Needed in Empirical Social Science: Numbers [‡]

Aaron Edlin[§] Michael Love[¶]

July 19, 2022

Abstract

Knowing the magnitude and standard error of an empirical estimate is much more important than simply knowing the estimate's sign and whether it is statistically significant. Yet, we find that even in top journals, when empirical social scientists choose their *headline results*—the results they put in abstracts—the vast majority ignore this teaching and report *neither* the magnitude nor the precision of their findings. They provide no numerical headline results for 63% ± 3% of empirical economics papers and for a whopping 92% ± 1% of empirical political science or sociology papers between 1999 and 2019. Moreover, they essentially never report precision (0.1% ± 0.1%) in headline results. Many social scientists appear wedded to a null hypothesis testing culture instead of an estimation culture. There is another way: medical researchers routinely report numerical magnitudes (98% ± 1%) and precision (83% ± 2%) in headline results. Trends suggest that economists, but not political scientists or sociologists, are warming to numerical reporting: the share of empirical economics articles with numerical headline results doubled since 1999, and economics articles with numerical headline results get more citations (+19% ± 11%).

[‡] We thank David Romer, John Donohue, Jonathan Kolstad, Andrew Gelman, Jonah Gelbach, Joseph Farrell, Scott Hemphill, Louis Kaplow, Bentley MacCleod, Andres Rodriguez, Catherine Roux, Christian Thoeni, Steven Tadelis, Stefano DellaVigna, Armen Schmutzler, and the participants at the Berkeley Initiative for Transparency in the Social Sciences 2021 annual conference for helpful comments, and David Romer for suggesting the title. For excellent research assistance, we thank Andrew Episkopos, Joshua Miner, Yifan Zhang, Arjun Bharat, Ishika Chawla, Yangchao Huang, Mary Kruberg, Keon Niknejad, Fuzail Shakir, Emma Sroat, and Kevin Wang.

[§] Department of Economics, and School of Law, University of California, Berkeley E-mail: aedlin@berkeley.edu

[¶] Department of Economics, University of California, Berkeley. E-mail: michael.love@berkeley.edu

1 Introduction

Most econometricians and other applied statisticians have, we suspect, long taught that knowing the magnitude and standard error (precision) of an empirical estimate is much more important than simply knowing the estimate's sign and whether it is statistically significant. Not only do magnitude and precision reveal statistical significance, but they reveal substantive significance and the uncertainty surrounding the estimate. They also avoid arbitrary distinctions between p=.049 and p=.051.

Despite these advantages, we find that when empirical social scientists choose their *headline results*—the results they put in abstracts—the vast majority ignore this teaching and report *neither* the magnitude nor the precision of their findings. In fact, as Figure 1 illustrates, authors provide no numerical magnitudes for headline results in 63% ± 3% of empirical economics papers and a whopping 92% ± 1% of empirical political science or sociology papers in the leading journals we study. (We report point estimates with a 90% confidence interval using "±.") Moreover, social scientists almost never provide confidence intervals or standard errors as part of headline results (0.1% ± 0.1%). We interpret this finding as implying that social scientists almost never think that the precision of their estimates is important enough to warrant placement in an abstract. In stark contrast, in medicine, which we use as a comparison discipline, almost all articles emphasize estimation by providing numerical magnitudes for headline results and the vast majority include precision measures as well (98% ± 1%, and 83% ± 2% respectively).

