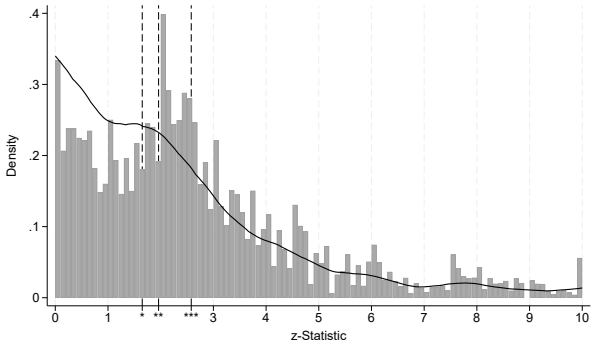
6 Figures and Tables

Figure 1: Distributions of z-statistics from NBER WPs with vs. without media attention

(a) NBER WPs w/ media attention (n=191 papers)



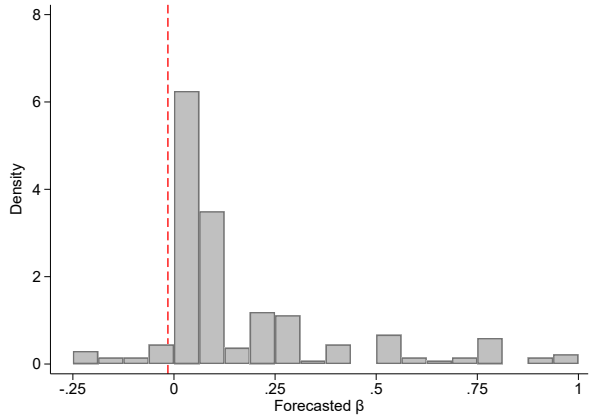
(b) NBER WPs w/out media attention (n=213 papers)



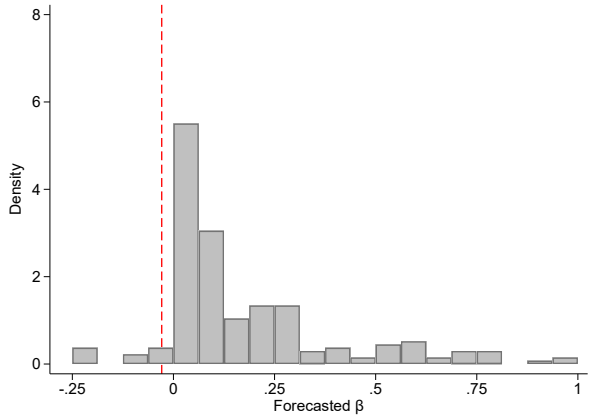
Notes: The first figure displays a histogram of test statistics for $z \in [0, 10]$, with bins of width 0.1, among NBER working papers that received at least some media attention after its release. The second figure covers NBER working papers that failed to receive media attention after its release. For each histogram, we superimpose an Epanechnikov kernel density curve. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 2: Distributions of responses from the prediction survey

(a) Predicted estimates at the 10% significance level



(b) Predicted estimates at the 5% significance level



Notes: Each panel presents a histogram of survey respondents' predicted coefficients (forecasted β) for the indicated significance level. The vertical red dashed line marks the actual estimated coefficient from our analysis (Figure 1, Column 1). The mean survey prediction is 0.161 (95% CI [0.1288, 0.1931]) in the 10% panel and 0.165 (95% CI [0.1326, 0.1971]) in the 5% panel. To improve readability, values below -0.25 were censored and recoded as -0.25 before constructing the histograms. Bins are of width 0.0625.

