Brodeur, Abel; Cook, Nikolai; Heyes, Anthony

Working Paper

We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell us about p-Hacking and Publication Bias in Online Experiments

GLO Discussion Paper, No. 1157

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Brodeur, Abel; Cook, Nikolai; Heyes, Anthony (2022) : We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell us about p-Hacking and Publication Bias in Online Experiments, GLO Discussion Paper, No. 1157, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/263216

We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell us about p-Hacking and Publication Bias in Online Experiments*

Abel Brodeur
University of Ottawa

Nikolai Cook
Wilfrid Laurier University

Anthony Heyes
University of Birmingham

August 17, 2022

Abstract

Amazon's Mechanical Turk is a very widely-used tool in business and economics research, but how trustworthy are results from well-published studies that use it? Analyzing the universe of hypotheses tested on the platform and published in leading journals between 2010 and 2020 we find evidence of widespread p-hacking, publication bias and over-reliance on results from plausibly under-powered studies. Even ignoring questions arising from the characteristics and behaviors of study recruits, the conduct of the research community itself erodes substantially the credibility of these studies' conclusions. The extent of the problems vary across the business, economics, management and marketing research fields (with marketing especially afflicted). The problems are not getting better over time and are much more prevalent than in a comparison set of non-online experiments. We explore correlates of increased credibility.

KEYWORDS: online crowd-sourcing platforms - Amazon Mechanical Turk - p-hacking - publication bias - statistical power - research credibility

JEL CODES: B41, C13, C40, C90.

*Authors: Brodeur: University of Ottawa and IZA, abrodeur@uottawa.ca. Cook: Wilfrid Laurier University, ncook@wlu.ca. Heyes: University of Birmingham, aheyes@uottawa.ca. Abigail Marsh and Susan Price provided excellent research assistance. We are grateful to Anna Dreber and Tom Stanley for helpful comments. Heyes acknowledges financial support for this research from the Canada Research Chairs program.

1 Introduction

The use of online platforms as a source of research participants in the social sciences has increased rapidly in recent years, and dominant among these platforms is Amazon Mechanical Turk (MTurk).

Although various advantages are claimed for MTurk as a research tool including giving access to a large pool of prospective subjects other than undergraduate students and significantly diversifying the demographic profile of respondents (see, for instance, Paolacci et al. (2010)), frequently cited is a practical one, namely giving researchers the ability to build large samples at low cost. It is perhaps not surprising that the platform has become so popular a venue for faculty and graduate student research over the past decade.

However, in parallel with the growth in use of MTurk has come a growing suspicion in some research communities about the reliability of results from studies using it. For example, after noting a 2117 percent increase in MTurk use in management research between 2012 and 2019, Aguinis et al. (2021) observed, in a review of the platform commissioned by the editorial board of the highly-rated Journal of Management, that "(A)mong scholars, though, there is a mixture of excitement about the practical and logistical benefits of MTurk and skepticism about the validity of the data". This skepticism is driven, it is claimed, by a variety of concerns about the behavior of MTurk respondents, and there is a literature probing these issues.[1]

We do not contribute to that literature - indeed we have nothing at all to say about the pros and cons of MTurk-ers as subjects. Rather the focus of this paper is quite different, providing the first systematic investigation of the statistical practices of the research community itself when using MTurk, and the extent to which those practices render MTurk-based empirical results untrustworthy. The three practices that we study are those that have become focal in recent assessments of research credibility elsewhere, namely (1) p-hacking, (2) publication bias (or selective publication) and, (3) the presentation of results from plausibly under-powered samples.

For our analysis, we analyze the universe of hypothesis tests from MTurk papers published in all journals categorized as either 4 or 4* in the 2018 edition of the

[1]For example, subjects recruited from the platform might pay insufficient attention to tasks because of the low rate of pay, might share information with other participants via online community tools, might be deliberately deceptive in responses, claim to be in one place but actually be working from another via a VPN, participate in a study multiple times using aliases, etc. Hauser et al. (2019) detail similar concerns. Both sets of authors go on to provide a set of recommendations about how methods can be improved to account for such considerations. Despite this, it is worth noting that several studies have shown, by running identical experiments on multiple subject pools, that results derived from MTurk samples do not look very different with those from samples from more conventional and expensive sources (for examples see Snowberg and Yariv (2021) and Horton et al. (2011)) and over time (Johnson and Ryan 2020).

Association of Business School's Academic Journal Guide between 2010 and 2020, around 23,000 in total. The Guide has broad coverage of business research and related fields such as economics, finance, management, marketing and sector studies (e.g., tourism, sociology) and is widely-used in the evaluation and assessment of researchers.

We first investigate the extent of p-hacking and publication bias in the body of research. Publication bias occurs if the likelihood that a hypothesis test is published depends upon the result, for example if a statistically significant treatment effect is more likely to be published than a null result. This could reflect the choices researchers make in deciding what to write up, and what to put in the "file drawer", and/or the processes by which journals select what to publish. The term p-hacking refers to research choices being made in such a way as to artificially inflate statistical significance.[2] Both phenomena lead to an artificial shortage of statistically insignificant or 'null' results in the published corpus. Anecdotally, many empirical researchers will recognize the attraction of statistical significance, consistent with the evidence from a randomized experiment conducted by Chopra et al. (2022, page 1) who finds that "(s)tudies with null results are perceived to be less publishable, of lower quality, less important, and less precisely estimated than studies with statistically significant results, even when holding constant all other study features."

It is increasingly acknowledged that p-hacking is an insidious problem. However, while it is difficult or impossible to detect or quantify in any individual study, it is possible to characterize the prevalence of the problem at meta-level by comparing the pattern of statistical significance observed in a body of research with what would be expected absent such behavior. A number of techniques have been developed to test for and quantify p-hacking and publication bias, either jointly or separately. None of the techniques are definitive, and each embeds particular assumptions, so we regard as a strength of our approach that we apply a wide set of them.

We plot the distribution of test statistics from MTurk articles and find them to exhibit patterns consistent with the presence of considerable p-hacking and publication bias. In particular, the distribution exhibits a pronounced global and local maximum around a z-statistic value of 1.96, corresponding to the inveterate threshold required for statistical significance at the 5% level, or a p-value of 0.05. This maximum is coupled with a shift of mass away from the marginally statistically insignificant interval, indicative of p-hacking. This pattern of test statistics is per-

[2]Approaches to p-hacking can take various forms, including in the way data is cleaned, variables defined, and specifications chosen. Concern about the propensity for p-hacking to undermine research credibility in recent years has led to a proliferation of robustness demonstrations in empirical work (whereby the author endeavours to show that qualitatively similar results would have followed from alternative yet plausible modeling choices) and increased interest in pre-registration (whereby the author commits to statistical modeling choices in advance of data collection).

sistent over time and roughly as present in papers published in elite (rank 4*) and top (rank 4) journals.

We use the method proposed by Brodeur et al. (2016) to estimate the extent of 'misallocated' test results, the proportion of purportedly significant results for which statistical significance is wrongly claimed, finding substantial variation in the prevalence of such results between research fields - being most frequent in marketing and least frequent in economics and finance.[3] Applying the method developed by Andrews and Kasy (2019), we also evidence severe publication bias in our sample. Other things equal a z-statistic greater than 1.96 is 4.61 times more likely to be published than a statistically insignificant result. We corroborate these findings using traditional caliper tests (Gerber and Malhotra 2008a), which look for a step in the frequency of tests statistics on either side of arbitrary significance thresholds, and the sophisticated battery of tests proposed by Elliott et al. (2022).

Finally, we examine sample size in the studies in our sample. The power of a statistical test is the probability that is detects an effect (rejecting the null hypothesis of no effect) should a true effect exist. The appropriate choice of sample size, and therefore level of power, is a central element of experimental research design. Of specific concern to us here is that low statistical power implies a high rate of false positives, by which effects are seemingly detected where none exist (Ioannidis 2005). Naturally such spurious results are unlikely to be reproducible, particularly if followed by a replication exercise with greater power. A literature populated by under-powered (small sample) studies is likely to feature disproportionately many 'surprising' results, face replication issues, and correspondingly face challenges to it's credibility.

In this context, we highlight two features of MTurk studies. First, many (most) MTurk studies use small samples. The median number of subjects in an experiment in our sample is 249.[4] A sceptical reader might find this surprising in light of how quickly and cheaply sample size can be built on the MTurk platform - indeed that is probably it's most widely-claimed benefit.[5] Based on a manual reading of each study we harvest the marginal cost of an additional subject or data point. In our sample of studies, the average cost of an additional data point is 1.30 USD, and in about 70 percent of cases it is less than 1 USD. So on what basis (unlikely to be

[3] Brodeur et al. (2020) document the extent of p-hacking in a sample of tests from 25 leading economics journals, which compares favorably (i.e., exhibits less p-hacking) than what has been documented in political science and sociology (Gerber and Malhotra (2008a); Gerber and Malhotra (2008b))

[4] With this sample size, a two-tailed comparison of means is only powered at 80% with a 5% confidence threshold if the underlying effect size is not less than 0.175 standard deviations.

[5] In observational settings sample size (and therefore statistical power) is often limited by the study setting. In conventional experimental settings (laboratory experiments, field-based randomized control trials) sample size is often limited by financial or other practical considerations. Neither of these constraints are likely to apply here.

cost) were such small sample sizes selected? This leads to our second observation; in most MTurk studies there is no justification given as to how a particular sample size was chosen.

The frequency of apparently small samples in MTurk studies leads to concerns about statistical power. Few of the studies in our sample include a formal power analysis, and there is no defensible way to impute the statistical power of a hypothesis ex post.[6] We explore systematically - both across the whole sample and within sub-samples - the relationship between sample size, the statistical significance of test results, cost per data point, and contextual data on whether a study provides a rationale for how sample size was determined.

The results here contribute to a literature discussing the credibility of research conducted on MTurk and other crowdsourcing platforms (Arechar et al. (2017); Berinsky et al. (2012); Coppock (2019); Buhrmester et al. (2011); Goodman et al. (2013); Horton et al. (2011); Johnson and Ryan (2020); Lee et al. (2018); Paolacci et al. (2010); Snowberg and Yariv (2021)).[7] While the existing literature focuses on the advantages and disadvantages of the platform itself, we instead document issues in the way MTurk experiments are collectively conducted.

Taken as a whole, the patterns that we identify in the data challenge the credibility of MTurk-based studies published in highly-rated journals across business and related research fields. However, this is not for the oft-cited reasons concerning the validity of responses provided by MTurk subjects (about which this study has nothing to say) but rather the dubious statistical practices of researchers.

More widely the analysis complements the growing literature on research credibility, in particular studies documenting the extent of p-hacking and publication bias (e.g., Andrews and Kasy (2019); Brodeur et al. (2016); Brodeur et al. (2020); Bruns et al. (2019); Camerer et al. (2019); DellaVigna and Linos (2022); Doucouliagos and Stanley (2013); Elliott et al. (2022); Furukawa (2019); Gerber and Malhotra (2008a); Havránek (2015); Vivalt (2019)),[8] and detailing the use of statistical power in economics and other disciplines (e.g., Ioannidis et al. (2017); Zhang and Ortmann (2013); Ziliak and McCloskey (2004)).

[6] Ioannidis et al. (2017) develop a method to estimate power of a test ex post, but only when the test involved can be conditioned on an effect size from a meta-analysis of studies estimating a common effect

[7] Other relevant studies discussing the credibility, representativeness and generalizability of experiments in economics include Camerer et al. (2016), Falk et al. (2013), Gillen et al. (2019), Guala and Mittone (2005), Harrison et al. (2009) and Levitt and List (2007), among others.

[8] See Christensen and Miguel (2018), Stanley and Doucouliagos (2014) and Swanson et al. (2020) for literature reviews and discussions of recent advancements in research transparency.

2 Data

Our first task is to construct a large sample of well-published studies that use MTurk participants across a broad set of research fields common in business school settings. As a starting point we adopt as a journal ranking the 2018 edition of the Academic Journal Guide published by the Chartered Association of Business Schools, a widely-used ranking which provides a unified rating approach across 22 research fields. An attractive feature of the Guide is that the ratings are based upon editorial, expert, and peer review, in addition to more standard (and perhaps problematic) journal impact and citation metrics.

We restrict attention to journals assigned the Guide's top ratings of 4* and 4, which across all fields represent the top 8.0% of journals - the outlets deemed by the Guide to be publishing the highest quality research.[9]

We use Google Scholar to search the entire body of every article published in every one of these rank 4 and 4* journals for the keyword "Mechanical Turk". We then manually inspect the text of each article featuring that keyword and keep only those that use a participant sample derived from MTurk and report resulting test statistics within either the text of the manuscript or in tables. As such we remove literature reviews, surveys, articles which include the keyword only in the bibliography, articles without numerical results, and so on. We also remove articles where MTurk workers are asked to categorize items (such as photographs) or are otherwise used as research assistants rather than as research subjects.

While we retain articles that contain both MTurk and non-MTurk analyses, we focus only on the former.[10] For example, if an article has two studies – one conducted on MTurk and another on an undergraduate sample – we collect only those test statistics that relate to the MTurk sample.

Our final sample includes 1,031 articles from 55 journals (see Appendix Table A1 for the complete journal list). Appendix Figure A1 plots the number of MTurk studies that we collect by year of publication, and evidences the rapid rise in use of the platform. Almost all articles contain multiple experiments and, in many cases, more than one MTurk sample.

We collect test statistics from both article text and tables (about half of test statistics are presented in tables). We collect only those test statistics that relate to main results, i.e. coefficients of interest, excluding, for example, control variables, constant terms, and test statistics derived from treatment-balance and placebo tests.

[9] Since our interest is in research 'core' to business fields we omit journals listed in the Guide's psychology section. While it is not unusual for business school academics to publish in psychology journals, a large majority of articles in these journals fall outside the scope of normal business research, and this approach circumvented the need for us to classify studies within-journal.

[10] About 58% of articles in our sample rely solely on MTurk, the rest also report other empirical analyses (e.g., from a laboratory experiment).

With these exclusions, we collect 22,989 test statistics.

Authors report test statistics and associated statistical significance in various ways. In our sample, about 28%, 28%, 18%, 7% and 1% report p-values, t-statistics and F-statistics, confidence intervals, and Wald or Chi-Squared Tests, respectively.[11] The remaining tests are reported as coefficients and standard errors. We follow Brodeur et al. (2016) in transforming all tests into equivalent z-statistics. For tests reported using coefficient and standard errors, we take their ratio and treat them as if they follow an asymptotically standard normal distribution under the null hypothesis.[12]

We collect additional contextual data. For each article, we record: the journal, year of publication, the number of authors, the year in which the MTurk data was collected, whether the article also presents results from non-MTurk samples, whether the paper provides a discussion of statistical power or a justification of sample size; mean remuneration paid per MTurk subject; the time required to perform the task(s) required by the researcher. For each test statistic, we record where it is reported in the paper (that is, whether in a table or only within the text) and in what form (e.g., t-statistic, p-value).

Finally, we deal with some technical complications noted in Brodeur et al. (2016). These include re-weighting articles with relatively more/less test statistics per article, and adjusting for the rounding by authors of statistics. We show in a series of robustness exercises that such adjustments have no meaningful impact on our overall conclusions.

Table 1 reports summary statistics for the whole sample. The unit of observation is a test statistic. 79% of the test statistics are published in 4* journals. 70% of articles include non-MTurk results (e.g., an additional lab experiment). The mean number of authors is 2.7. The mean number of participants for a MTurk experiment is 450 though the range is very large, from 20 to 15,166. We report throughout the number of participants per MTurk experiment rather than the number of observations or the total number of participants per article.[13]

Business research encompasses a wide range of disciplinary traditions, and in a number of places in the analysis it will be useful to dis-aggregate the whole sample into narrower research areas. The Guide itself reports journals under a number of

[11] For F-statistics, we also collected the numerator and denominator degrees of freedom in order to compute an associated z-statistic.

[12] Two qualifications are worth noting here. First, we also search all articles for the keywords "one sided" and "one-sided" and find few contain either term. If neither is mentioned (the large majority of cases) we assume all tests are two-sided. Second, a non-small proportion of p-values are coarsely reported (e.g., $p < 0.05$). We omit these p-values from our main analysis.

[13] Note that we use the terminology number of participants to refer to the number of participants used by the authors in their statistical analyses, not the number collected. Typical reasons given for omitting participants from the analysis include failing attention checks and demographic criteria.

different headings, and the top panel in Table 2 provides certain summary statistics for those, under the heading "ABS-defined Fields". While the whole sample is too coarse for our purposes, however, these categorizations prove too granular, delivering many sparsely-populated sub-samples. We therefore define four consolidated fields; 'Economics Finance', 'Marketing', 'Management Accounting' and 'Sector Social Studies'. Summary statistics for these broader fields are reported in the lower panel in Table 2, and the composition of those fields in Appendix Table A1.

3 Distribution of Test Statistics

We first visually investigate the distribution of test statistics for the whole sample of MTurk articles and then various sub-samples.

Before presenting, it is worth thinking about what we might expect such a distribution to look like absent publication bias or p-hacking. A number of authors offer expectations grounded in statistical theory - for example Elliott et al. (2022) show that for any distribution of true effects the p-curve should be non-increasing and continuous under the null of no p-hacking, under a broad set of circumstances. Another example is the logic underpinning the caliper test - absent publication bias we should not expect to see any particular 'clumping' of significance tests just above arbitrary thresholds set for statistical significance.