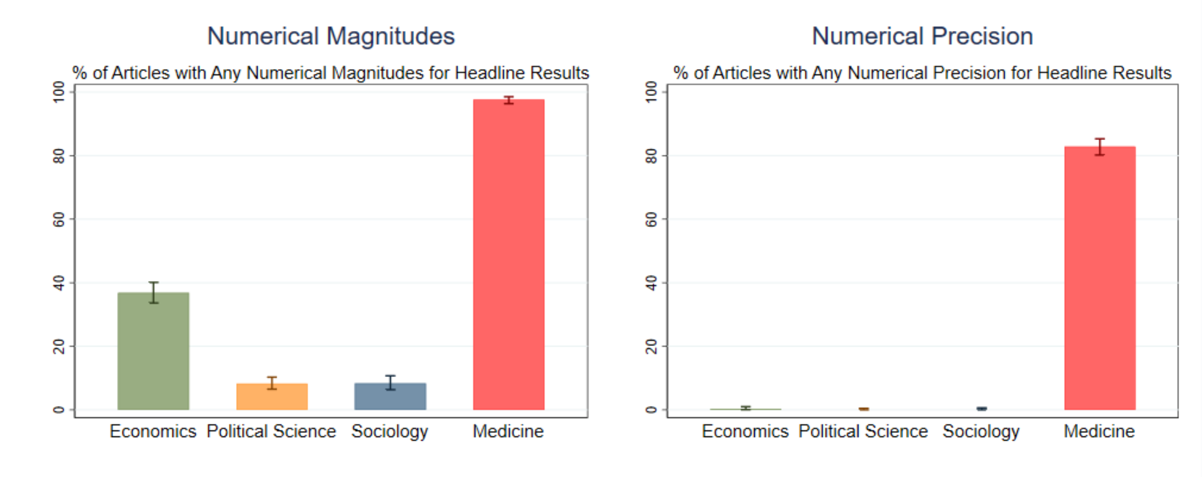


Figure 1: Headline Results with Numerical Magnitude or Numerical Precision

We examined the headline results of 2370 empirical papers in three leading journals in each of economics, political science, sociology, and for comparison medicine, over the period 1999-2019. We focus on abstracts because they reveal what authors think are the most important takeaways for readers.

Compare a typical social science headline result,

> *...we find that an increase in mortgage purchases by [federal] agencies boosts mortgage lending, in particular refinancing, and lowers mortgage rates.*[1]

to a typical medical headline result:

> *The estimated overall survival at 42 months was 70.2% (95% confidence interval [CI], 63.5 to 76.0) in the ribociclib [treatment] group and 46.0% (95% CI, 32.0 to 58.9) in the placebo group ....*[2]

Medical scientists present the magnitude and precision of their empirical findings front-and-center. In contrast, not only do social scientists invariably exclude precision from headline results, but they often don't even report magnitude; instead, they frequently sell their findings as what we call *sign* results, emphasizing only the sign or direction of the estimated effect—like the "boosts mortgage lending" result above.

The phenomenon is problematic because magnitudes matter and sign results create ambiguity. Consider the finding that the "performance of Italian firms that sent their managers to the US increased."[3] The increase might be large, in which case the statement is fair, but not particularly informative. On the other hand, if the increase were trivial (but statistically significant), then the statement would be technically accurate but misleading; a statement of no effect or negligible effect would be closer to the mark in that case. The implications of the two possibilities are opposite: if the estimate is large, then more managers should be sent, and if it is small then fewer should be. In this sense, sign results are ambiguous.

Sign results can also make entirely consistent studies appear contradictory. Consider how two hypothetical studies estimating the elasticity of earnings with respect to class size as -4 ±5 and -6 ±5 might report their findings:

> *Study 1: We do not find an effect of class size on earnings.*
>
> *Study 2: We find that earnings increase as class size falls.*

Described this way, the two studies seem contradictory. But both studies estimate enormous elasticities of earnings with respect to class size, have similar precision, and reinforce each other's conclusions. Their similarity becomes apparent if magnitudes and standard errors or confidence intervals are reported instead of providing sign reports.

Despite the problems with sign results, half of the articles in political science and sociology have only sign headline results ($51\% \pm 3\%$ and $54\% \pm 4\%$, respectively). Economics does better, but still over a quarter of articles have only sign results ($26\% \pm 3\%$). Unfortunately, the high proportion of sign results has not gone down over the time period we studied across these disciplines, and has even increased in sociology and political science.

[1] Fieldhouse, Mertens, and Ravn (2018).

[2] Im, Lu, Bardia, Harbeck, Colleoni, Franke, Chow, Sohn, Lee, Campos-Gomez, et al. (2019).

[3] Giorcelli (2019).