Table 1: Received media attention: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Received media attention	-0.015	-0.020	-0.010	-0.012	-0.012	-0.013
	(0.044)	(0.046)	(0.046)	(0.056)	(0.059)	(0.061)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Received media attention	-0.029	-0.054	-0.045	-0.031	-0.057	-0.044
	(0.041)	(0.044)	(0.041)	(0.054)	(0.059)	(0.055)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Received media attention	-0.007	-0.018	-0.018	-0.009	-0.034	-0.030
	(0.047)	(0.055)	(0.051)	(0.058)	(0.065)	(0.062)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table 2: Abstract views and downloads: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
ln(abstract views)	-0.005	-0.004	-0.012	-0.028	-0.031	-0.032
	(0.026)	(0.026)	(0.025)	(0.036)	(0.036)	(0.034)
ln(downloads)	-0.014	-0.015	-0.016	-0.031	-0.032	-0.035
	(0.018)	(0.018)	(0.018)	(0.020)	(0.022)	(0.021)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
ln(abstract views)	-0.018	-0.010	-0.014	-0.003	0.007	0.008
	(0.023)	(0.023)	(0.022)	(0.033)	(0.033)	(0.033)
ln(downloads)	0.022	0.023	0.016	0.028	0.027	0.023
	(0.016)	(0.017)	(0.015)	(0.022)	(0.023)	(0.020)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
ln(abstract views)	0.020	0.021	0.031	-0.006	0.001	0.009
	(0.028)	(0.028)	(0.029)	(0.033)	(0.032)	(0.033)
ln(downloads)	-0.040**	-0.040**	-0.039**	-0.037*	-0.036	-0.034
	(0.018)	(0.018)	(0.018)	(0.023)	(0.022)	(0.023)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table 3: ChatGPT scores: Caliper test, significance at the 10%, 5%, and 1% levels

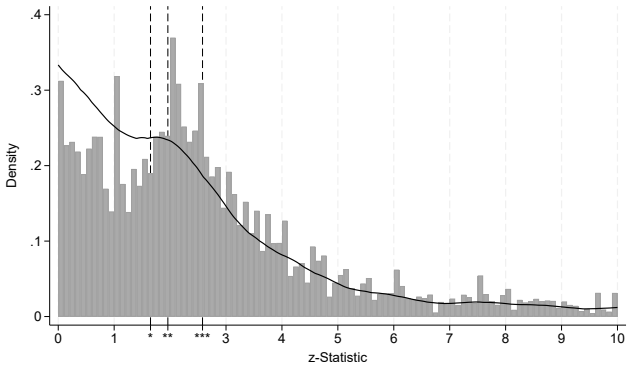
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Media score (0-10)	0.011	0.010	0.016	0.011	0.016	0.021
	(0.021)	(0.020)	(0.020)	(0.022)	(0.021)	(0.022)
Interesting score (0-10)	-0.082*	-0.080*	-0.061	-0.135***	-0.130**	-0.131**
	(0.045)	(0.045)	(0.047)	(0.051)	(0.051)	(0.054)
Influential score (0-10)	-0.010	-0.012	-0.014	0.036	0.036	0.038
	(0.024)	(0.024)	(0.025)	(0.031)	(0.031)	(0.030)
Convincing score (0-10)	-0.015	-0.012	-0.005	0.002	-0.001	0.017
	(0.030)	(0.030)	(0.029)	(0.034)	(0.034)	(0.034)
Observations	1222	1222	1222	716	716	716
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Media score (0-10)	-0.017	-0.023	-0.020	-0.018	-0.030	-0.025
	(0.018)	(0.018)	(0.018)	(0.022)	(0.020)	(0.021)
Interesting score (0-10)	0.014	0.010	0.033	0.102**	0.097**	0.136***
	(0.034)	(0.036)	(0.032)	(0.048)	(0.044)	(0.043)
Influential score (0-10)	-0.042**	-0.044**	-0.055***	-0.044*	-0.048*	-0.067***
	(0.019)	(0.019)	(0.018)	(0.025)	(0.025)	(0.022)
Convincing score (0-10)	0.005	0.008	0.005	0.004	0.012	-0.002
	(0.031)	(0.031)	(0.028)	(0.033)	(0.031)	(0.032)
Observations	1190	1190	1190	747	747	747
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Media score (0-10)	-0.056**	-0.057**	-0.053**	-0.061**	-0.059**	-0.059**
	(0.025)	(0.024)	(0.022)	(0.028)	(0.028)	(0.027)
Interesting score (0-10)	0.007	0.009	0.026	0.021	0.034	0.047
	(0.041)	(0.040)	(0.038)	(0.051)	(0.045)	(0.043)
Influential score (0-10)	0.012	0.012	0.013	0.025	0.031	0.031
	(0.018)	(0.019)	(0.017)	(0.026)	(0.026)	(0.026)
Convincing score (0-10)	-0.002	-0.002	-0.004	0.032	0.031	0.018
	(0.029)	(0.029)	(0.028)	(0.043)	(0.040)	(0.038)
Observations	957	957	957	584	584	584
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