Readers might also find it useful to have some empirical benchmarks in mind; three of these are presented in Figure 1. The first panel, using data from Brodeur et al. (2016), plots the z-statistics relating to hypothesis tests in all laboratory experiment articles published in three of the most prestigious economics journals in 2005 through 2011. The middle panel does the same for field-based experiments and randomized control trials published in those journals for the same period. The last panel, using data from Brodeur et al. (2020), does the same but for randomized control trials published in the top 25 highest rated economics journals in 2015 and 2018. Each of these papers argue that these depicted distributions of test statistics are consistent with a (relative to their target literature) low propensity of p-hacking - in other words, those sub-samples whose results are derived from experiments tend to be the best performing.[14]

3.1 Test Statistics for Full Sample

Figure 2 plots the raw distribution of z-statistics for our full sample for (truncated at $z = 10$). Histogram bins are 0.1 wide. An Epanechnikov kernel, which smooths

[14]Other strands of empirical work common in these economics journals are shown to be much more problematic, in particular those employing instrumental variables. Brodeur et al. (2020) discuss why randomized control trials appear largely unaffected by the p-hacking problem.

the distribution, is superimposed, as are 95% confidence intervals (the latter, by virtue of our large sample size, are difficult to see).

The shape of the distribution is visually striking. Unlike any of the comparison distributions from Figure 1, the distribution in Figure 2 features a large peak just above $z = 2$. The distribution exhibits a local minimum around 1.75, consistent with a dearth of observations just below the 5% statistical significance threshold.[15]

Two aspects of Figure 2 are worth noting.

First, studies in the MTurk sample collected for the current paper are much less likely to feature null results than the non-MTurk experiments that underpin Figure 1. This is reflected in the mass in the left-hand side of the distribution. For the MTurk sample, this mass begins at 0.2 on the vertical axis and falls monotonically (as expected from no publication bias or p-hacking) until the large peak (expected in the presence of publication bias or p-hacking). In comparison, the mass begins at 0.3 and 0.4 for the comparison distributions - meaning the presence of much more statistically *in*significant published results.

Second, as noted, the very sharp peak of results with z-statistics in a tight range just above 1.96, and an apparent lack of results in the range just below that value - the valley to the left of the peak - are typical symptoms of p-hacking, consistent with a subset of marginally insignificant results being 'nudged' over the threshold. The comparison distributions do not have similarly sized peaks (note the global maximum is contained in the null mass), and are closer to the monotonically decreasing distributions implied by the theoretical predictions of Elliott et al. (2022) under the null of no p-hacking or publication bias.

3.2 Test Statistics by Field

Examining the whole sample distribution gives a "high level" view of statistical practices with respect to MTurk studies in journals covered by the Guide. At the same time, this aggregation may mask important differences in practices between research fields. We investigate this here by looking at sub-samples.

As already noted, the Guide categorises journals more finely than is appropriate for our comparison purposes here. To streamline our presentation, and to ensure each of the sub-samples are well enough populated to allow meaningful conclusions to be drawn, we consolidate the 22 categories in the Guide into four broader fields. We report in Table 2 the important differences between the conduct of MTurk studies across the four research fields. Plausibly important in terms of outcomes of statistical testing is that the mean number of participants used in an experiment is

[15] In additional exercises we verify that neither de-rounding nor applying article weights substantially disturbs the distributions (Appendix Figures A2 and A3). If anything, article-weights make the peaks more pronounced, so our resulting focus on the un-adjusted results is conservative.

much larger in 'Economics and Finance' (mean = 1,908) and 'Management' (517) than in 'Marketing' (382) and 'Sector and Social Studies' (390). With this concurrent difference in mind Figure 3 plots the distributions of z-statistics for each of the sub-fields. Although the histograms are less smooth than in Figure 1 (reflecting the smaller sample size in each sub-field) it is immediately apparent that the distributions in the four panels look quite different from one another.

The kernel density plotted for the Economics & Finance distribution is downward-sloping over its whole support, has substantial mass at the left-hand end (corresponding to consistently publishing null results), and little if any discernible excess mass immediately to the right of 1.96. The Management & Accounting kernel features a peak in the range above the 1.96 threshold and a corresponding valley immediately below, though these are not especially pronounced. The Marketing panel reveals a pattern of statistical significance quite different in character to the other fields. There is a very low mass of published non-statistically significant results and a very sharp peak of z-statistics at values immediately above 1.96. The final panel displays test statistics from Sector & Social studies which has a shape similar in character to Marketing, but much less pronounced. The data for this panel is drawn from sector studies (such as highly ranked service industry and tourism journals) and sociology.

Overall the dis-aggregation in Figure 3 suggests the problems of publication bias and p-hacking are not inherent to the use of MTurk, but rather differences in how researchers in different fields use the platform.

3.3 Test Statistics by Primacy of MTurk

The role played by MTurk samples varies between articles.[16] Around 40 percent of the sample of z-statistics are collected from the 30 percent of articles that rely exclusively on MTurk as the article's source of data. The remaining 60 percent from articles that also report results from experiments conducted on non-MTurk subjects (for example student samples or consumer panels).

In Figure 4 we separately plot the distribution of test statistics for part-MTurk articles (left-hand panel) and all-MTurk articles (right-hand panel). Qualitatively the distributions are quite similar, exhibiting a non-monotonic peak at 1.96 after a valley between 1.5 and 1.96. However, those characteristics (the valley and peak) are much more pronounced in the part-MTurk.

Why might such a pattern emerge? One conjecture would be that a researcher holding results from multiple experiments that use different subject pools may find

[16]Multiple experiments are common in our sample. For example, a later experiment could examine a refinement of the previous experiment's protocol, extend its hypothesis, or use a different participant source as a robustness exercise.

it easier to exclude insignificant results than would a researcher holding just MTurk results.[17] Another could reflect the peer-review stage itself: if an experiment already offers the reviewer 'conventional' lab experimental evidence, an additional highly statistically significant MTurk result may act as a bolster and perhaps not face the same scrutiny that a 'main' result otherwise would. We are cautious not to over-interpret our descriptive findings.

3.4 Test Statistics by Reporting Method

There is also variation across our sample in where in the article test statistics are reported. This might reflect differences in norms between fields, style requirements of journals, or discretionary choices made by individual researchers.

Figure 5 plots the distribution of z-statistics where the statistic is reported in the manuscript text (left) versus in a dedicated results-table (right panel). The distributions are similar yet still slightly different in character. For the text-reported results, the distribution reveals a larger dearth of insignificant results and a larger peak, the characteristic symptomatic of a p-hacked body of research. The distribution for results reported in tables features a greater prevalence of null results and a smaller peak, in relative terms.

How should we interpret this? First it is important to notice that this is an association, and the reader should be wary of drawing causal conclusions. One conjecture would be that the discipline of a journal or an area of research in which results are expected to be tabulated might make it harder to exclude null outcomes. An alternative would be that this difference is purely artefactual, possibly picking up differences in reporting conventions between fields which may have differently p-hacked bodies of research. For instance the percentage of statistics derived from tables in the 'Economics & Finance' subsample is much higher than that in 'Marketing', and we noted from Figure 1 the greater prevalence of p-hacking in the latter than the former. To explore this further we plot, in Appendix Figure A4, the distributions of table-reported versus text-reported test statistics but *only* in the 'Marketing' subsample. It is interesting to observe that even *within* the field - across studies that plausibly share disciplinary norms with respect to reporting conventions - the overall pattern of p-hacking continues to appear more pronounced in articles that report test results in the text body rather than in a dedicated table.

[17] Of note, we did not collect test statistics from MTurk experiments that were labeled explicitly as a pilot by the authors.

3.5 Test Statistics by Journal Rank

Recall that in constructing our sample of articles we collected both journals rated as 4* and 4 by the Guide. The 4-rated journals are high quality, but the 4* designation is the highest and awarded to a relatively small number of the most prominent and highly selective journals in each area (for example there are six in economics and five in marketing). Does the higher selectivity of the 4* journals exacerbate the problems that we are investigating here, or perhaps does a more rigorous standard of review allow those journals to filter out potentially spurious results?

In Figure 6 we present the distribution of statistics from the 4* and 4 journals separately. The two panels reveal qualitatively similar patterns, each with an apparent excess of statistics in the range just above 1.96, and a dearth in the range just below. The 'spike' is somewhat sharper in the 4* sub-sample, though we are again wary to over-interpret that difference as causal.

3.6 Test Statistics Over Time

Amazon launched the Mechanical Turk platform late in 2005 and its use in academic research in the social sciences expanded rapidly over our study period, reflected in Appendix Figure A1. While our main results point to substantial p-hacking in the overall sample, in this section we consider whether things are improving over time. There has been an increase in awareness over time both of the challenges relating to research credibility, p-hacking and publication bias which could plausibly have influenced behaviors of researchers as well as reviewers and editors. Furthermore research practices might have evolved as individual and collective experience with the new platform accumulated.

To probe this we divide the full set of articles into those published in the first half of our study window (2010 through 2015) and those in the second half (2016 through 2021). While this divides the sample in half chronologically, the increase in use of the platform means that the second subset contains more z-statistics.

Figure 7 presents the distributions of z-statistics for each of the time periods, the left-hand for the earlier articles and the right-hand for the later. There is little discernible difference between the two panels, offering little evidence that evolution of research and editorial practices concerning p-hacking and publication bias - at least with respect to MTurk-based articles - are changing through time. This mirrors the pessimistic finding in Brodeur et al. (2020) who study the application of causal inference methods in articles published in the top 25 economics journals for 2015 and 2018.

3.7 Economists Publishing in Non-Economics Journals

The analysis above has pointed to important differences in observed patterns of statistical significance between journals of different types. An intriguing supplemental question to that is whether the research behaviors suggested by the patterns are driven by the disciplinary roots of researchers, or by the norms of the field in which they find themselves working. Is it inherent to the way that economists work, for example, that they tend to generate a much higher proportion of null results than do marketing researchers?

To investigate this further we look at hypothesis tests from studies published in non-economics fields but authored by economists.

For this purpose we label an author as an economist if their name appears in at least one of two RePEc lists: 'Authors in Cognitive Behavioural Economics' and 'Authors in Experimental Economics'.[18] Joining these and removing duplicate names provides a list of 2,241 economists. We then need to compare this list of economists with the authors of the articles that make up our test statistic sample.[19]

The bottom line in Table 2 report summary statistics for studies authored by economists but published in non-economics journals. The number of test statistics by economist-authored papers in non-economics journals is 1,186.

In Appendix Figure A5 we plot test statistics for three separate instances. In the first panel, we present the test-statistic distribution produced by economics authors published in economics journals. While we note the now larger width of the 95% confidence interval bounding the kernel density estimates, the shape seems to be more or less monotonically decreasing. In the middle panel, test statistics from economics authors published *not* in an economics journals. Compared to the first panel it is apparent that there are now a reduced prevalence of null results (where $z = 0$ or nearby) and a change from the monotonically decreasing distribution to one with a discernible bump right after $z = 2$.[20]

In connecting these visual results to some of the underlying aspects of experiment design we note elsewhere in the paper, economists in economics are the most likely to include a power analysis (39%) and perform equal to the best non-economics authors (management) when publishing in non-economics journals by including a power analysis 23% of the time. A similar result holds true for the cost of remuneration -

[18]To be included in either of these lists requires the author to have an account on RePEc. RePEc is a large-scale set of tools for dissemination of economics research. It's bibliographic database includes around 3.8 million research items from 64,000 registered authors).

[19]To account for potential differences in names (e.g., John A Smith and John Smith), we use a Jaccard similarity score when comparing the two lists - any name from the author list that has a similarity score of 0.8 and above for a name from the list of economists we flag as an economist (a score of approximately 0.8 would be assigned to the previous example).

[20]For completeness, the right panel displays the remainder of test statistics - the test statistic distribution of non-economists (as defined here) published in any journal.

economists spend more per article regardless of where they publish.

The results here should only be taken as suggestive for a number of reasons, not least that the sample sizes are relatively small. However they suggest field norms play an important role in the mapping from research behaviors to patterns of statistical significance, in addition to the potential importance of the disciplinary 'home' of the researcher.

4 Formal Tests and Metrics for p-Hacking and Publication Bias

At present, there is no single definitive method to test for, or metric to quantify, the presence or extent of p-hacking and publication bias in a corpus of research. In light of this we apply several considered to be the state-of-the art. This is pragmatic, and we believe renders our qualitative assessments not overly sensitive to the assumptions underlying, or criticisms of, any particular approach.

4.1 Testing for p-Hacking Using the Tests Proposed by Elliott et al. (2022) (Econometrica)

This subsection applies an identical series of analyses to that introduced in Elliott et al. (2022) now applied to our current sample. Elliott et al. (2022) derive testable restrictions for the expected distribution of test statistics absent p-hacking which resulted in more powerful tests (than previously used in the literature) against their null hypothesis of no p-hacking. Of note, their tests are joint tests for p-hacking and publication bias in the presence of publication bias.

While our analysis has focused on the z-curve, Elliott et al. (2022) examine its mechanical counterpart, the p-curve (a histogram of p-values). For the sake of ease of translation, this subsection examines the p-curve which would be derived from our sample(s).

Figure 8 illustrates our p-curve which is truncated above p=0.15 (z=1.440). As a benchmark, we compare our results to those found when Elliott et al. (2022) applied their analysis to the data provided by Brodeur et al. (2016); our sample is arguably of similar statistical power, considering 1,181 test statistics contained in [0.04,0.05] compared to the Elliott et al. (2022) application with 1,175 in the same interval. Using Brodeur et al. (2016)'s data (who examined top economics journal articles) a visible discontinuity existed between the [0.040,0.045] bin and the [0.045,0.050] bin, with the latter (the more right of the two on a p-curve) larger than the *more* statistically significant former bin. In our setting, the relative magnitude of this discontinuity is larger.

Figure 8 contains both types of test proposed by Elliott et al. (2022), namely those based on the expected non-increasingness of the p-curve and those testing for

discontinuities. As a preview, all tests except one (Fisher's) reject the null of no p-hacking.[21]

We briefly discuss each test in what follows, starting with those based on the non-increasingness of the p-curve. The interested reader is directed to the paper for technical details of the methods.[22]

First, for our sample the binomial test rejects the null hypothesis that the p-curve is non-increasing.[23] Second (as mentioned above) Fisher's Test does not reject the null hypothesis. Third, CS1 (an application of the conditional chi-squared test introduced in Cox and Shi (2022)) rejects the null hypothesis. Fourth, CS2B (a histogram based test, this time for 2-monotonicity and bounds on the p-curve and its first two derivatives) also rejects the null hypothesis (see Elliott et al. (2022) for details on this powerful test). Fifth, (directly following the null hypothesis that the p-curve is non-increasing is that the CDF of p-values is concave) we apply a test based on the least concave majorant (LCM), which also rejects the null. Sixth, Figure 8 also provides a discontinuity test (an application of the density discontinuity test from Cattaneo et al. (2020)) which rejects the null hypothesis of no discontinuity in the p-curve at p=0.05.

Finally, we follow Elliott et al. (2022) by selecting a random subset of our data set in order to deal with possible within-article p-value dependence. Specifically, we randomly select one test statistic per article in our sample. This reduces our sample of test statistics by a factor of around 23. Despite this, the binomial test is only marginally insignificant at p=0.126, while the discontinuity and two monotonicity tests remain statistically significant at conventional levels.[24]

Appendix Table A2 reports the same analyses as in Figure 8, but separately by field. Using the same battery of tests there is no detectable p-hacking in the Economics & Finance subsample, and we also note the lack of a visually discernible discontinuity on either side of p=0.05 (not pictured).[25] For Management, we find a statistically significant result for the presence of a discontinuity, and statistically significant results for the two histogram-based tests of non-increasingness, indicating that p-hacking is likely present. For Marketing, all tests (with the consistent

[21] In Elliott et al. (2022), Fisher's test never rejects the null of no p-hacking, in economics or in different disciplines examined by Head et al. (2015).

[22] The p-curve should be non-increasing under rather general conditions following Theorem 1 in Elliott et al. (2022).

[23] For the binomial test, we follow Elliott et al. (2022) and split [0.04,0.05] into two subintervals [0.040,0.045] and (0.045,0.050]. Under the null of no p-hacking, the fraction of p-values in (0.045,0.050] should be smaller than or equal to one half.

[24] This is in marked contrast to when applying these tests to the sample in Brodeur et al. (2016), where Elliott et al. (2022) found that no tests for p-hacking or publication bias could reject the null in a random sub-sample, concluding that these insignificant results were plausibly due to power issues in applying the tests to small samples.

[25] Due to such a small number of tests in Economics & Finance in the [0.04,0.05] interval however, both CS1 and C2SB (the histogram based tests) were unable to be computed.

exception of Fisher's) reject their null hypothesis of no p-hacking. For the remaining field of Sector and Social Studies, we note only a statistically significant discontinuity test (p=0.016), and the powerful histogram test (that of 2-monotonicity) detect publication bias or p-hacking.

In summary, applying the test battery developed by Elliott et al. (2022) serves to confirm formally our conclusions drawn from our earlier visual inspection.

4.2 Estimating Excess Test Statistics using Method of Brodeur et al. (2020) (American Economic Review)

This subsection describes and applies the excess test statistics methodology found in Brodeur et al. (2020).[26] The underlying intention is to compare the observed distribution of test statistics to a bespoke or context-specific counterfactual distribution which we would expect absent publication bias or p-hacking. To construct the counterfactual, a non-central t-distribution is calibrated to resemble the observed distribution for the range z>5 (where p-hacking is unlikely to a meaningful problem). The additional assumptions allow us move beyond rejecting a null hypothesis of no p-hacking to estimating the *amount* of distortion.

Figure 9 presents the observed distribution of test statistics from our MTurk sample as a solid line (and corresponds exactly to the kernel density in 2). We then calculate a counterfactual non-central t-distribution whose tail matches the tail of the observed distribution (in this case the best fitting counterfactual uses 2 degrees of freedom and a non-centrality parameter of 1.6) and plot the distribution as the dashed line. In the figure, we have included the mass difference between the observed and counterfactual distribution by statistical significance region in rotated text below the horizontal axis. For example, in the statistically insignificant region of $0 < z < 1.645$, the observed distribution is "missing" 16.1% of the total mass when compared to the counterfactual. These "missing" test statistics can almost wholly be found above the two-star statistical significance threshold (where there is 9.5% more total mass than expected) and above the three-star threshold (where there is 5% more total mass than expected). In the marginally significant one-star interval (which allows a study to only claim significance at the 10 percent level) and in the very significant interval (where a study would begin to claim p less than 0.00000058), there is no appreciable mass difference between the observed and counterfactual distributions. Applying this method by field confirms the results we present elsewhere and are presented in Appendix Table A3. Economics & Finance are 'missing' less than 1% of their insignificant test statistics. In contrast, Manage-

[26] This methodology expands upon the framework introduced in Brodeur et al. (2016), who introduced as a counterfactual the central t-distribution with one degree of freedom. For additional technical details the interested reader is directed there.

ment & Accounting, Marketing, and Sector & Social Studies are missing between 11 and 17% of statistically insignificant test statistics. These are mostly found in the 'two-star' or $p = 0.05$ interval, where 52% of Management & Accounting's, 79% of Marketing's, and 59% of Sector & Social Studies' misallocated tests are found.