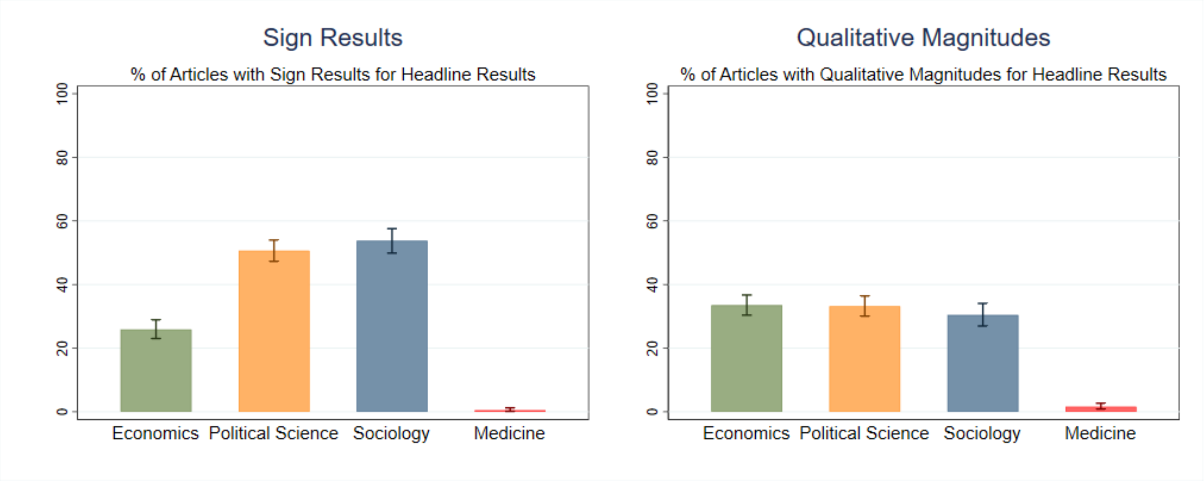


Figure 2: Articles with only Sign Headline Results or at most Qualitative Magnitudes

Many social scientists take an intermediate option and report their headline result with what we call *qualitative* magnitudes, saying that X is *large*, that X is *substantially* smaller than Z, or the like. This approach goes part of the way toward dealing with the ambiguity issue, but still conceals much easily deliverable information. Roughly one third of the articles in each of the three social science disciplines report headline results with qualitative magnitudes but not numerical magnitudes (33% ± 3%, 33% ± 3% and 30% ± 3% in economics, political science and sociology, respectively).

So, why do most social scientists avoid numerical headline results and instead choose a sign or qualitative headline? Spin is one possible explanation for sign results: some no doubt find small effects and fear that highlighting magnitude would make their paper less impressive. It is possible, for example, that if a paper finding that golfers are "more likely to make puts for par or over par than puts for birdie or eagle," told readers up front that golfers are only "0.2% more likely" to make such puts that readers or referees might be less impressed.[4] Frequently, however, spin can not be the motive. Another AER paper, for example, tells us that "contracts with termination options are more common when research is non-contractible" instead of wowing us with the size of its finding that termination options are 97% more common.[5] We suspect that social science writers are typically not motivated by spin, but instead fail to emphasize numerical magnitudes because of a deeply-seated cultural tradition of hypothesis testing in which researchers strive to reject the null hypothesis of no effect and no difference. The pull of this tradition is so strong that one of us emphasized sign results even while working on this paper: Edlin, Roux, Schmutzler, and Thöni (2019) write that "[alternative policies] increase consumer welfare," instead of "[alternative policies] increase consumer welfare *by 20%*."

Some colleagues have suggested to us that social scientists care deeply about the

[4] Pope and Schweitzer (2011)

[5] Lerner and Malmendier (2010)

magnitude and precision of their estimates, but that space constraints in abstracts prevent them from including estimates with precision in their headline results. One colleague, for example, bet that at least 2/3 of economics articles would provide a numerical precision *somewhere* in the introduction, because space is less of a factor in an introduction than an abstract. It turned out only 2% ± 2% do.[6] Perhaps even more surprising, we find that if point estimates are not described in the abstract, they are highly unlikely to be reported in the introduction either: among economics articles reporting only the sign of empirical results in the abstract, only 26% ± 7% reported a numerical magnitude in the introduction.[7] Many social scientists simply do not care deeply enough about magnitude and precision to report them prominently.