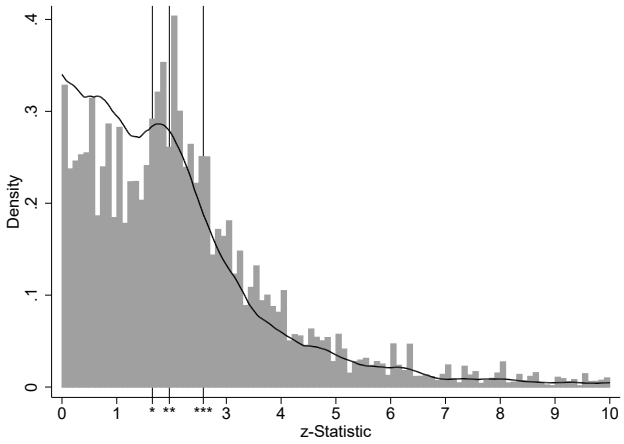
A Appendix Tables and Figures

Figure A1: Distributions of z-statistics from our sample of NBER WPs vs. z-statistics from Brodeur et al. (2023) and Brodeur et al. (2020)

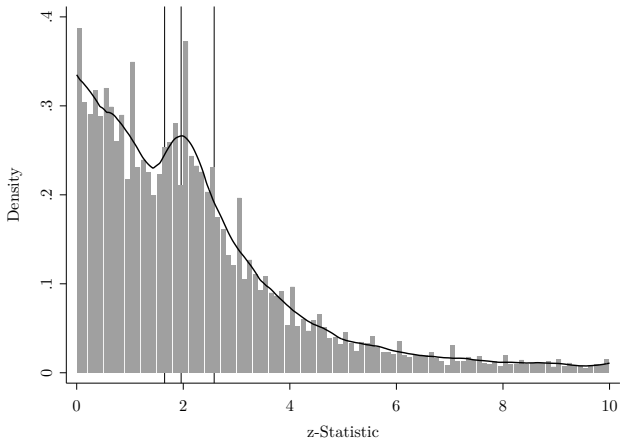
(a) Sample of NBER WPs



(b) Brodeur et al. (2023) - Initial submissions to JHR

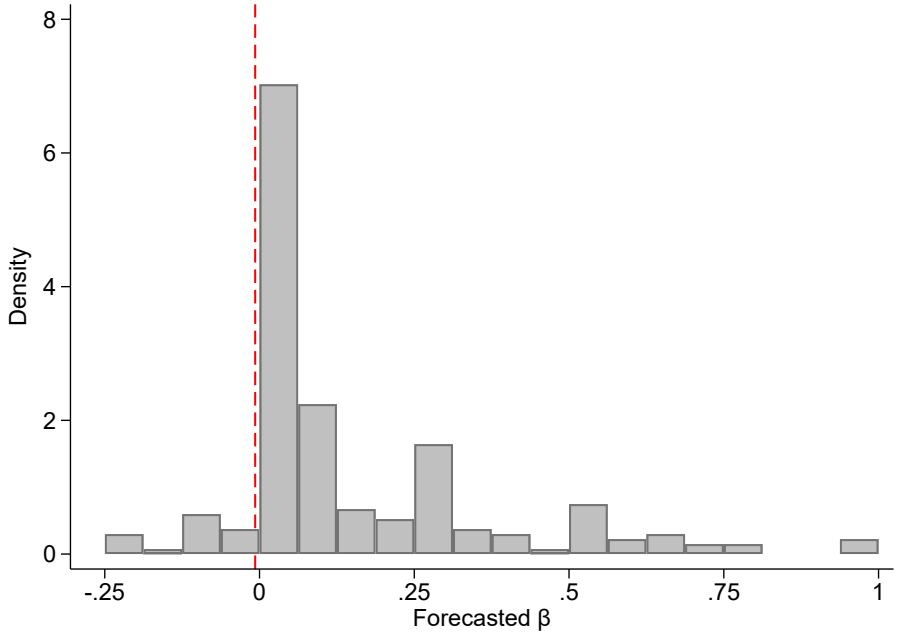


(c) Brodeur et al. (2020) - Top-ranked 25 journals



Notes: The first figure displays a histogram of test statistics for $z \in [0, 10]$, with bins of width 0.1, among our random sample of NBER working papers. As a comparison, the second figure plots the corresponding histogram of z-statistics from initial submissions to the *Journal of Human Resources* (JHR) from Brodeur et al. (2023), while the third figure plots the corresponding histogram from the top-ranked 25 economics journals published in 2015 and 2018 from Brodeur et al. (2020). For each histogram, we superimpose an Epanechnikov kernel density curve. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A2: Distribution of responses from the prediction survey (1% significance level)



Notes:The figure displays a histogram of survey respondents' predicted coefficients (forecasted β) at the 1% significance level. The vertical red dashed line marks the actual estimated coefficient from our analysis (Figure 1, Column 1). The mean survey prediction is 0.1455 (95% CI [0.1081, 0.1829]). To improve readability, values below -0.25 were censored and recoded as -0.25 before constructing the histograms. Bins are of width 0.0625.

Table A1: Summary statistics at the paper level