4.3 Testing for Publication Bias using Caliper Tests Following Gerber (2008b) (Sociological Methods and Research)

Turning more particularly to publication bias, in this section we rely on caliper tests following Gerber and Malhotra (2008b). A caliper test examines test statistics in a narrow band just above and below a statistical significance threshold. The underlying logic is that if there is no manipulation (whether publication bias or p-hacking) then we would expect the frequency of tests statistics falling just below a threshold, (e.g., $z = 1.96$) to be very similar to the frequency just above. Here, as elsewhere, we focus on the 5% statistical significance threshold, specifically:

$$R_{-,h} = [1.96 - h, 1.96], R_{+,h} = [1.96, 1.96 + h] \tag{1}$$

for a variable bandwidth parameter h, we estimate probit models where the dependent variable is a dummy variable that takes the value one if a test statistic is statistically significant at the 5%-level, and zero otherwise. In our main specification we report standard errors clustered at article-level.

One advantage of caliper tests over others presented so far is that this method allows us to control for potentially confounding factors. Here we do so for the journal of publication, number of authors, how test statistics are reported, and where in an article they are presented. Our variables of interest are an indicator for 4* journals (the effect of rank), a post-2015 dummy variable (the effect of time), field indicators (differences between research fields), and an indicator for the presence of any discussion of statistical power (the effect of design transparency).

The estimates are reported in Table 3. (See Appendix Tables A4 and A5 for the other statistical significance thresholds.) In columns 1 and 2, we restrict the sample to [1.96±0.50], a window which contains 6,826 test statistics. In columns 3 and 4 we repeat the specification in column 2 but with narrower bandwidths - in columns 3 ([1.96±0.35]) and 4 ([1.96±0.20]). In columns 5 and 6, we use the inverse of the number of tests presented in the same article to weight observations, preventing articles with more tests from having a disproportionate effect on inference.

In Table 3, we find that test statistics in Marketing are about 15 percentage points more likely to be statistically significant than an estimate in the field of Economics & Finance (the omitted field). The estimates are statistically significant in all columns. Similarly, we find that test statistics in the fields of Management and

Accounting (Sector and Social Studies) are between 11.2-14.2 (11.5-17.6) percentage points more likely to be statistically significant than an estimate in the field of Economics & Finance. In other words, field of publication matters.

In contrast, the coefficient estimates for 4* journals and our post-2015 dummy variable are small in value and statistically insignificant, suggesting that the extent of p-hacking is not changing across quality of journal - or at least across the 4 to 4* threshold - nor over time.

4.4 Estimating Publication Bias using the Relative Publication Probabilities Method of Andrews and Kasy (2019) (American Economic Review)

Next we conduct an analysis following Andrews and Kasy (2019). An appealing feature of this method is that it focuses explicitly on publication probabilities, and so allows for an estimate of publication bias isolated from any preceding p-hacking. Recall that publication bias occurs if the outcome of a study is related to the decision (whether by researcher or review process) to publish. The interested reader is directed to the original paper for methodological details and derivations, a detailed description of which space considerations preclude here.

The primary estimated parameter generated by Andrews and Kasy (2019) is the relative publication probability of a statistically significant result being published compared to a statistically *in*significant result.[27] In our full MTurk sample, a statistically significant result is estimated to be 4.61 times *more* likely to be published than a statistically insignificant result, other things equal, indicating a very substantial degree of publication bias.[28]

5 Sample Size

In this section, we consider issues around sample size, statistical power and the cost of data in MTurk research.

As noted, an important underlying concern in assessing research credibility is the propensity for under-powered studies to generate false positives - delivering (by chance) statistically significant evidence in support of an effect when in fact none exists. This would be a problem even if the subset of studies that ended up published were to be chosen at random, but the resulting loss of credibility in the overall corpus of published research is exacerbated if the set of studies that

[27] In the notation of Andrews and Kasy (2019) we refer to β_p.

[28] As a benchmark, Brodeur et al. (2020) only finds such a strong result for studies using instrumental variables (with an estimated relative publication probability of 4.72 times) which can be compared to their baseline of randomized control trial studies which have an estimated relative publication probability of 1.52 times.

researchers choose to publish, or that journals select for publication, also depends positively on statistical significance.

The introduction of Mechanical Turk as a source of research participants was lauded for its ability to access a large number of participants both quickly and in a cost-effective manner (Buhrmester et al. 2011).

In our full sample, the mean number of participants for an MTurk experiment is 450. However, there is great variation. The standard deviation is 814 and sample sizes range from 20 to 15,166. Figure 10 illustrates the sample size distribution for our entire sample. Over a third of studies are derived from experiments with less than 200 participants.

The figure also masks large differences across fields, which we plot in Appendix Figure A6. The number of participants in the Economics & Finance sub-sample (1,908) is not just substantially larger than in, for example, Marketing (382) but the the distribution of sample size is relatively unskewed, and centred around 500. Note there are only a handful of sample sizes less than 200. The bar-charts in the other panels are quite different in character, with distributions strongly favoring smaller samples. For all three other panels, the typical sample size is below 250.

The left panel of Figure 11 plots the evolution of mean sample size used in MTurk-based experiments by year of publication (it is worth noting that publication lags make it likely that in many cases the year of publication may not coincide with the year in which the experiment was conducted). The connected line indicates a strong upward trend in sample size over the study period, moving from an average of around 200 in 2011 to 500 in 2020.

5.1 Sample Size Justification

Sample size, by determining what we can reliably learn from observed variation in outcomes, is central to statistical inference from experimental data. A die failing to fall as a 6 in ten throws should not convince us that the die is loaded, but if it fails in 100 throws then the evidence is close to overwhelming. In empirical work using naturally-occurring or artefactual data the researcher's sample size is often constrained by what data exists, but in experimental work sample size is a key element of research design.

So how do researchers using MTurk as a participant pool justify or rationalize the sample sizes that they choose? The short answer is that, typically, they don't.

Regular readers of MTurk-based research might have noted a dearth of careful discussion of statistical power or sample size considerations in many or most studies. The terse treatment of sample size by Desai and Kouchaki (2017), in introducing the MTurk sample in their study of whether moral symbols can reduce unethical

behavior in the workplace, published in the 4* (elite) Academy of Management Journal, is fairly typical: "We recruited 128 individuals from the United States to participate in an online study through Amazon's Mechanical Turk website. Ten individuals who failed to follow instructions, failed attention checks, or did not respond to questions regarding the study variables of interest were excluded from analyses. The final sample consisted of 118 participants." (Desai and Kouchaki (2017), page 14)). No further discussion of sample size is provided.

To investigate this systematically we adopt a textual approach. We manually categorize each article according to whether or not it provides - any where in the paper - the reader with a rationale for the sample size used. We make no assessment with respect to the merit of any rationale, simply that it was provided.[29] In some cases it may be couched in the formal language of statistical power, and accompanied by a formal power calculation, in others reference made to a fixed sampling time window, or norms about what constituted an 'adequate' sample in a particular sub-field. Each of these was coded the same.

We report in Table 1 that in the full sample 12% of experiments provide the reader with any justification for sample size. However, Table 2 exposes substantial differences between the consolidated research fields, with the percentage of studies discussing power to be highest in Economics & Finance (39%), followed by Management & Accounting (23%), Sector & Social Studies (13%), and lowest in Marketing (8%).

The right panel of Figure 11 evidences that the likelihood the reader of a study is provided a justification of sample size is encouragingly increasing through time, though even in 2020 it is only the case for less than one-fifth of articles.

5.2 Sample Size, Power and *p*-Hacking

To further explore the relationship between sample size, power, and p-hacking we present, in Figure 12, the distribution of test statistics for experiments employing sample sizes below (left panel) and above (right) the median. Both panels exhibit the characteristic valley-peak shape - a distribution with two local maxima, one close to zero, and one in the vicinity just above 1.96, the threshold for 5% statistical significance. However the significance spike is more pronounced in the left-hand panel than the right-hand panel (a Kolmogorov–Smirnov test rejects the null of equality of distributions with $p < 0.000$).

[29] We were skeptical about some of the statistical logic on offer; "Collecting too many observations might increase the likelihood of an overpowered study (i.e., results are deemed significant statistically, but only because of the large number of observations), so we kept the number of participants close to the minimum of 50 per condition suggested by previous work." (Hahl et al. 2018).

Appendix Figure A7 splits our data into quartiles. The first quartile includes MTurk articles with the smallest sample size ($n < 190$), while the fourth quartile includes articles with the largest sample size ($n > 442$). The pattern discussed above still holds with a more distinct peak for studies with the smallest sample sizes, and becomes successively smoother for each successively larger quartile.[30]

Of course, these are correlations and care is needed in their interpretation. In particular, we have already observed that large samples are more prevalent in the Economics & Finance sub-sample, where we have found evidence consistent with less p-hacking.

Turning to sample size justification, Figure 13 presents the distribution of z-statistics for articles which do provide (left panel) or do not provide (right panel) the reader with a justification of sample size.[31] Visually, the distributions are quite different with a discernibly larger peak around the 5% significance threshold for articles that do not include a discussion of sample size (a Kolmogorov–Smirnov test also rejects the null hypothesis that the distributions are the same with $p < 0.000$).

Our caliper analysis also points to articles discussing statistical power being less p-hacked. Our estimates in Table 3 for the 5% statistical significance threshold suggest that test statistics in articles providing a discussion of statistical power or a justification of sample size are about 4 to 7 percentage points less likely to be marginally statistically significant.[32]

6 Cost of Data

As a final exercise we explore the cost of data in research using Mechanical Turk.

In much empirical research using artefactual in the social sciences, sample size is limited by external constraints on data availability. For example, when using administrative data the researcher may be limited by the number of records). With methods that involve data creation, such as the conduct of surveys and running of experiments, additional data points can be expensive such that sample limited by budget constraints. Our prior is that neither of these is likely an important consideration for most experiments conducted on MTurk. Running experiments

[30] Appendix Table A2 formalizes this difference where the null of no p-hacking is rejected by an additional test in the below-median subsample. In Appendix Table A3, we rely on the excess test statistics methodology of Brodeur et al. (2020) and provide further evidence that the extent of p-hacking is nearly 3 times larger for the below-median sub-sample.

[31] Appendix Figures A8 and A9 show findings are robust to weighting and de-rounding.

[32] In Appendix Table A2, we find that articles which do not discuss power or provide a justification of sample size are likely more afflicted by publication bias and/or p-hacking using almost all Elliott et al. (2022)'s tests. In contrast, there is only limited evidence of publication bias and/or p-hacking for the sub-sample of articles that discuss power as only three of six tests are significant at the 5% level. Using Brodeur et al. (2020)'s excess test statistics method yields similar conclusions. See Appendix Table A3.

on Mechanical Turk is, in most cases, cheap and quick in comparison to other commonly-used participant pools. We have already reported in Tables 1 and 2 that across the full sample the mean remuneration per participant is 1.30 USD. In the Marketing sub-sample the mean remuneration is around 80 cents. Due to selective averaging both of these figures is almost certainly an over-estimate.[33]

The low marginal cost of data collection, when combined with a publishing environment that incentivizes statistical significance at the article level, rather than credibility of inference at the published corpus level, makes it plausible that there is a greater volume of unpublished insignificant results that we do not observe than would be the case if data were more costly. The implications that this could have for the credibility of published studies is advanced anecdotally by Calin-Jageman (2018): "MTurk makes running studies so easy that it exacerbates the publication bias problem. There are so many researchers running so many studies. My theory is a retelling of the publication bias story ... an incredibly pernicious one. What is new, I think, is the way MTurk has made the opportunity costs for conducting a study so negligible: its like fuel being poured on the publication dumpster fire. MTurk dramatically increases the number of people running studies and the number of studies run by each researcher. Moreover, the low opportunity cost means it is less painful to simply move on if results didn't pan out. With MTurk it costs very little to fill your file drawer while mining noise for publication gold".[34]

We explore first the question of the cost of the experiment, the remuneration of subjects, and how it relates to the likelihood that a study reports a marginally statistically significant result.[35]

We create two measures: remuneration per participant, which we interpret as the marginal cost of an additional data point, and the total remuneration per experiment. We measure the former as the average wage reported, summing any

[33]The elasticity of labor supply facing an individual researcher on the platform has also been shown by Dube et al. (2020) to be very small - around 0.1.

[34]The author continues with anecdotal evidence of the low cost of using MTurk from his efforts to collect unpublished results for a meta-analysis: "For the red-romance meta-analysis, one lab sent us 6 unpublished online studies representing 956 participants (all conducted in 2013). The lab actually sent us the data in a chain of emails because digging up data from one study reminded them of the next and so on. It had been 5 years, but *the lab leader reported having completely forgotten about the experiments.* That to me indicates the incredibly low opportunity cost of MTurk. If I had churned through 956 in-person participants I would remember it, and I would have had so much sunk cost that I would have wanted to find some outlet for publishing the result. ... But when you can launch a study and see 300 responses roll in within an hour, your investment in eventually writing up is weak."

[35]See Camerer and Hogarth (1999) for a discussion of when and why experimental subjects should be paid and a review of financial performance-based incentives for 74 experimental papers. In our sample the MTurk-article-reader is rarely provided guidance as to how the researcher set remuneration. Even when provided the information is scant. A typical example is provided in Tzini and Jain (2018): "participants were paid a participation fee of $0.10 according to the norm for payments of the platform".

participation or show-up fee with average reported bonus or performance-related payment.

In our sample, about one third of test statistics are derived from articles that numerically report remuneration per participant as a monetary amount.[36] Around a quarter of our sample describe their participant's remuneration as "nominal" or similar adjective, these we code as 'nominal'. Due to the imprecise nature of the description, we do not use these in our calculations of cost. The remaining majority of articles are silent on remuneration.[37] For articles reporting costs per participant, the median cost of an experiment is about $260 and the median cost of a participant is about $0.80. Appendix Figures A12 and A13 illustrate the cost distributions for our entire sample. Over 15 percent of the test statistics that we collect are based on experiments that cost less than $100. Over 40 percent of the test statistics are in articles where the cost per participant is $0.50 or less.[38]

Again, there are striking differences between fields. The average cost of an experiment is over $2,500 in Economics & Finance, with mean remuneration over $1.80. The mean cost per experiment is $693, $437 and $378 for Management & Accounting, Marketing and Sector & Social Studies, respectively. The mean remuneration per participant is $1.85, $0.80 and $0.94 for Management and Accounting, Marketing and Section and Social Studies, again respectively. Appendix Figures A15 and A16 illustrate the distribution by field.

Figure 14 presents the distributions of test statistics for MTurk experiments that cost less than the median (left panel) and more than the median (right panel). The spike to the right of 1.96 is more pronounced for experiments that cost less than the median. We further decompose the sample in Appendix Figure A17 by splitting the sample into quartile buckets by total cost. Here, we find that the bunching around $z = 1.96$ becomes steadily less pronounced as we move from the lowest to the highest cost quartile of experiments.[39]

7 Conclusion

We provide the first analysis of the statistical practices of researchers using Amazon's Mechanical Turk as a source of subjects for experiments across business re-

[36] Notably, all studies on MTurk provide some remuneration, in its minimal form as a 'show up' fee. This reflects the nature of the platform as a micro-task employment site.

[37] The proportion of articles reporting cost by year is illustrated in Appendix Figure A10. The distributions of test statistics by type of cost reporting are illustrated in Appendix Figure A11.

[38] Appendix Figure A14 evidences that the likelihood the reader of a study is provided cost information increased over the study window.

[39] For completeness we apply the other methods from earlier to this decomposition. In Appendix Table A2, we find evidence that both experiments that cost less and more than the median are p-hacked. We document that the extent of p-hacking is larger for experiments that cost less than the median using Brodeur et al. (2020)'s excess test statistics method. See Appendix Table A3.

search fields, with special reference to the patterns of statistical significance found in the corpus of studies published in a wide set of highly regarded journals.

Applying a set of state-of-the art methodologies, including several only recently developed, we find consistent and persuasive evidence of widespread p-hacking and publication bias. This is particularly true for research in Marketing and Sector Studies, and less so in the Economics & Finance and Management & Accounting fields. There are far fewer null results in the published literature than there should be, and a high proportion of published non-null results may have had their statistical significance artificially inflated and/or come from under-powered tests (that is from studies that use samples that are too small).

Our findings are in one sense pessimistic and in another optimistic. The credibility of results contained in the existing corpus of research using the MTurk platform is substantially compromised. If a reader were to pick at random a study from our sample, our analysis points to this result being unlikely to be replicable. However, going forward the flaws we identify relate to the way in which MTurk experiments are conducted, and results selected for publication, by the research community, rather than flaws inherent to the platform itself (Hauser et al. (2019)). This distinction is important as it suggests that there is no reason - at least from this perspective - for researchers to discontinue to use MTurk and other similar platforms. Rather more rigorous attention to statistical practice, in particular use of larger samples to provide appropriately powered experiments, need to become more common. Fortunately, this is an area of research where - as we have shown - data points are cheap.

References

Aguinis, H., Villamor, I. and Ramani, R. S.: 2021, MTurk Research: Review and Recommendations, *Journal of Management* **47**(4), 823–837.