	No media attention		Received media attention	
	Mean	SD	Mean	SD
Number of test statistics	18.91	20.98	15.68	15.07
Month of NBER WP release	6.55	3.45	6.46	3.33
Year of NBER WP release	2012.94	4.36	2016.30	3.59
# of NBER WPs released in week	24.41	8.57	25.32	8.57
NBER program:				
-Labor Studies	0.52	0.50	0.56	0.50
-Economics of Education	0.18	0.38	0.24	0.43
-Economics of Health	0.37	0.48	0.35	0.48
-Environment and Energy Economics	0.15	0.35	0.16	0.36
# of authors	2.81	1.08	2.96	1.28
# of prior NBER WPs	34.98	32.16	30.51	26.84
Identification strategy:				
-Difference-in-differences	0.61	0.49	0.61	0.49
-Instrumental variables	0.18	0.38	0.17	0.37
-Randomize control trial	0.16	0.37	0.18	0.39
-Regression discontinuity	0.05	0.21	0.04	0.19
Additional outcomes:				
-Altmetric Attention Score	0.00	0.00	13.97	34.76
-Abstract views (in 1st month)	5.29	5.76	5.80	6.65
-Downloads (in 1st month)	8.83	12.75	12.80	18.86
ChatGPT score (0-10):				
-Media attention	2.08	1.13	2.40	1.04
-Interesting	7.41	0.53	7.53	0.53
-Influential	6.63	1.17	6.70	0.99
-Convincing	7.46	0.77	7.46	0.86
Observations	213		191	

Notes: This table presents summary statistics for our sample at the paper level, split by whether the paper received no media attention or some media attention (as collected by Altmetric). ChatGPT failed to produce an "interesting" score for five papers which did not receive media attention and one paper which received media attention, an "influential" score for one paper which did not receive media attention, and a "convincing" score for 42 papers which did not receive media attention and 28 papers which did receive media attention.

Table A2: Correlations with receiving media attention (at the paper level)

	(1)	(2)	(3)
-Abstract views (in 1st month)	-0.015	0.040	0.033
	(0.032)	(0.030)	(0.031)
-Downloads (in 1st month)	0.041*	0.033*	0.034*
	(0.022)	(0.020)	(0.020)
ChatGPT score (0-10):			
-Media attention	0.074***	0.074***	0.078***
	(0.025)	(0.022)	(0.023)
-Interesting	0.109**	0.101**	0.098**
	(0.051)	(0.047)	(0.047)
-Influential	0.013	0.017	0.017
	(0.025)	(0.022)	(0.022)
-Convincing	-0.028	-0.023	-0.026
	(0.034)	(0.030)	(0.031)
Observations	342	342	342
Time-varying controls		X	X
Paper controls			X

Notes: ChatGPT failed to produce an "interesting" score for five papers which did not receive media attention and one paper which received media attention, an "influential" score for one paper which did not receive media attention, and a "convincing" score for 42 papers which did not receive media attention and 28 papers which did receive media attention. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A3: Received media attention: Caliper tests, keeping ambiguous estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Received media attention	-0.010	-0.015	-0.008	-0.001	-0.005	-0.007
	(0.043)	(0.045)	(0.044)	(0.055)	(0.057)	(0.059)
Observations	1523	1523	1523	899	899	899
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Received media attention	-0.025	-0.052	-0.040	-0.023	-0.049	-0.038
	(0.041)	(0.044)	(0.041)	(0.053)	(0.058)	(0.055)
Observations	1503	1503	1503	932	932	932
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Received media attention	-0.000	-0.012	-0.013	-0.002	-0.026	-0.022
	(0.046)	(0.053)	(0.049)	(0.057)	(0.063)	(0.060)
Observations	1212	1212	1212	751	751	751
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A4: Altmetric attention score: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
ln(Altmetric score)	-0.002	-0.003	-0.002	-0.011	-0.013	-0.017
	(0.020)	(0.021)	(0.021)	(0.024)	(0.026)	(0.026)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
ln(Altmetric score)	-0.009	-0.020	-0.011	0.002	-0.007	0.002
	(0.018)	(0.019)	(0.018)	(0.022)	(0.025)	(0.024)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
ln(Altmetric score)	0.010	0.007	0.006	0.014	0.004	0.005
	(0.020)	(0.024)	(0.024)	(0.022)	(0.026)	(0.027)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A5: Abstract views: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
ln(# of abstract views)	-0.015	-0.014	-0.023	-0.046	-0.050	-0.055
	(0.025)	(0.026)	(0.025)	(0.036)	(0.036)	(0.035)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
ln(# of abstract views)	-0.004	0.004	-0.004	0.015	0.025	0.024
	(0.022)	(0.022)	(0.021)	(0.030)	(0.029)	(0.029)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
ln(# of abstract views)	-0.006	-0.004	0.007	-0.027	-0.020	-0.010
	(0.029)	(0.029)	(0.028)	(0.033)	(0.031)	(0.031)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A6: Paper downloads: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
ln(# of downloads)	-0.016	-0.016	-0.019	-0.038*	-0.039*	-0.043**
	(0.017)	(0.018)	(0.018)	(0.020)	(0.021)	(0.021)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
ln(# of downloads)	0.017	0.020	0.012	0.027	0.029	0.025
	(0.015)	(0.016)	(0.015)	(0.020)	(0.020)	(0.018)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
ln(# of downloads)	-0.034*	-0.034*	-0.031*	-0.039*	-0.036*	-0.032
	(0.018)	(0.018)	(0.018)	(0.022)	(0.021)	(0.022)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A7: ChatGPT media score: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Media score (0-10)	0.004	0.002	0.006	-0.001	-0.001	0.006
	(0.019)	(0.019)	(0.019)	(0.021)	(0.020)	(0.022)
Observations	1486	1486	1486	880	880	880
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Media score (0-10)	-0.005	-0.009	-0.011	0.001	-0.004	-0.001
	(0.016)	(0.016)	(0.016)	(0.020)	(0.019)	(0.019)
Observations	1462	1462	1462	909	909	909
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Media score (0-10)	-0.030	-0.029	-0.025	-0.040	-0.039	-0.036
	(0.027)	(0.027)	(0.026)	(0.027)	(0.026)	(0.025)
Observations	1165	1165	1165	717	717	717
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A8: ChatGPT interesting score: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Interesting score (0-10)	-0.079**	-0.079**	-0.067*	-0.116**	-0.116**	-0.112**
	(0.040)	(0.040)	(0.041)	(0.048)	(0.047)	(0.052)
Observations	1456	1456	1456	862	862	862
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Interesting score (0-10)	0.039	0.040	0.049	0.107**	0.106**	0.115**
	(0.034)	(0.034)	(0.031)	(0.044)	(0.043)	(0.045)
Observations	1427	1427	1427	888	888	888
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Interesting score (0-10)	-0.021	-0.020	-0.009	0.001	0.010	0.021
	(0.038)	(0.038)	(0.037)	(0.045)	(0.041)	(0.041)
Observations	1129	1129	1129	693	693	693
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A9: ChatGPT influential score: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Influential score (0-10)	-0.008	-0.008	-0.009	0.040	0.042	0.042
	(0.020)	(0.020)	(0.022)	(0.028)	(0.028)	(0.028)
Observations	1484	1484	1484	879	879	879
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Influential score (0-10)	-0.030*	-0.030*	-0.042**	-0.040*	-0.041*	-0.055**
	(0.017)	(0.017)	(0.016)	(0.023)	(0.024)	(0.022)
Observations	1457	1457	1457	905	905	905
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Influential score (0-10)	0.020	0.021	0.021	0.032	0.036	0.038
	(0.019)	(0.020)	(0.019)	(0.026)	(0.026)	(0.025)
Observations	1159	1159	1159	714	714	714
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