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

Arechar, A. A., Kraft-Todd, G. T. and Rand, D. G.: 2017, Turking Overtime: How Participant Characteristics and Behavior Vary Over Time and Day on Amazon Mechanical Turk, *Journal of the Economic Science Association* **3**(1), 1–11.

Berinsky, A. J., Huber, G. A. and Lenz, G. S.: 2012, Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk, *Political Analysis* **20**(3), 351–368.

Brodeur, A., Cook, N. and Heyes, A.: 2020, Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics, *American Economic Review* **110**(11), 3634–60.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y.: 2016, Star Wars: The Empirics Strike Back, *American Economic Journal: Applied Economics* **8**(1), 1–32.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M. et al.: 2019, Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research, *Research Policy* **48**(9), 103796.

Buhrmester, M., Kwang, T. and Gosling, S. D.: 2011, Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?, *Perspectives on Psychological Science* **6**(1), 3–5.

Calin-Jageman, R.: 2018, The Perils of MTurk, Part 1: Fuel to the Publication Bias Fire?, https://thenewstatistics.com/itns/2018/05/02/the-perils-of-mturk-part-1-fuel-to-the-publication-bias-fire/. Accessed: 2022-07-06.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. et al.: 2016, Evaluating Replicability of Laboratory Experiments in Economics, *Science* **351**(6280), 1433–1436.

Camerer, C. F., Dreber, A. and Johannesson, M.: 2019, Replication and other practices for improving scientific quality in experimental economics, *Handbook of*

Research Methods and Applications in Experimental Economics, Edward Elgar Publishing, pp. 83–102.

Camerer, C. F. and Hogarth, R. M.: 1999, The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework, *Journal of Risk and Uncertainty* **19**(1), 7–42.

Cattaneo, M. D., Jansson, M. and Ma, X.: 2020, Simple Local Polynomial Density Estimators, *Journal of the American Statistical Association* **115**(531), 1449–1455.

Chopra, F., Haaland, I., Roth, C., Stegmann, A. et al.: 2022, The Null Result Penalty. Mimeo: University of Bonn and University of Cologne.

Christensen, G. and Miguel, E.: 2018, Transparency, Reproducibility, and the Credibility of Economics Research, *Journal of Economic Literature* **56**(3), 920–80.

Coppock, A.: 2019, Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach, *Political Science Research and Methods* **7**(3), 613–628.

Cox, G. and Shi, X.: 2022, Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models, *Review of Economic Studies* .

DellaVigna, S. and Linos, E.: 2022, RCTs to Scale: Comprehensive Evidence from Two Nudge Units, *Econometrica* **90**(1), 81–116.

Desai, S. D. and Kouchaki, M.: 2017, Moral Symbols: A Necklace of Garlic Against unethical Requests, *Academy of Management Journal* **60**(1), 7–28.

Doucouliagos, C. and Stanley, T. D.: 2013, Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity, *Journal of Economic Surveys* **27**(2), 316–339.

Dube, A., Jacobs, J., Naidu, S. and Suri, S.: 2020, Monopsony in Online Labor Markets, *American Economic Review: Insights* **2**(1), 33–46.

Elliott, G., Kudrin, N. and Wüthrich, K.: 2022, Detecting p-Hacking, *Econometrica* **90**(2), 887–906.

Falk, A., Meier, S. and Zehnder, C.: 2013, Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples, *Journal of the European Economic Association* **11**(4), 839–852.

Furukawa, C.: 2019, Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method. MIT Mimeo.

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

Gerber, A. S. and Malhotra, N.: 2008b, Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?, *Sociological Methods & Research* **37**(1), 3–30.

Gillen, B., Snowberg, E. and Yariv, L.: 2019, Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study, *Journal of Political Economy* **127**(4), 1826–1863.

Goodman, J. K., Cryder, C. E. and Cheema, A.: 2013, Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples, *Journal of Behavioral Decision Making* **26**(3), 213–224.

Guala, F. and Mittone, L.: 2005, Experiments in Economics: External Validity and the Robustness of Phenomena, *Journal of Economic Methodology* **12**(4), 495–515.

Hahl, O., Kim, M. and Zuckerman Sivan, E. W.: 2018, The Authentic Appeal of the Lying Demagogue: Proclaiming the Deeper Truth about Political Illegitimacy, *American Sociological Review* **83**(1), 1–33.

Harrison, G. W., Lau, M. I. and Rutström, E. E.: 2009, Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments, *Journal of Economic Behavior & Organization* **70**(3), 498–507.

Hauser, D., Paolacci, G. and Chandler, J.: 2019, Common concerns with MTurk as a participant pool: Evidence and solutions, *Handbook of Research Methods in Consumer Psychology*, Routledge/Taylor & Francis Group, p. 319–337.

Havránek, T.: 2015, Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting, *Journal of the European Economic Association* **13**(6), 1180–1204.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. and Jennions, M. D.: 2015, The Extent and Consequences of p-Hacking in Science, *PLoS Biology* **13**(3), e1002106.

Horton, J. J., Rand, D. G. and Zeckhauser, R. J.: 2011, The Online Laboratory: Conducting Experiments in a Real Labor Market, *Experimental Economics* **14**(3), 399–425.

Ioannidis, J. P.: 2005, Why Most Published Research Findings Are False, *PLoS Medecine* **2**(8), e124.

Ioannidis, J., Stanley, T. and Doucouliagos, H.: 2017, The Power of Bias in Economics Research, *Economic Journal* **127**, F236–F265.

Johnson, D. and Ryan, J. B.: 2020, Amazon Mechanical Turk Workers Can Provide Consistent and Economically Meaningful Data, *Southern Economic Journal* **87**(1), 369–385.

Lee, Y. S., Seo, Y. W. and Siemsen, E.: 2018, Running Behavioral Operations Experiments using Amazon's Mechanical Turk, *Production and Operations Management* **27**(5), 973–989.

Levitt, S. D. and List, J. A.: 2007, On the Generalizability of Lab Behaviour to the Field, *Canadian Journal of Economics/Revue canadienne d'économique* **40**(2), 347–370.

Paolacci, G., Chandler, J. and Ipeirotis, P. G.: 2010, Running Experiments on Amazon Mechanical Turk, *Judgment and Decision making* **5**(5), 411–419.

Snowberg, E. and Yariv, L.: 2021, Testing the Waters: Behavior Across Participant Pools, *American Economic Review* **111**(2), 687–719.

Stanley, T. D. and Doucouliagos, H.: 2014, Meta-regression Approximations to Reduce Publication Selection Bias, *Research Synthesis Methods* **5**(1), 60–78.

Swanson, N., Christensen, G., Littman, R., Birke, D., Miguel, E., Paluck, E. L. and Wang, Z.: 2020, Research Transparency Is on the Rise in Economics, *AEA Papers and Proceedings*, Vol. 110, pp. 61–65.

Tzini, K. and Jain, K.: 2018, Unethical Behavior Under Relative Performance Evaluation: Evidence and Remedy, *Human Resource Management* **57**(6), 1399–1413.

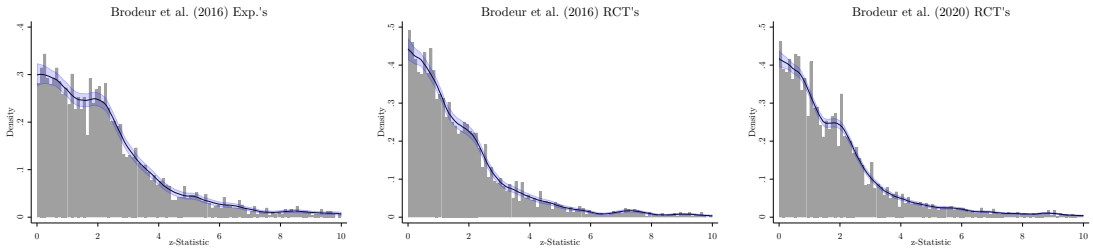
Vivalt, E.: 2019, Specification Searching and Significance Inflation Across Time, Methods and Disciplines, *Oxford Bulletin of Economics and Statistics* **81**(4), 797–816.

Zhang, L. and Ortmann, A.: 2013, Exploring the Meaning of Significance in Experimental Economics. UNSW Australian School of Business Research Paper 2013-32.

Ziliak, S. T. and McCloskey, D. N.: 2004, Size Matters: the Standard Error of Regressions in the American Economic Review, *Journal of Socio-Economics* **33**(5), 527–546.

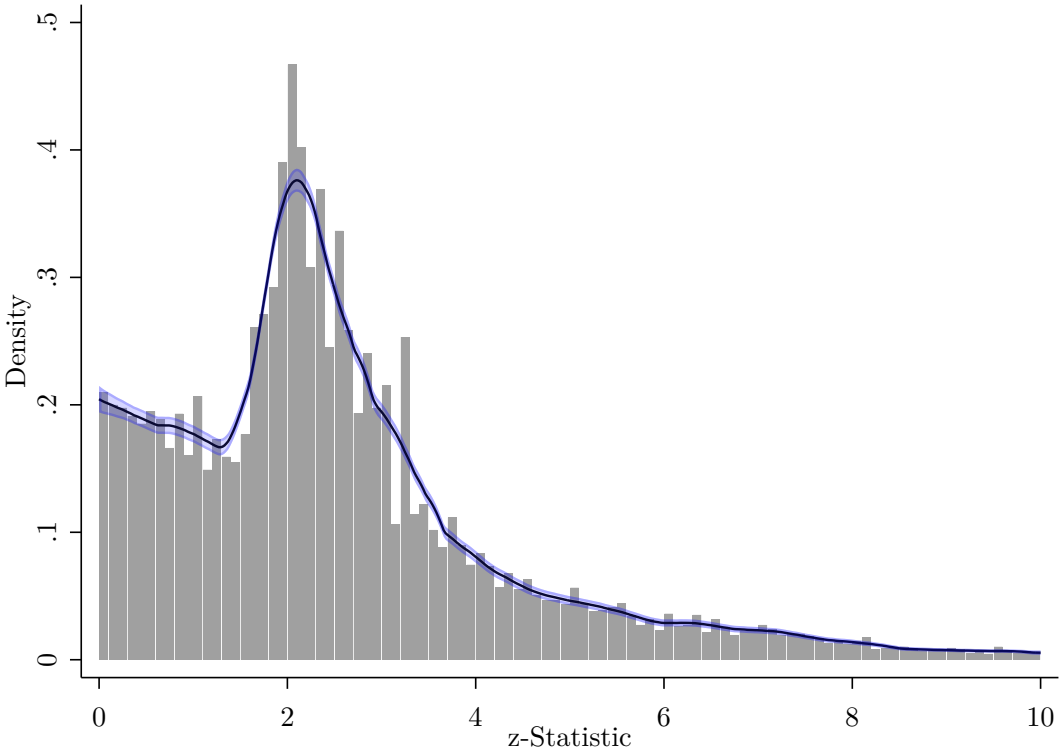
8 Figures

Figure 1: z-Statistic Distributions from Related Studies



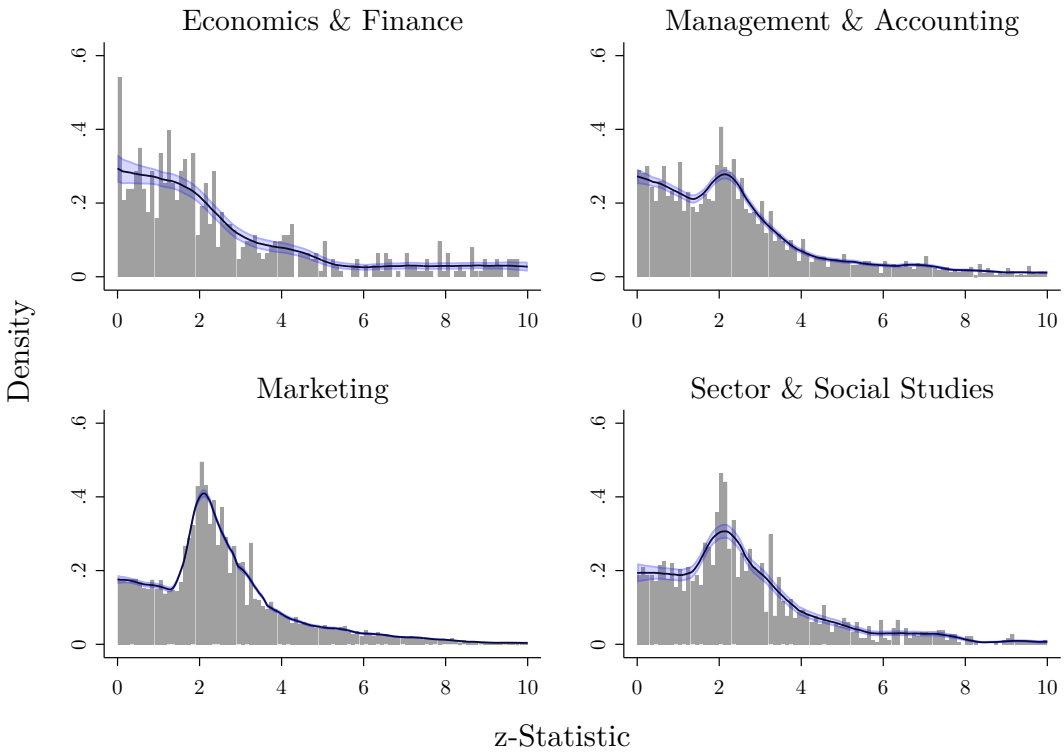
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Bins are 0.1 wide. Epanechnikov kernel superimposed. The top panel presents all laboratory experiment test statistics from the *American Economic Review*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics* published 2005 through 2011 (from Brodeur et al. (2016)). The middle panel presents all RCT test statistics from the same journals and period. The bottom panel presents all RCT test statistics from the top-ranked 25 economics journals published 2015 through 2018 (from Brodeur et al. (2020)). No weights applied.

Figure 2: z-Statistics in Mechanical Turk Articles



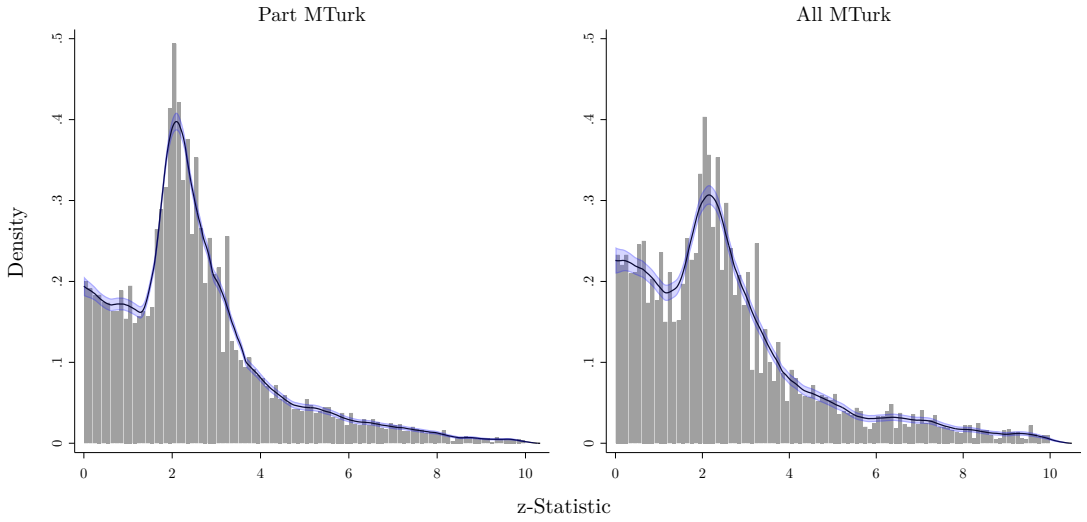
Notes: This figure displays a histogram of test statistics for the full sample of MTurk test statistics. $z \in [0, 10]$. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 3: z-Statistics by Field



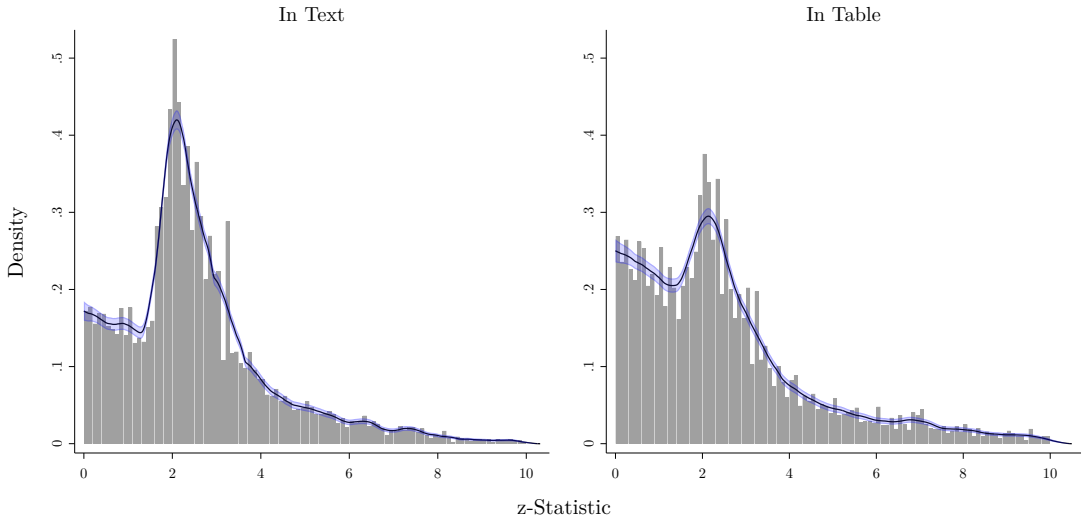
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by field. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 4: z-Statistics by MTurk Share