Table A10: ChatGPT convincing score: Caliper test, significance at the 10%, 5%, and 1% levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% significant						
Convincing score (0-10)	-0.024	-0.021	-0.012	-0.004	-0.005	0.008
	(0.031)	(0.031)	(0.030)	(0.037)	(0.035)	(0.035)
Observations	1254	1254	1254	735	735	735
z sample bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel B: 5% significant						
Convincing score (0-10)	-0.005	-0.003	-0.000	0.005	0.010	0.007
	(0.032)	(0.032)	(0.029)	(0.034)	(0.032)	(0.033)
Observations	1230	1230	1230	772	772	772
z sample bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$
Time-varying controls		X	X		X	X
Paper controls			X			X
Panel C: 1% significant						
Convincing score (0-10)	-0.002	-0.001	-0.003	0.034	0.039	0.027
	(0.029)	(0.029)	(0.029)	(0.041)	(0.040)	(0.038)
Observations	999	999	999	611	611	611
z sample bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.28, 2.88]$	$[2.28, 2.88]$	$[2.28, 2.88]$
Time-varying controls		X	X		X	X
Paper controls			X			X

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A/B/C is a dummy for whether the test statistic is significant at the 10/5/1 percent level, respectively. "Time-varying controls" include month dummies, year dummies, and the total number of NBER WPs that were released in the same week as the observed paper. "Paper controls" include dummies for the programs the NBER WP belongs to, the paper's identification strategy, and linear controls for number of authors on the paper and the total number of NBER WPs written by the coauthor team. In columns 1-3, we restrict the sample to $z \pm 0.50$. Columns 4-6 restrict the sample to $z \pm 0.30$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations. *, **, and *** denote significance at the 10%-, 5%-, and 1%-level, respectively.

B Prediction Survey

Why is this study being done?
We are researchers at the University of Pittsburgh conducting a prediction survey — we are investigating whether researchers can predict the correlation between media attention and statistical significance across economics working papers.

We are recruiting participants who have a PhD in economics or finance or are PhD students in economics or finance. In addition to providing your predictions, you will also complete a brief demographic questionnaire. We aim to understand your beliefs about the relationship between statistical significance and media attention. You will receive a base payment of $5 for completing the survey, plus you may earn an additional $15 based on the accuracy of your predictions of the actual relationship between media attention and statistical significance.