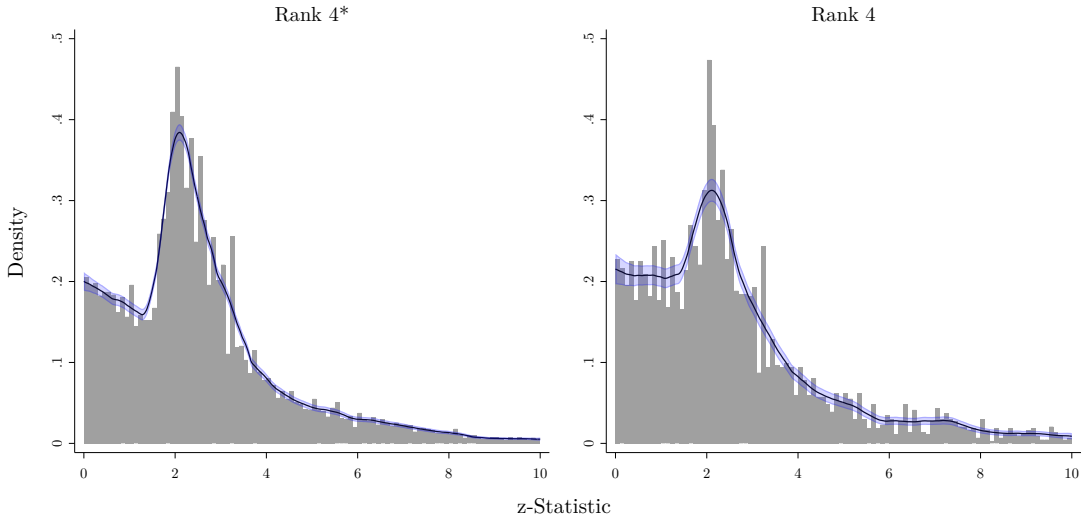
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by MTurk share in article. The left panel restricts the sample to articles reporting MTurk AND non-MTurk experiments. The right panel restricts the sample to articles on MTurk only. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 5: z-Statistics by Reporting Method



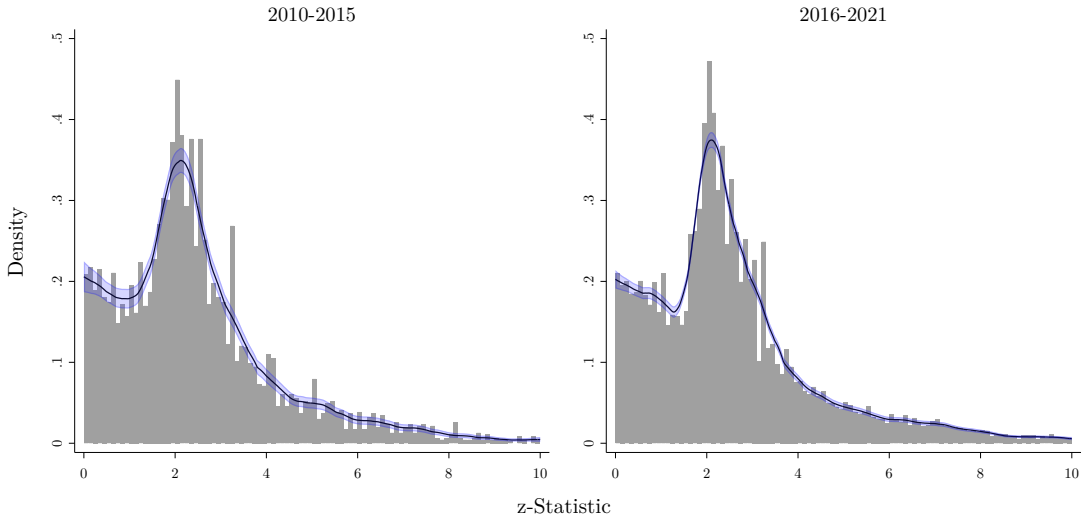
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for test statistics reported only in the manuscript text (left panel) and a table (right panel). Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied. Appendix Figure A4 shows results only for marketing subsample.

Figure 6: z-Statistics by Journal Rank



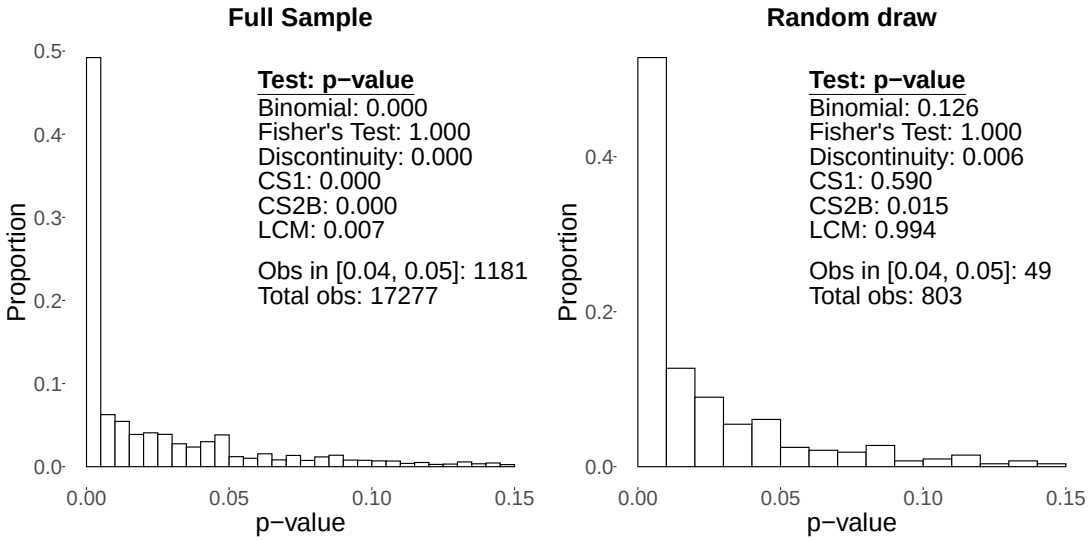
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by journal ranking. The left panel restricts the sample to 4-rated journals, the right panel restricts the sample to 4*-rated journals. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 7: z-Statistics by Time Period



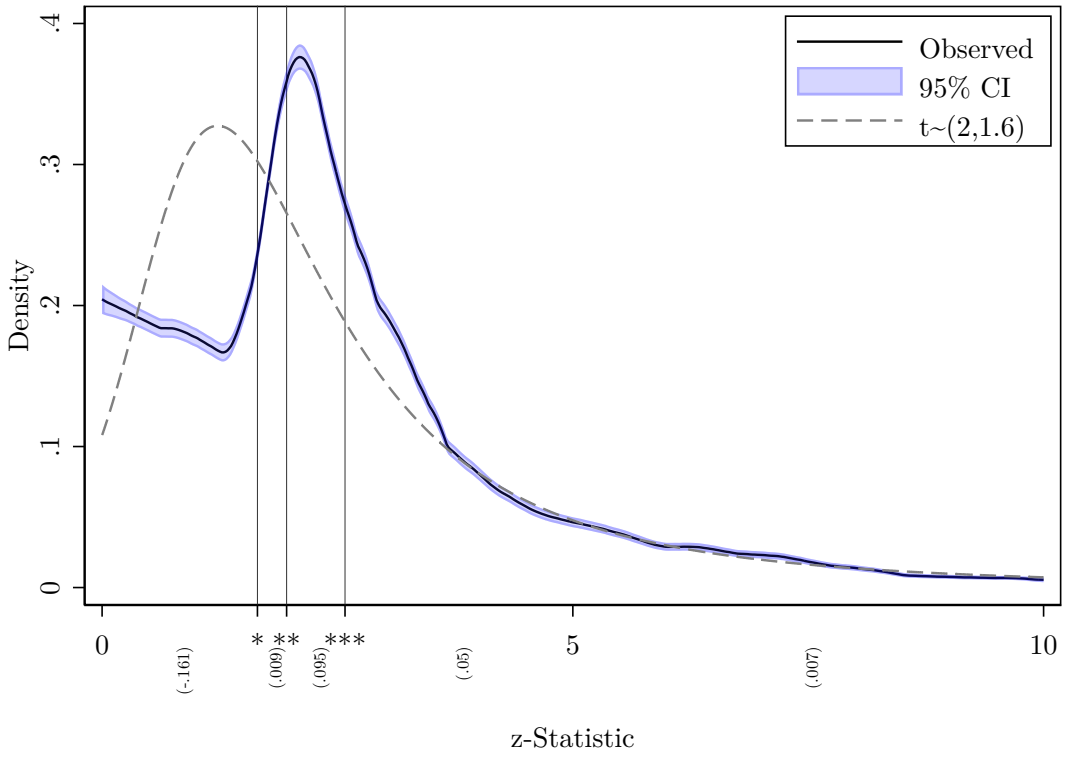
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ over time. The left panel restricts the sample to articles published in 2010–2015. The right panel restricts the sample to articles published in 2016–2021. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 8: Application of Elliott et al. (2022)



Notes: The left panel is a direct application of Elliott et al. (2022)'s p-hacking test battery to our full sample. The right panel draws a random sample of one test statistic per article, following Elliott et al. (2022), to account for within-article dependencies in test statistics. Section 4.1 discusses the p-curve histograms and included tests in detail.

Figure 9: Excess Test Statistics



Notes: This figure presents the observed distribution of test statistics for our MTurk sample as a solid line. The counterfactual distribution we would expect to observe in the absence of publication bias or p-hacking (see Brodeur et al. (2016) and Brodeur et al. (2020)) is the dashed line. Below the horizontal axis we include the difference in mass between statistical significance thresholds. For example, the difference in mass between the observed and counterfactual distribution tails is 0.007 where we expect minimal distortion (above $z > 5$).

Figure 10: Sample Sizes

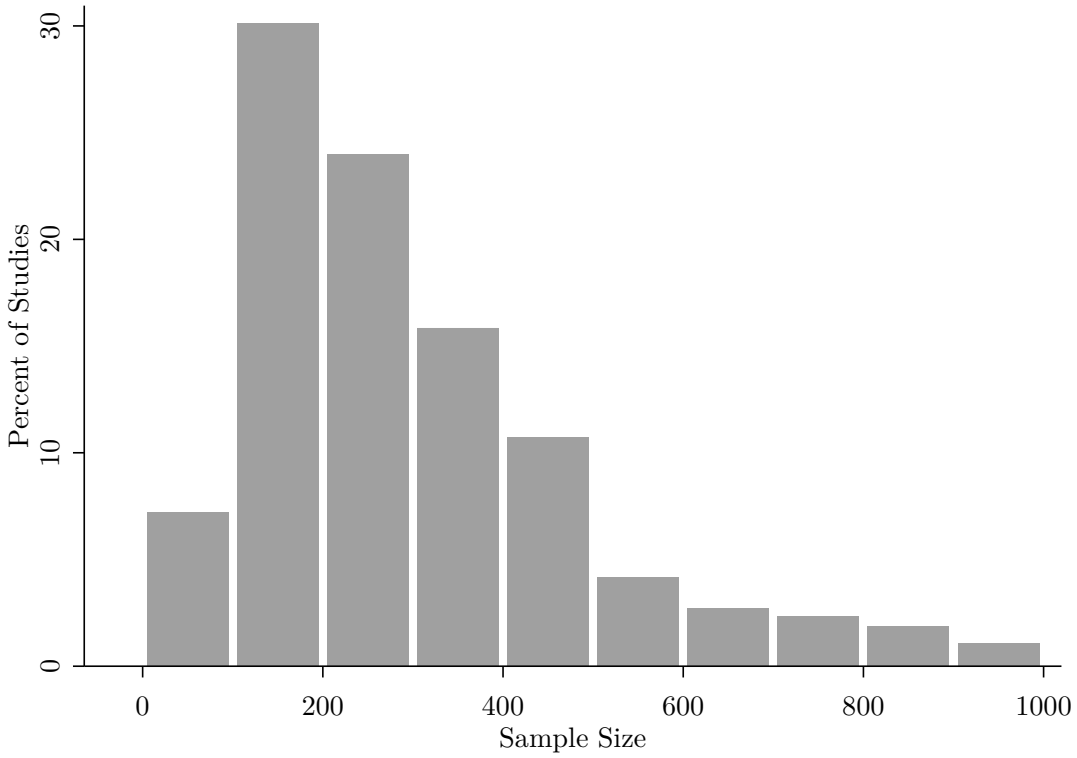


Figure 11: Number of Participants and Power Discussion Over Time

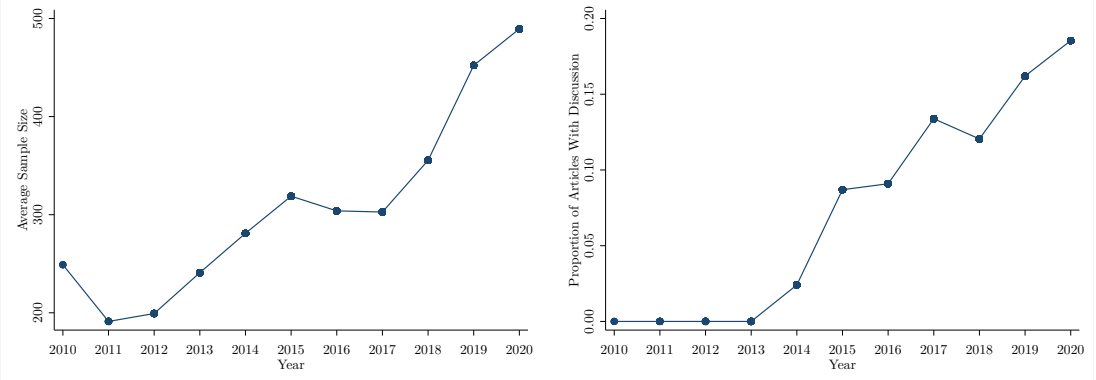
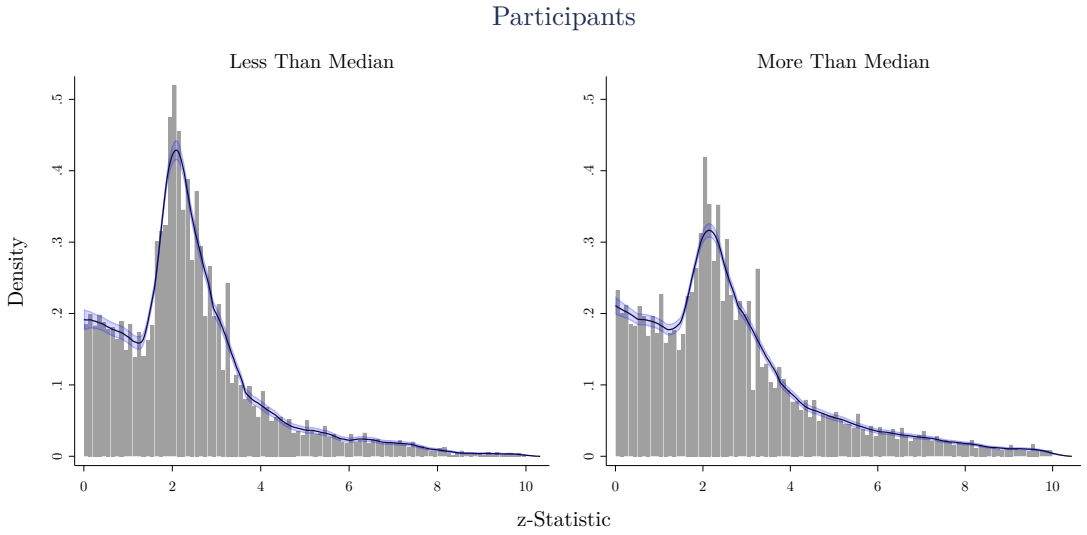
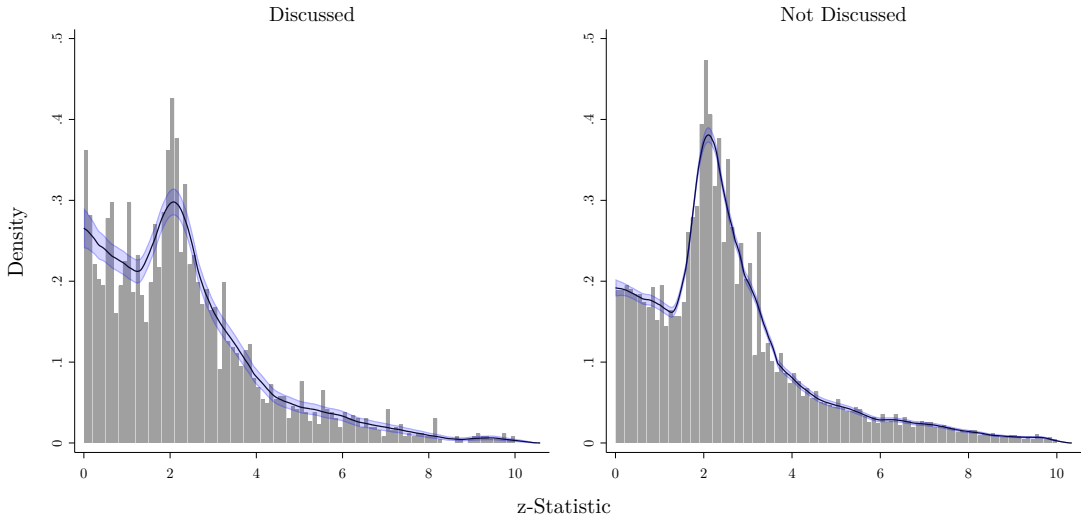


Figure 12: z-Statistics by Sample Size



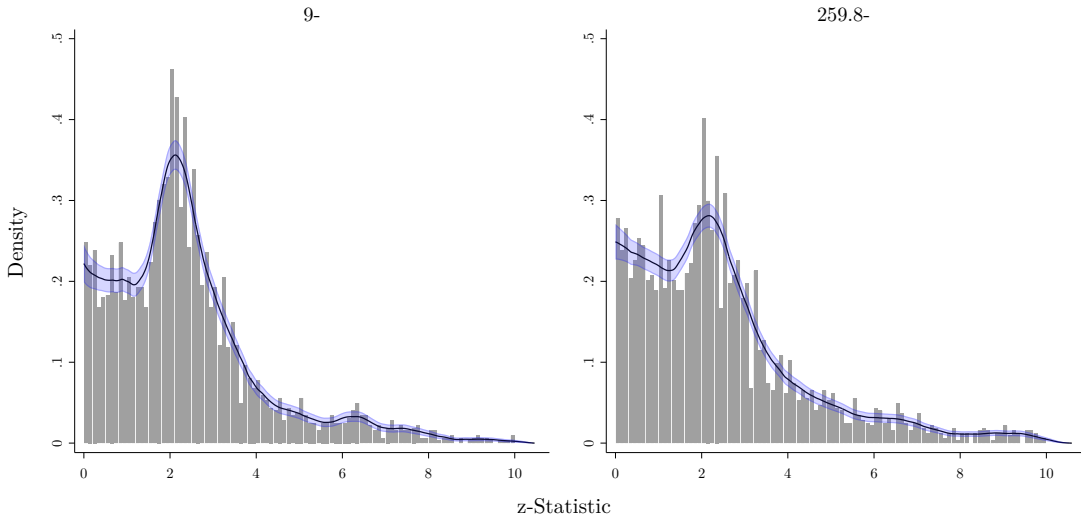
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by sample size. The left (right) panel contains test statistics from samples less than 291 participants, the median sample size. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 13: z-Statistics by Existence of Sample Size Discussion



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by statistical power status. The left (right) panel restricts the sample to articles providing (not providing) a discussion of/justification for sample size. Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights applied.

Figure 14: z-Statistics by Participant Remuneration



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by total participant remuneration. The left (right) panel contains test statistics costing from experiments where remuneration was above (below) median (259.80 USD). Bins are 0.1 wide. Epanechnikov kernel superimposed. No weights attached.

9 Tables

Table 1: Summary Statistics

	Mean	Std. Dev.	Min	Max	N
	(1)	(2)	(3)	(4)	(5)
4* Journal	0.79	0.41	0	1	22,989
Number of Authors	2.66	0.84	1	7	22,989
Year	2017.6	2.38	2010	2021	22,989
Test in Table	0.60	0.49	0	1	22,989
Participants	450	814	20	15,166	22,297
Discussion Statistical Power	0.12	0.33	0	1	22,989
Cost per Experiment	632	1,044	9	7,984	6,998
Cost per Participant	1.30	1.63	0.04	21.5	6,998
Presence of non-MTurk Results	0.70	0.46	0	1	22,989

Notes: This table provides summary statistics. The unit of observation is a test statistic.

Table 2: Summary Statistics by Field

ABS-Defined Fields	Number Part.	Power Disc.	Cost Expe.	Cost Part.	Articles	Tests
	(1)	(2)	(3)	(4)	(5)	(6)
Accounting	279	10%	534	2.01	39	1,118
Economics	1,922	40%	2,603	1.79	18	738
Entrepreneurship	1,128	2%	1,862	5.64	4	44
Finance	400	0%	1,900	4.75	2	21
General Management	336	24%	449	1.85	38	472
Human Resource Management	251	45%	147	0.57	7	158
Information Management	589	7%	389	0.93	19	312
Innovation	175	88%	650	3.80	2	8
Marketing	382	8%	270	0.80	685	15,705
Operations Research	932	34%	907	1.39	32	1,213
Operations & Tech. Mgmt	325	30%	1,018	3.49	10	208
Organization Studies	496	22%	753	1.50	38	765
Public Sector and Healthcare	660	36%	669	1.10	9	175
Sector Studies	253	9%	210	0.70	89	1,363
Social Sciences	683	22%	525	1.14	36	648
Strategy	386	0%	1,679	3.64	3	41
Consolidated Fields						
Economics and Finance	1,908	39%	2,596	1.82	20	759
Management and Accounting	517	23%	693	1.85	201	4,514
Marketing	382	8%	437	0.80	270	15,705
Other: Sector & Social Sci.	390	13%	378	0.94	125	2,011
Economists in non-Econ Journals	526	23%	866	1.36	33	1,186

Notes: This table alphabetically presents sample by field as defined bu the from the 2018 edition of the Academic Journal Guide. Fields that did not report any MTurk estimates are excluded. The unit of observation is a test statistic, except for column 1. Column 1 reports the number of articles that contribute to sample. Column 2 reports number of test statistics. Column 3 mean sample size. Column 4 reports percentage of test statistics in articles providing discussion of sample size. Column 5 and 6 report the average participant remuneration per experiment and average remuneration per participant respectively. The last row restricts the sample to economists identified using the RePEc list who published in non-economics journals.

Table 3: Caliper Test Relating to Statistical Significance at the 5 Percent Level

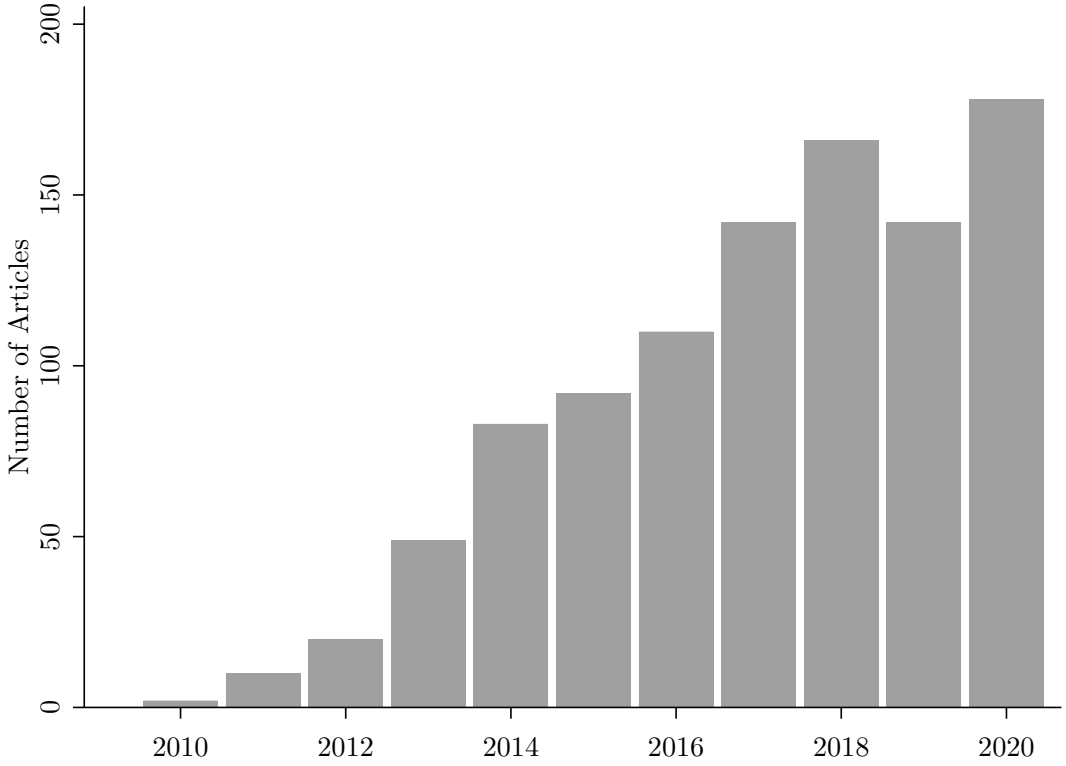
	(1)	(2)	(3)	(4)	(5)	(6)
Management & Account	0.134	0.119	0.142	0.121	0.112	0.122
	(0.042)	(0.047)	(0.058)	(0.076)	(0.053)	(0.084)
Marketing	0.161	0.143	0.161	0.156	0.164	0.141
	(0.045)	(0.050)	(0.059)	(0.078)	(0.062)	(0.087)
Sector & Social Sci.	0.144	0.138	0.176	0.164	0.115	0.155
	(0.050)	(0.054)	(0.064)	(0.081)	(0.065)	(0.090)
Year > 2015	0.012					
	(0.017)					
Journal 4*	-0.017	-0.002	-0.014	-0.045	-0.043	-0.038
	(0.021)	(0.022)	(0.025)	(0.033)	(0.026)	(0.047)
Power Discussed		-0.038	-0.045	-0.062	-0.067	-0.075
		(0.022)	(0.026)	(0.032)	(0.024)	(0.033)
Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Report Text/Table	Y	Y	Y	Y	Y	Y
Number of Authors	Y	Y	Y	Y	Y	Y
Journal FE		Y	Y	Y	Y	Y
Observations	6,826	6,826	5,213	3,098	6,826	3,098
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]	[1.96±0.50]	[1.96±0.20]
Weight Articles					Y	Y

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

10 ONLINE APPENDIX

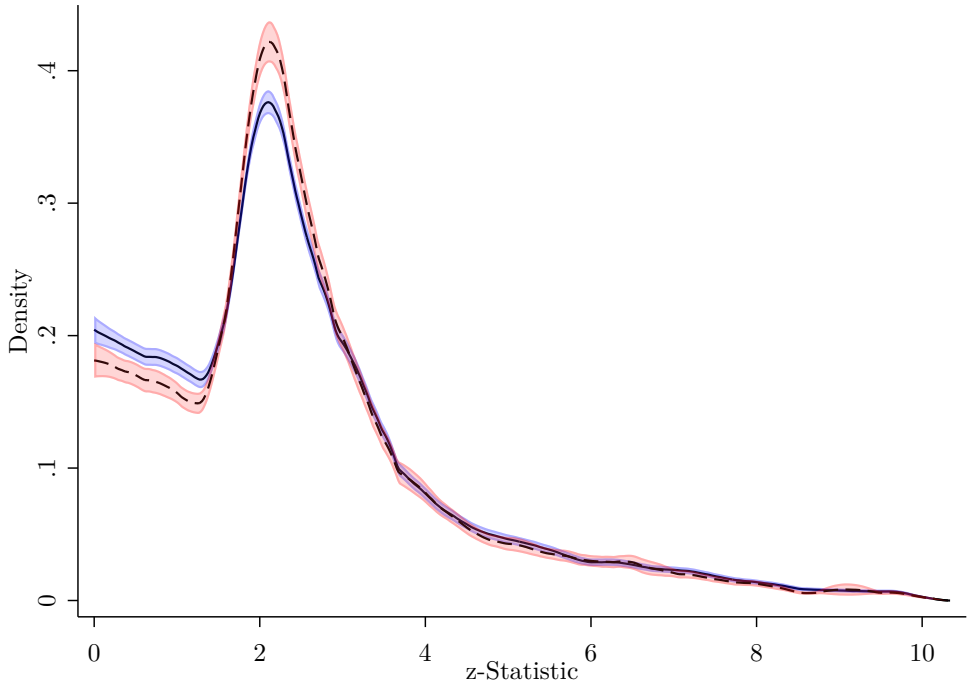
11 Appendix Figures

Figure A1: Year of Publication



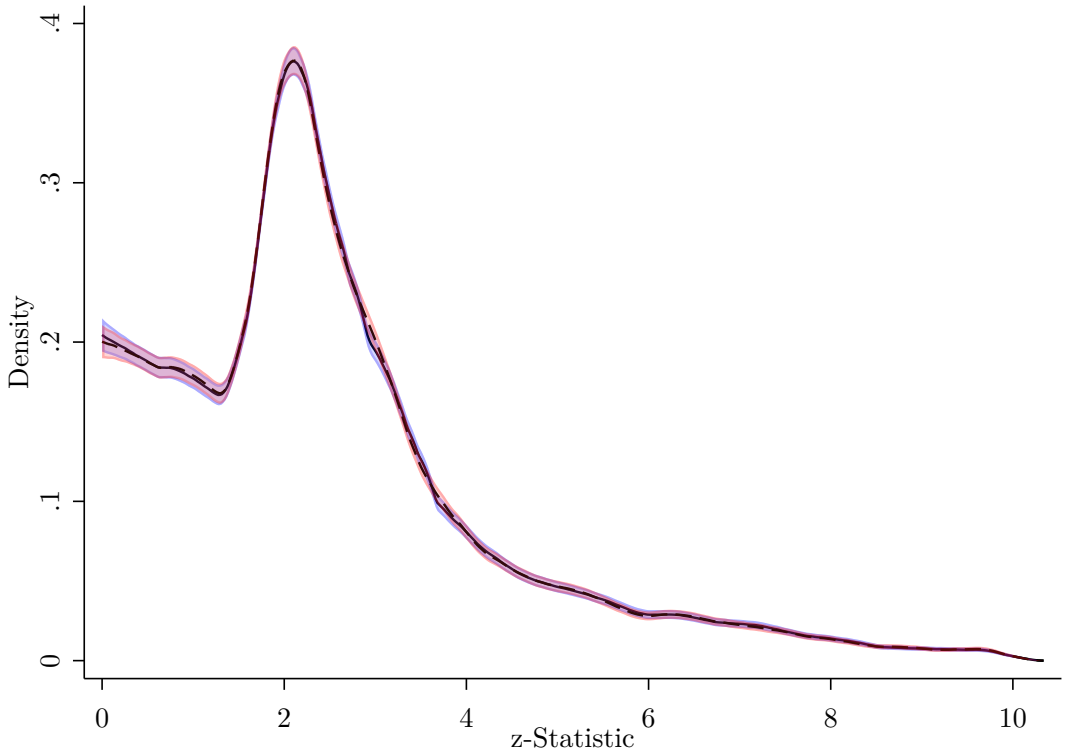
Notes: An observation is a single article. 2021 is an incomplete collection year and is excluded.

Figure A2: z-Statistics in Mechanical Turk Articles: Article Weights



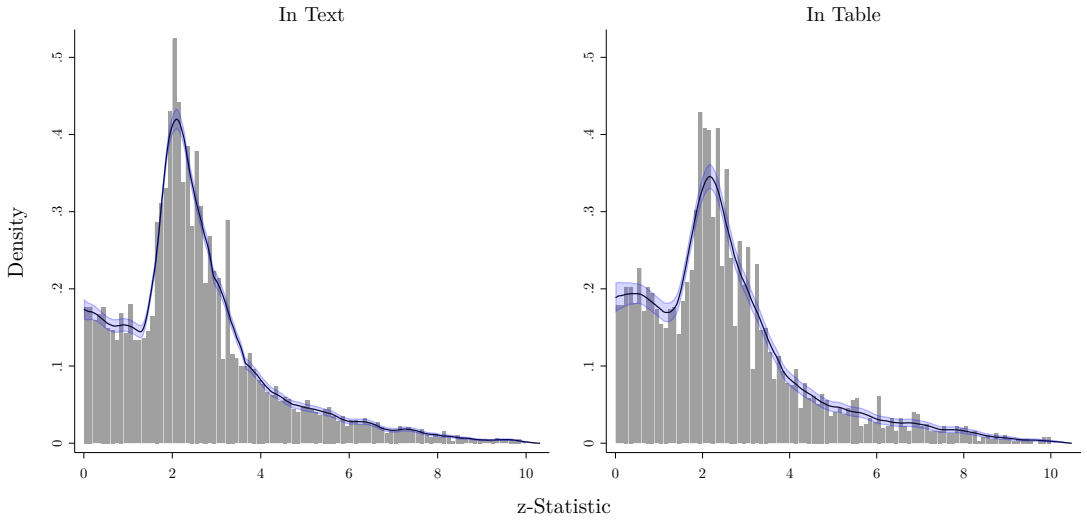
Notes: This figure displays a histogram of test statistics for our full sample of Mechanical Turk test statistics. $z \in [0, 10]$. Bins are 0.1 wide. Solid line and blue confidence intervals correspond to unweighted results. Dashed line and red confidence intervals correspond to article weights. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A3: z-Statistics in Mechanical Turk Articles: Derounding



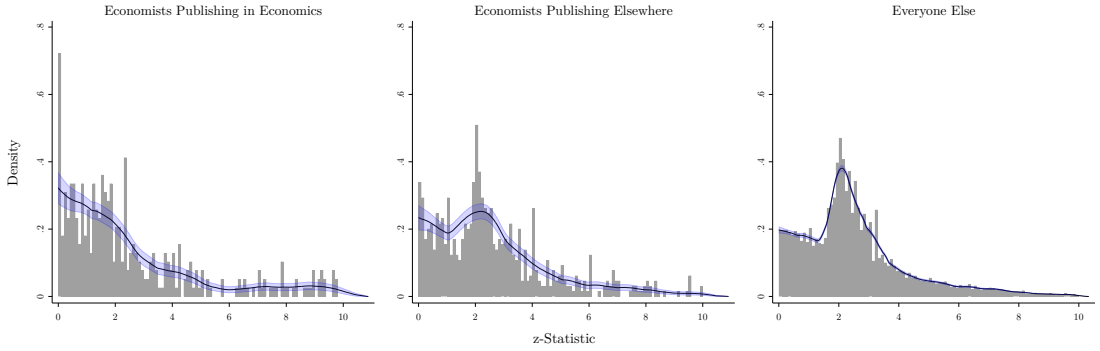
Notes: This figure displays a histogram of test statistics for $z \in [0, 10]$. Bins are 0.1 wide. Solid line and blue confidence intervals correspond to unchanged results. Dashed line and red confidence intervals correspond to derounded test statistics following Brodeur et al. (2016).

Figure A4: z-Statistics by Reporting Method - Marketing Only



Notes: This figure displays histograms of test statistics *in marketing* for $z \in [0, 10]$ for test statistics in a table (left panel) and in the manuscript text (right panel), respectively. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A5: Economists in Economics and in Other Contexts



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Bins are 0.1 wide. Epanechnikov kernel superimposed. The left panel presents economists (as identified through RePEc) publishing in an economics journal. The middle panel presents economists publishing in all other journals. The right panel presents non economists (the remainder after those identified using the RePEc list - see text) publishing in both economics and other contexts. No weights applied.