Taking part in this study is your choice.
Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analyzed.

What will happen if I decide to take part in this study?
You will be asked to predict the point estimates from three separate logit regressions that utilize data from a random sample of 1,000 National Bureau of Economic Research (NBER) working papers (WPs) that belong to at least one of the following NBER programs: Labor Studies, Economics of Education, Economics of Health, Environment and Energy Economics. From these 1,000 papers, we will focus strictly on papers identifying a causal effect with a clear primary identification strategy: Difference-in-Differences (DID), Instrumental Variables (IV), Randomized Controlled Trials or Experiments (RCT), or Regression Discontinuity Designs (RDD).

We will then code the "main" estimates from each of these papers (typically, point estimates and standard errors), focusing on what we determine to be the main results from the primary coefficient(s) of interest. Each test statistic will then be converted into their corresponding z-score. z-scores greater than 1.65 indicate statistical significance at the 10% level; z-scores greater than 1.96 indicate statistical significance at the 5% level; z-scores greater than 2.58 indicate statistical significance at the 1% level.

Each NBER WP is linked to data from Altmetric.com to determine whether the NBER WP received any media attention.

Our analyses will estimate three logit regressions, each following the equation:

$$\Pr\left(Significant_{pe} = 1\right) = \Phi(\alpha + \beta Media_{p})$$

where each observation is unique at the paper p estimate e level. We will estimate this equation three times, one for each statistical threshold: 10%, 5%, and 1%. Significant is an indicator for whether the estimate e on paper p is statistically significant at the corresponding threshold (10%, 5%, 1%). Media is an indicator for whether paper p received media attention. We will use the inverse of the number of estimates presented in the same article to weight observations (i.e. each paper will receive equal weight in the regression).

These regressions are sometimes referred to as Caliper tests, which test for differences in the number of test statistics in a narrow range above and below a considered statistical threshold. For our analyses, at each threshold, we will focus on test statistics that are within 0.5 of the corresponding critical value. That is,

- At 10%, our sample will include estimates with z-scores between 1.15 and 2.15
- At 5%, our sample will include estimates with z-scores between 1.46 and 2.46
- At 1%, our sample will include estimates with z-scores between 2.08 and 3.08

β will be estimated as marginal effects from the logit regressions. Thus, β can be interpreted as the change in probability (in percentage points) in the likelihood an estimate is statistically significant if the paper received media attention. β>0 (<0) indicates a positive (negative) relationship between media attention and statistical significance.

For example:

- β = 0 means that an estimate from a paper that received media attention is not any more or less likely to be statistically significant
- β = 0.05 means that an estimate from a paper that received media attention is 5 percentage points more likely to be statistically significant
- β = -0.05 means that an estimate from a paper that received media attention is 5 percentage points less likely to be statistically significant
- β = 0.20 means that an estimate from a paper that received media attention is 20 percentage points more likely to be statistically significant
- β = -0.20 means that an estimate from a paper that received media attention is 20 percentage points less likely to be statistically significant

Your task will be to predict our three estimates for β, i.e. the predicted change in the likelihood an estimate is statistically significant, at the 10%, 5%, and 1% levels if the estimate comes from a NBER WP that received media attention.

At the end of the survey we will also ask you some background questions and for your e-mail address. The e-mail address will only be used for paying the incentives and will be deleted from the data once the payments have been made so that no identifying information is saved.

What are the risks and benefits of taking part in this study?
We believe there is little risk to you for participating in this research project. You can stop taking the survey or you can withdraw from the project altogether at any point.

There is a small risk of breach of confidentiality. This is mitigated by not connecting your survey responses with your email. There are no direct benefits. If you choose not to participate, or if you do not complete the study, this will have no effect on your relationship with the University of Pittsburgh. Although every reasonable effort has been taken, confidentiality during Internet communication activities cannot be guaranteed and it is possible that additional information beyond that collected for research purposes may be captured and used by others not associated with this study.