Figure A6: Sample Size by Field

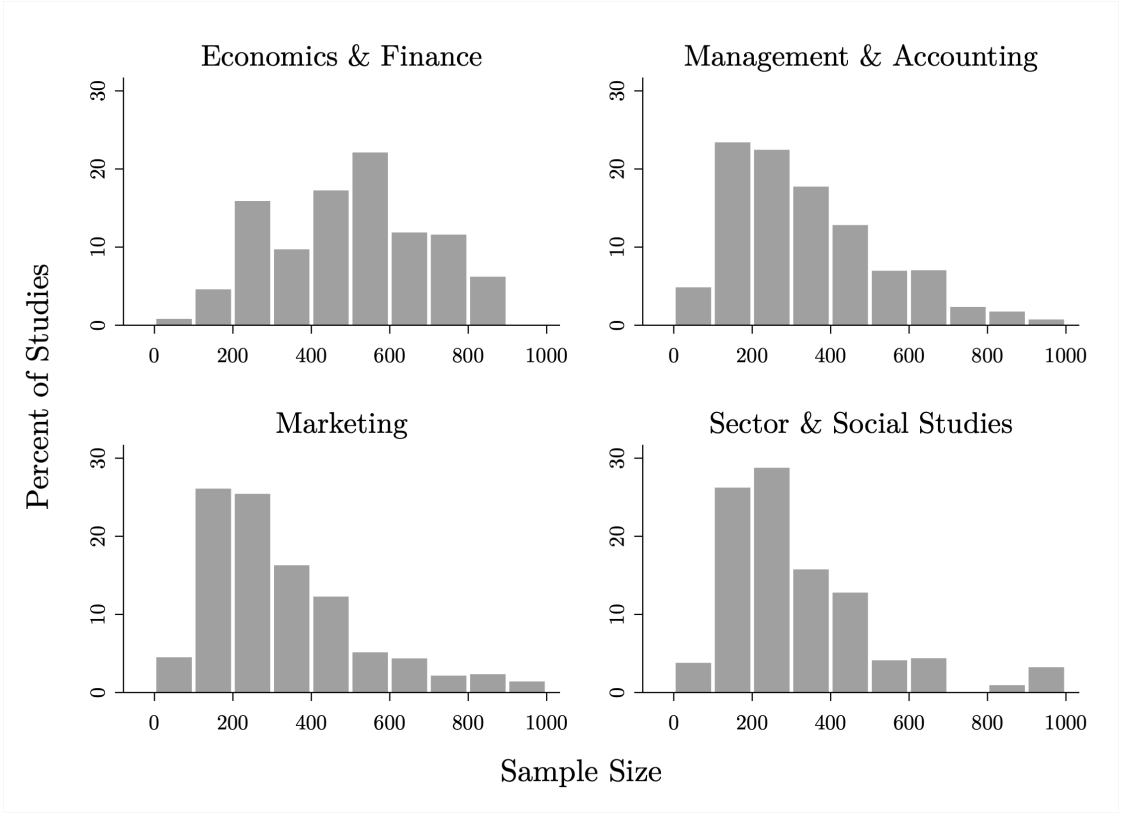
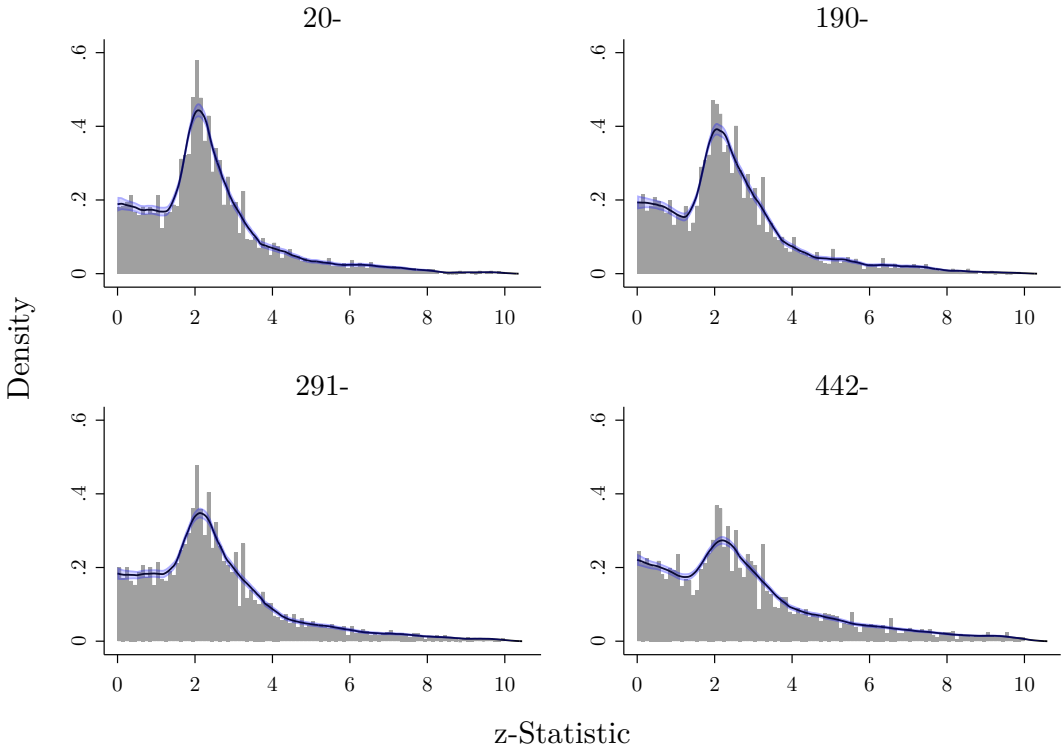
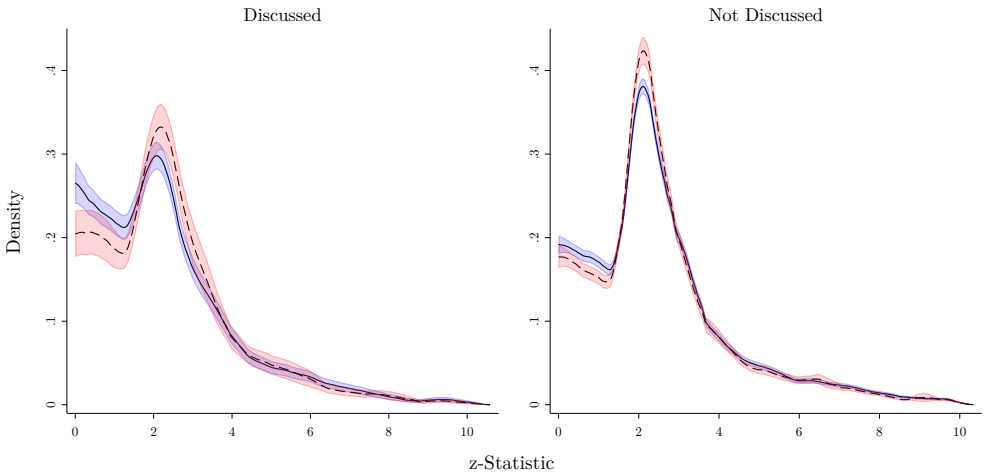


Figure A7: z-Statistics by Sample Size Quartile



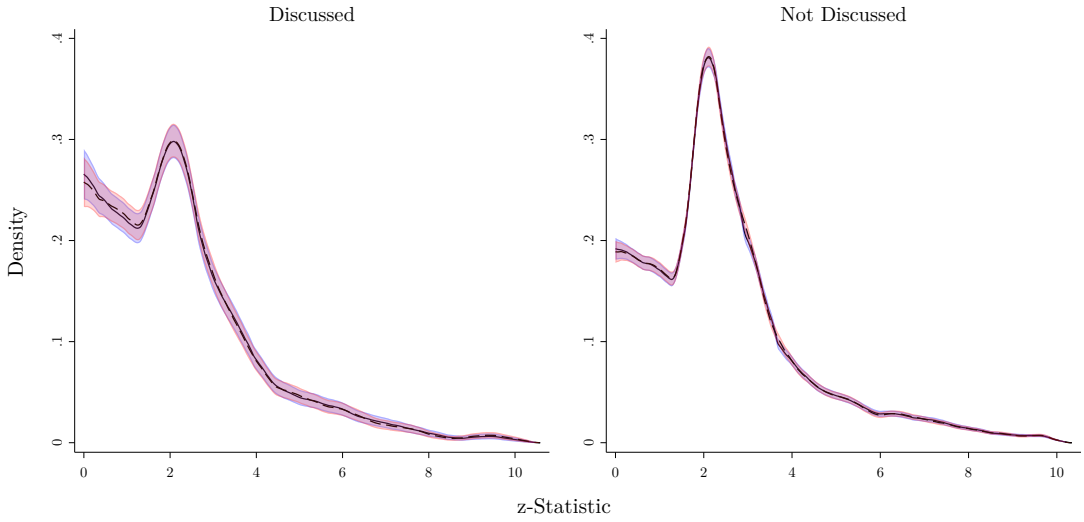
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by participant number quartile e.g., the top left panel contains test statistics using 20 to 189 participants. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A8: z-Statistics: Article Weights and Power Discussion



Notes: This figure displays a histogram of test statistics for $z \in [0, 10]$ by statistical power status. Bins are 0.1 wide. Solid line and blue confidence intervals correspond to unweighted results. Dashed line and red confidence intervals correspond to article weights. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A9: z-Statistics: De-Rounding by Power Discussion



Notes: This figure displays a histogram of test statistics for $z \in [0, 10]$ by discussion of power. Bins are 0.1 wide. Solid line and blue confidence intervals correspond to unchanged results. Dashed line and red confidence intervals correspond to de-rounded test statistics (following Brodeur et al. (2016)).

Figure A10: Proportion of Articles Reporting Cost by Year

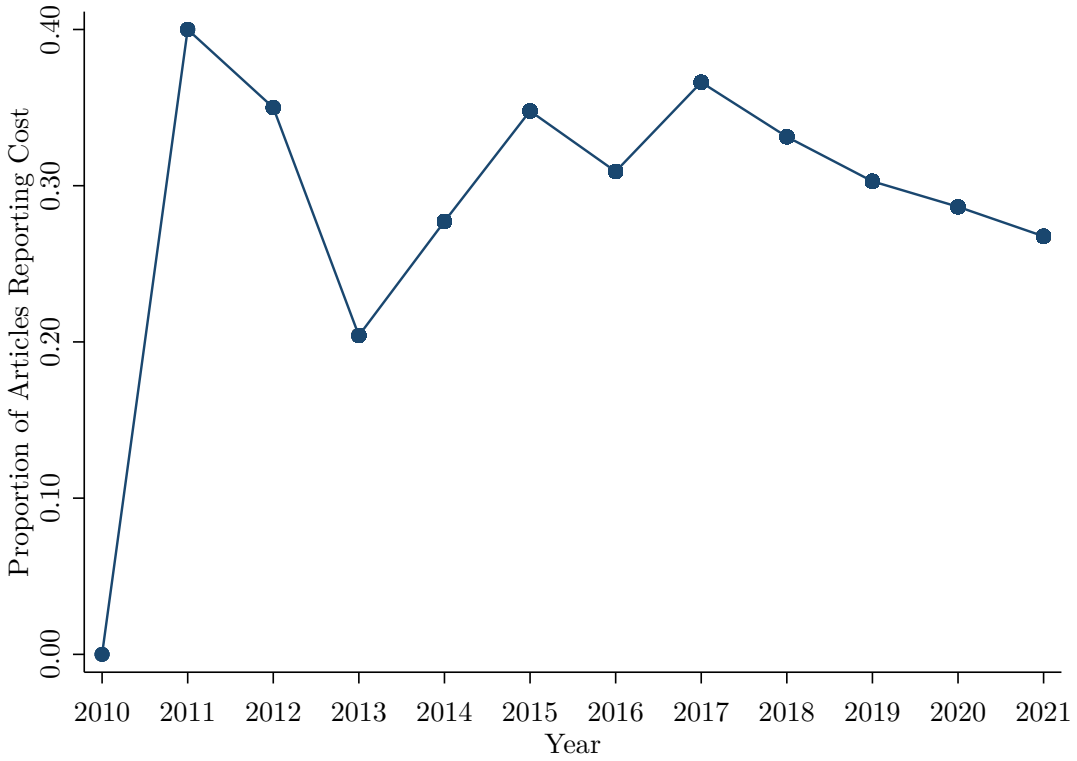
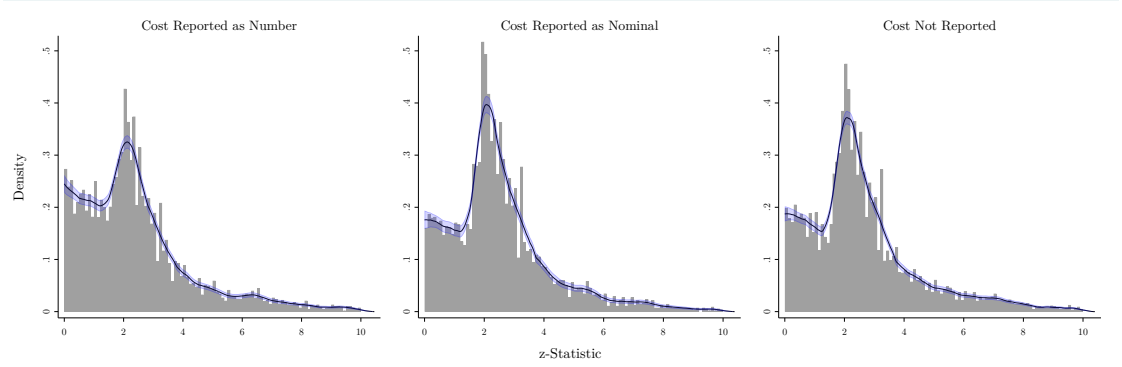


Figure A11: z-Statistics by Cost Reporting



Notes: This figure displays histograms of test statistics for three subsamples of articles: costs are reported (left), costs are reported as nominal (center) and costs not reported (right). $z \in [0, 10]$. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A12: Total Participant Remuneration per Experiment

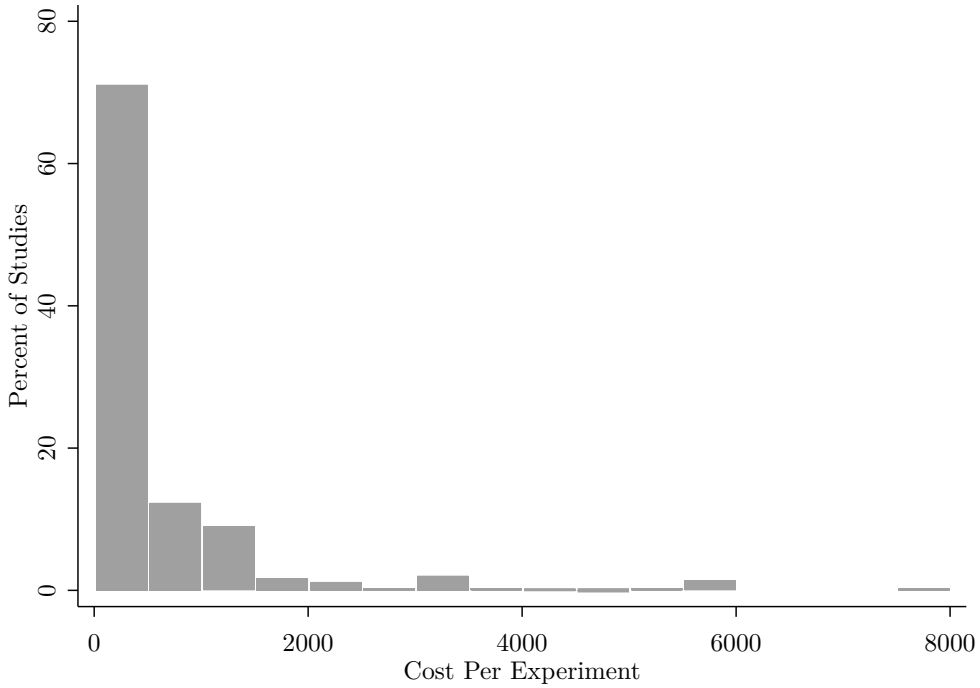


Figure A13: Remuneration per Participant

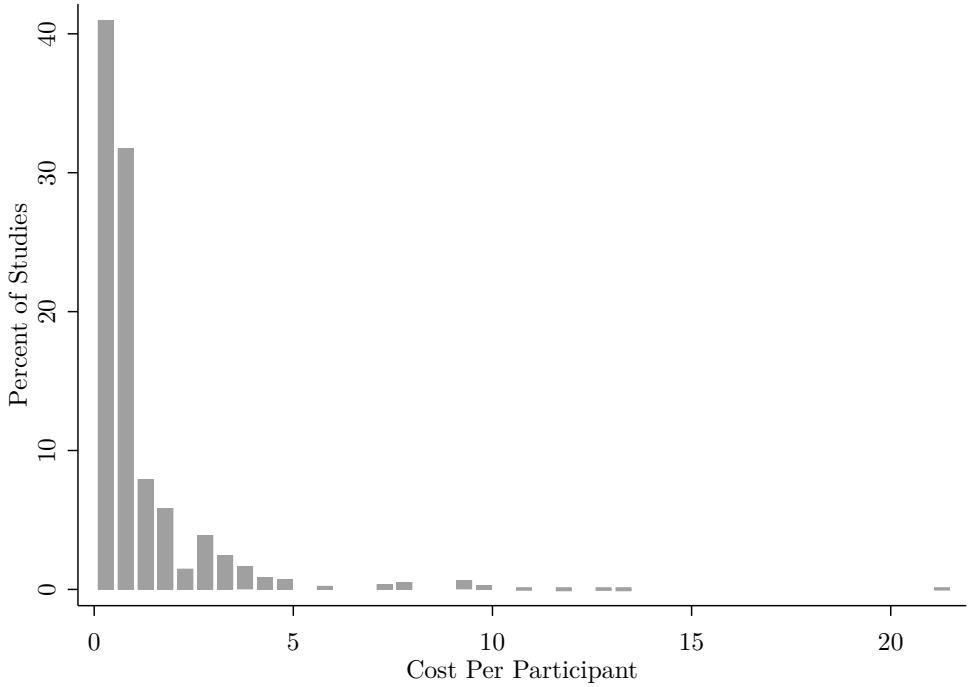


Figure A14: Mean Total Participant Remuneration per Experiment by Year

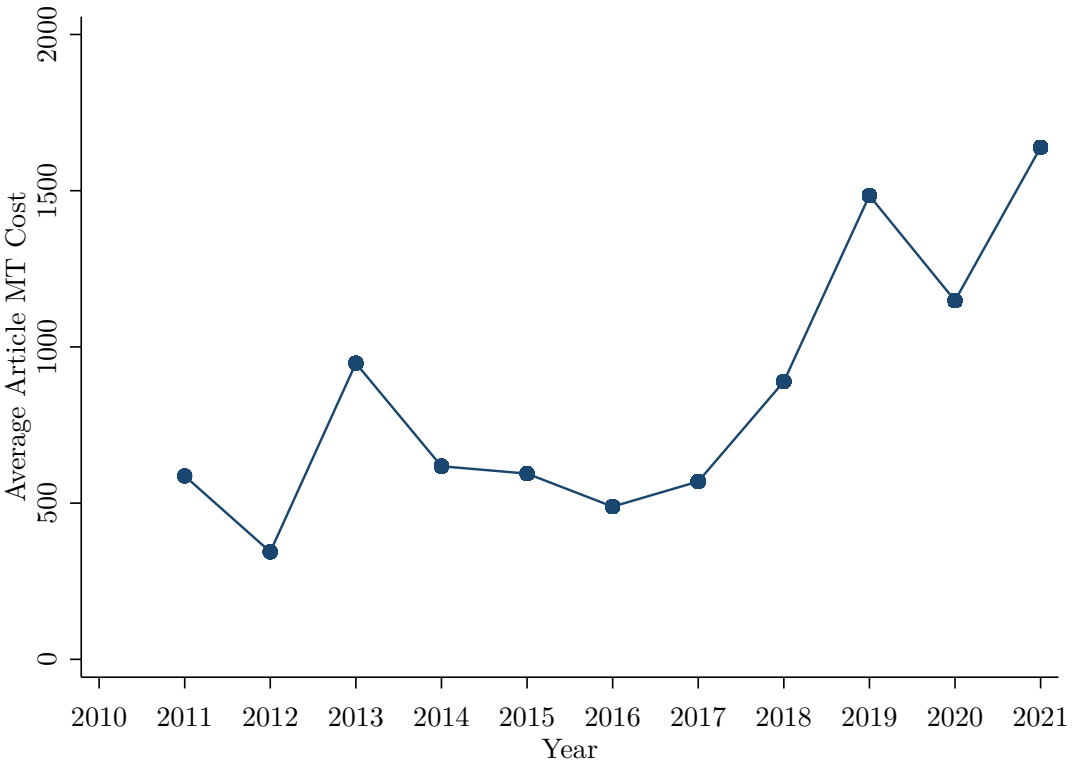


Figure A15: Mean Total Participant Remuneration per Experiment by Field

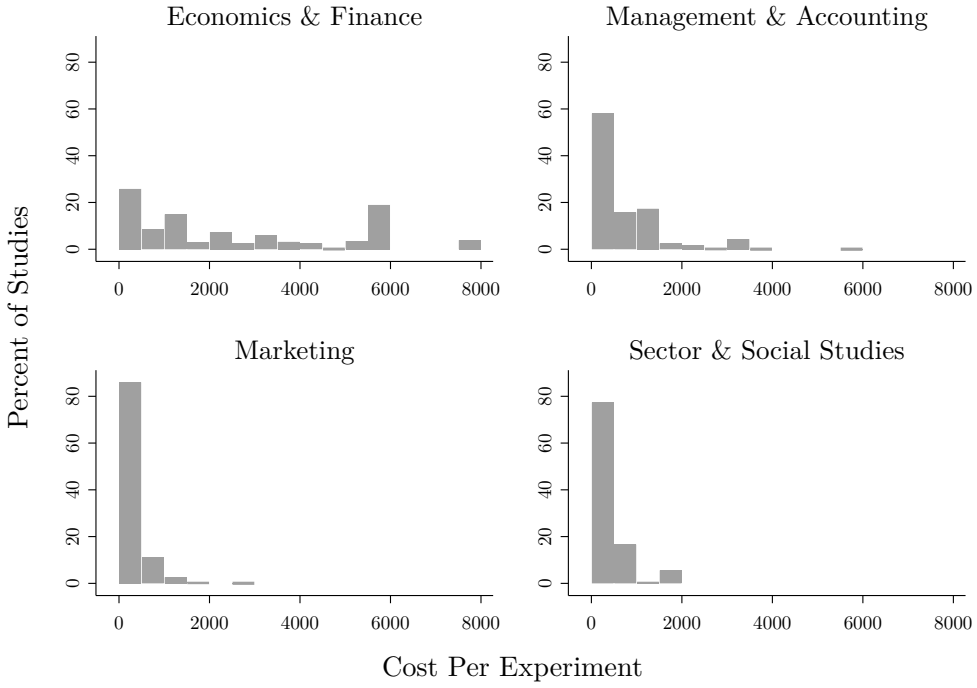


Figure A16: Remuneration per Participant by Field

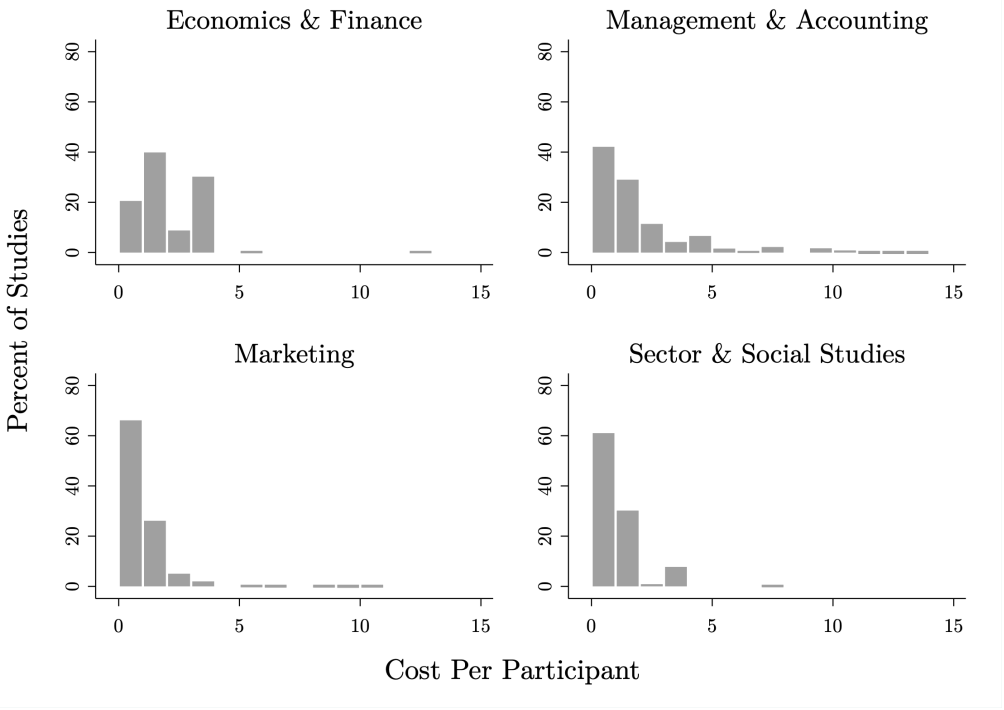
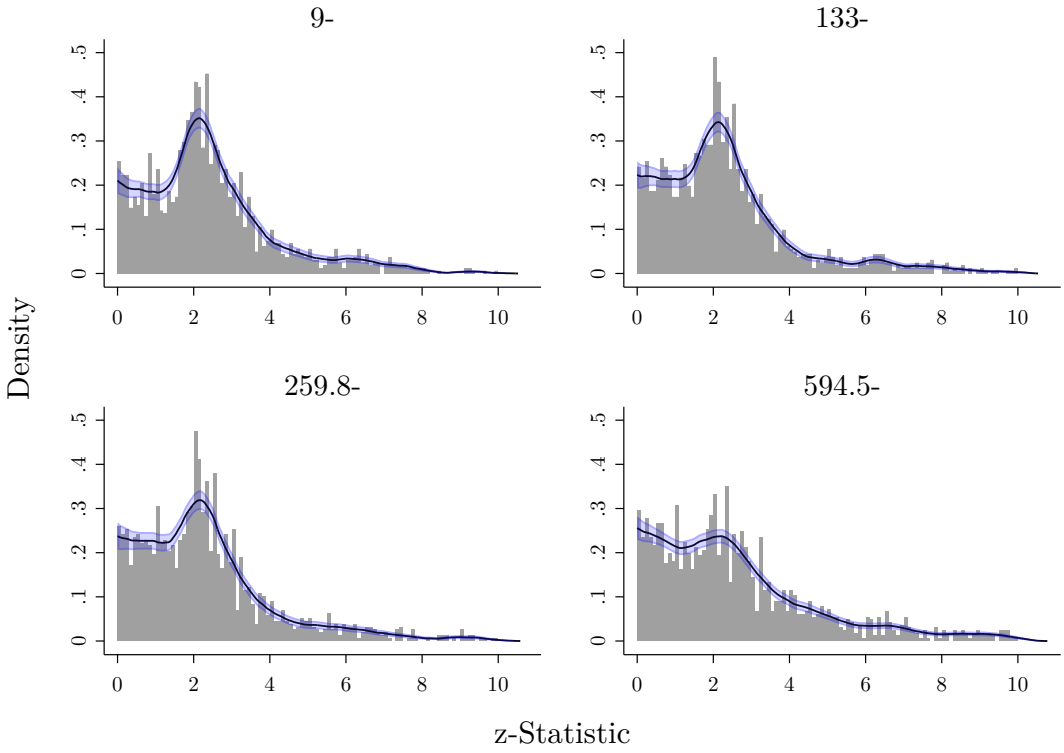


Figure A17: z-Statistics by Total Participation Remuneration by Quartile



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by experiment cost quartile e.g., the top left panel contains test statistics costing 9 to 133. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

12 Appendix Tables

Table A1: Summary Statistics by Journal

Journals	Broad Field	Articles	Tests
	(1)	(2)	(3)
Academy of Management Journal	Management	17	237
Accounting Organizations & Society	Management	8	252
Accounting Review	Management	19	469
Administrative Science Quarterly	Management	2	6
American Economic Review	Econ & Finance	5	215
American Journal of Sociology	Sector & Social	3	75
American Sociological Review	Sector & Social	9	98
Annals of Tourism Research	Sector & Social	12	133
British Journal of Management	Management	2	65
Business Ethics Quarterly	Management	4	62
Contemporary Accounting Research	Management	8	271
Economic Journal	Econ & Finance	2	143
Entrepreneurship, Theory and Practice	Management	1	23
European J. of Operational Research	Management	2	42
Human Relations	Management	3	17
Human Resource Management	Management	6	156
Human Resource Management Journal	Management	1	2
Information Systems Research	Management	7	108
Intl. J. of Research in Marketing	Marketing	35	1,078
Journal of Accounting Research	Management	1	26
Journal of Accounting and Economics	Management	1	3
Journal of Business Venturing	Management	2	3
Journal of Consumer Psychology	Marketing	110	2,347
Journal of Consumer Research	Marketing	313	6,761
Journal of Econometrics	Econ & Finance	1	38
Journal of Management	Management	12	100
Journal of Mgmt Information Systems	Management	9	177
Journal of Management Studies	Management	1	2
Journal of Marketing	Marketing	74	1,710
Journal of Marketing Research	Marketing	115	2,891
Journal of Operations Management	Management	5	120
Journal of Political Economy	Econ & Finance	1	4
Journal of Product Innovation Mgmt	Management	1	1
J. of Public Admin Research & Theory	Management	6	106
Journal of Service Research	Sector & Social	23	437
Journal of Travel Research	Sector & Social	31	498
J. of the Academy of Marketing Science	Marketing	27	737
J. of the European Econ Association	Econ & Finance	4	206
Leadership Quarterly	Management	18	317
MIS Quarterly	Management	3	27
Management Science	Management	30	1,171
Marketing Science	Marketing	11	189
Organization Science	Management	15	412
Organizational Research Methods	Management	2	19
Production and Operations Management	Management	5	88
Public Administration Review	Management	3	69
Research Policy	Management	1	7
Review of Accounting Studies	Management	2	97
Review of Economic Studies	Econ & Finance	3	108
Review of Economics and Statistics	Econ & Finance	2	24
Review of Financial Studies	Econ & Finance	2	21
Risk Analysis	Sector & Social	24	475
Strategic Entrepreneurship Journal	Management	1	18
Strategic Management Journal	Management	3	41
Tourism Management	Sector & Social	23	295

Notes: This table alphabetically presents summary statistics for each journal that contributed at least one test statistic to sample

Table A2: Results from Application of Elliott et al. (2022)

Sample	Bin.	Discont.	CS1	CS2B	LCM	Obs in [0.04,0.05]	Total
Full Sample	0.000	0.000	0.000	0.000	0.007	1181	17277
By Field							
Econ. & Finance	0.500	0.368	888	888	0.985	13	508
Manag. & Account.	0.294	0.001	0.007	0.000	0.967	166	3089
Marketing	0.000	0.000	0.000	0.000	0.010	902	12164
Sector & Social	0.136	0.016	0.147	0.000	0.956	100	1516
By Participants							
Below Median	0.002	0.000	0.000	0.000	0.039	655	8205
Above Median	0.002	0.000	0.000	0.000	0.153	504	8591
By Power Discussion							
Not Discussed	0.000	0.000	0.000	0.000	0.027	1049	15344
Discussed	0.049	0.123	0.021	0.000	0.631	132	1933
By Experiment Cost							
Below Median	0.650	0.000	0.011	0.000	0.468	168	2519
Above Median	0.031	0.012	0.003	0.000	0.345	139	2474

Notes: Each panel is a direct application of Elliott et al. (2022)'s p-hacking test battery to a sub sample. Section 4.1 discusses the p-curve histograms and included tests in detail.

Table A3: Results from Application of Brodeur et al. (2020)

	Location	Scale	[0.00,1.65)	[1.65,1.96)	[1.96,2.58)	[2.58,5.00)	[5.00,10.00]
Full Sample	2	1.6	-0.161	0.009	0.095	0.050	0.007
By Field							
Econ. and Finance	1	1.6	0.007	0.007	0.003	-0.027	0.011
Manag. and Account.	1	1.2	-0.174	0.013	0.091	0.056	0.014
Marketing	3	1.9	-0.107	0.005	0.085	0.011	0.006
Sector & Social	2	1.7	-0.117	-0.002	0.069	0.045	0.006
By Participants							
Below Median	3	1.7	-0.148	0.013	0.103	0.025	0.007
Above Median	2	1.9	-0.066	-0.006	0.058	0.013	0.000
By Power Discussion							
Not Discussed	2	1.6	-0.172	0.010	0.100	0.054	0.008
Discussed	3	1.9	0.002	-0.006	0.031	-0.034	0.007
By Experiment Cost							
Below Median	2	1.4	-0.191	0.019	0.107	0.057	0.008
Above Median	2	1.7	-0.062	-0.005	0.047	0.014	0.006

Notes: Each panel is a direct application of Brodeur et al. (2020)'s excess test statistics methodology to a sub sample. For each of the statistical significance intervals, the table displays the difference between the observed and calibrated-counterfactual distribution. Section 4.2 discusses the method in detail.

Table A4: Caliper Test for Statistical Significance at the 10 Percent Level

	(1)	(2)	(3)	(4)	(5)	(6)
Management & Account	0.134	0.119	0.142	0.121	0.112	0.122
	(0.042)	(0.047)	(0.058)	(0.076)	(0.053)	(0.084)
Marketing	0.161	0.143	0.161	0.156	0.164	0.141
	(0.045)	(0.050)	(0.059)	(0.078)	(0.062)	(0.087)
Sector & Social Sci.	0.144	0.138	0.176	0.164	0.115	0.155
	(0.050)	(0.054)	(0.064)	(0.081)	(0.065)	(0.090)
Year > 2015	0.012					
	(0.017)					
Journal 4*	-0.017	-0.002	-0.014	-0.045	-0.043	-0.038
	(0.021)	(0.022)	(0.025)	(0.033)	(0.026)	(0.047)
Power Discussed		-0.038	-0.045	-0.062	-0.067	-0.075
		(0.022)	(0.026)	(0.032)	(0.024)	(0.033)
Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Report Text/Table	Y	Y	Y	Y	Y	Y
Number of Authors	Y	Y	Y	Y	Y	Y
Journal FE		Y	Y	Y	Y	Y
Observations	6,826	6,826	5,213	3,098	6,826	3,098
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]	[1.96±0.50]	[1.96±0.20]
Weight Articles					Y	Y

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A5: Caliper Test for Statistical Significance at the 1 Percent Level

	(1)	(2)	(3)	(4)	(5)	(6)
Management & Account	0.024	0.024	0.001	-0.084	0.023	-0.088
	(0.065)	(0.065)	(0.076)	(0.089)	(0.062)	(0.097)
Marketing	0.039	0.037	0.006	-0.101	0.062	-0.086
	(0.065)	(0.067)	(0.077)	(0.090)	(0.065)	(0.093)
Sector & Social Sci.	0.082	0.078	0.080	0.104	0.004	0.155
	(0.070)	(0.070)	(0.083)	(0.094)	(0.067)	(0.104)
Year > 2015	0.049					
	(0.018)					
Journal 4*	0.041	0.040	0.067	-0.056	0.042	0.033
	(0.024)	(0.025)	(0.029)	(0.037)	(0.036)	(0.042)
Power Discussed		-0.009	-0.009	-0.020	-0.041	-0.085
		(0.026)	(0.033)	(0.041)	(0.038)	(0.051)
Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Report Text/Table	Y	Y	Y	Y	Y	Y
Number of Authors	Y	Y	Y	Y	Y	Y
Journal FE		Y	Y	Y	Y	Y
Observations	5,906	5,906	4,089	2,318	5,906	2,318
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.35]	[2.58±0.20]	[2.58±0.50]	[2.58±0.20]
Weight Articles					Y	Y

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.