For participating in this survey, you will receive compensation as described below. The results of this project may be published in a journal article.

Compensation
There will be a monetary bonus based on the accuracy of your predictions. The survey consists of three questions, each of which will solicit your beliefs for β and will be incentivized.

To be eligible for the bonus payout, you must:

- Complete all questions in this survey.
- Provide your email address for payment processing.

The size of your bonus payout will be determined by the accuracy of your predictions. More specifically, they will be determined by the below quadratic scoring rule, where we'll calculate the average of the squared errors across your three predictions for three different β's. The squared error is the squared difference between your prediction for β and the β we estimate.

$$\$15 - (\overline{Sq.Error} \times 1500)$$

Confidentiality and Privacy
We will not store any of your identifying information once the payments have been made.

Future Research Studies
After removing identifiers, we may share the data we collect in this study with other researchers doing future studies. If we share your data, there will not be any identifying information about you or other participants.

Questions
If you have any questions about this study, please email Lester Lusher at lesterlusher@pitt.edu. You can always revisit this page by moving backwards in the survey.

Please indicate, in the box below, that you are at least 18 years old, you have or are currently pursuing in PhD in economics of finance, have read and understand this consent form, and you agree to participate in this online research study

☑ I am at least 18 years old, I have or am currently pursuing a PhD in economics of finance, I have read and understand this consent form, and agree to participate in this online research study.

As a reminder:

Our sample consists of estimates collected from main results tables from a random sample of NBER working papers. Our analyses will estimate three logit regressions, each following the equation:

$$Pr(Significantpe = 1) = \Phi(\alpha + \beta Mediap)$$

where each observation is unique at the paper p estimate e level. We will estimate this equation three times, one for each statistical threshold: 10%, 5%, and 1%. Significant is an indicator for whether the estimate e on paper p is statistically significant at the corresponding threshold (10%, 5%, 1%). Media is an indicator for whether paper p received media attention. We will use the inverse of the number of estimates presented in the same article to weight observations (i.e. each paper will receive equal weight in the regression).

For our analyses, at each threshold (10%, 5%, 1%), we will focus on test statistics that are

within 0.5 of the corresponding critical value. That is,

- At 10%, our sample will include estimates with z-scores between 1.15 and 2.15
- At 5%, our sample will include estimates with z-scores between 1.46 and 2.46
- At 1%, our sample will include estimates with z-scores between 2.08 and 3.08

β will be estimated as marginal effects from the logit regression. β can be interpreted as the change in probability (in percentage points) in the likelihood an estimate is statistically significant if the paper received media attention.

For example:

- β = 0 means that an estimate from a paper that received media attention is not any more or less likely to be statistically significant
- β = 0.05 means that an estimate from a paper that received media attention is 5 percentage points more likely to be statistically significant
- β = -0.05 means that an estimate from a paper that received media attention is 5 percentage points less likely to be statistically significant
- β = 0.20 means that an estimate from a paper that received media attention is 20 percentage points more likely to be statistically significant
- β = -0.20 means that an estimate from a paper that received media attention is 20 percentage points less likely to be statistically significant

What is your prediction for the estimate of β for statistical significance at the 10% level? (Must be between -1 and 1)

What is your prediction for the estimate of β for statistical significance at the 5% level? (Must be between -1 and 1)

What is your prediction for the estimate of β for statistical significance at the 1% level? (Must be between -1 and 1)

What is your current position?

- PhD Student
- Postdoc
- Assistant Professor
- Associate Professor
- Full Professor
- Other position in academia
- Other position outside of academia

What is your field? (pick the option most closely related to your field or "Other" if different)

- Labor/Public/Health/Environment/Education
- Development/Political Economy
- Macroeconomics
- International Trade
- Other

In which region are you working?

- North America
- Central or South America
- Asia
- Africa
- Australia/New Zealand

What is your gender?

- Male
- Female
- Other/prefer not to say

Please provide an email address that we can use to contact you about your bonus compensation.