Of course, the vast majority of social science papers provide both magnitudes and some measure of precision (at a minimum a t-statistic or p-value) *somewhere* in the paper, at least in a table. But these numerical results usually are not what social science authors choose to emphasize.

Here though is what we consider good news: economists, anyway, may be gradually coming to the conclusion that numerical headline results are preferable to sign or qualitative results, as they provide substantially more headline numbers today than two decades ago. We estimate that about half of the articles published in leading economics journals in 2018 and 2019 have at least one numerical magnitude in their headline results (58% ± 11%), more than doubling the proportion from 20 years prior (25% ± 9%). By comparison, we observe substantially smaller changes in the prevalence of numerical headline results in other disciplines, and these changes are imprecisely estimated (over the same period we estimate a 6p.p. ± 6p.p. increase in political science, 4p.p. ± 7p.p. in sociology, and 3p.p. ± 3p.p. in medicine).

Citation analysis suggests that the growth of numerical magnitudes in headline results could reflect authors' reactions to incentives. We estimate that economics articles with numerical magnitudes in their headline results get 19% ± 11% more Google Scholar citations than articles without numerical headline results (after controlling for journal and year). Political science and sociology also have positive point estimates for extra citations associated with the inclusion of numerical magnitudes in headline results, but as these disciplines so seldom report numerical magnitudes in headline results, these effects are imprecisely estimated (+1% ± 29% and +24 ± 30%, respectively).

Our paper contributes to a current discussion in the statistics and applied statistics community about how to interpret and present empirical findings. There is a longstand-

[6] We selected 100 random economics articles from our dataset whose abstracts reported numerical magnitudes, but not precision. We then read through the introductions of these articles to see if a measure of precision was reported there.

[7] We selected 100 random economics articles from our dataset whose abstracts reported sign headline results. We then read through the introductions of these articles to see if numerical point estimates for any headline results were reported there.

ing culture of null hypothesis significance testing ("NHST") that usually privileges the null of $\beta = 0$ (see, e.g., Nickerson, 2000). Growing opposition to NHST has argued that it is more important to report estimates and to recognize uncertainty regardless of the statistical significance of a coefficient with respect to a null of 0. In particular, a recent statement of the American Statistical Association says that "some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals." (Wasserstein and Lazar, 2016). Some statisticians have gone farther, concluding that the term "statistical significance" should be abandoned along with "significantly different," and that we must embrace uncertainty by "accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate." (Wasserstein, Schirm, and Lazar, 2019).[8] Economists have taken steps in this direction as well: in roughly the middle of the period we study, the American Economic Review amended its style guide to prohibit the use of asterisks to denote statistically significant results.[9] Our paper contributes to this discussion by offering the first large scale attempt we know of to measure across the social sciences the extent to which authors emphasize estimation (magnitudes) as well as uncertainty (precision), and how these vary by discipline.

Our finding on the scarcity of both numerical magnitudes and precision complements Romer (2020), who recently found that of 105 economics papers studied from 2019, 63 percent do not put either standard errors or confidence intervals anywhere in the text of their papers. Our paper looks beyond Romer's focus on precision, and finds that economists usually don't even emphasize the numerical magnitudes that they estimate. But we do find that economists emphasize magnitudes much more than their fellow social scientists, and much more than economists of two decades ago.

If the absence of numbers is a malady, we think the cure is for social science journals to follow medical journals and adopt style guides that require or encourage reporting numerical results with precision in abstracts, introductions, and elsewhere, unless there is a good reason not to do so. To be sure, the nature of medical research differs from social science research and this may justify different reporting standards: for example, instead of mandating numerical reporting, social science guidelines might encourage numerical headline results, allow qualitative headline results, and discourage or even prohibit sign headline results. Without editorial guidelines of some type, it could take almost a century at current rates for the social sciences to stop being so number-shy and look more like medicine.

[8] Gelman and Stern (2006) point out that the difference between statistically significant and insignificant empirical estimates is often not statistically significant

[9] The style guide instructs authors: "Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses." AER Guidelines for Accepted Articles, downloaded October 5, 2021 from https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide#IVH.

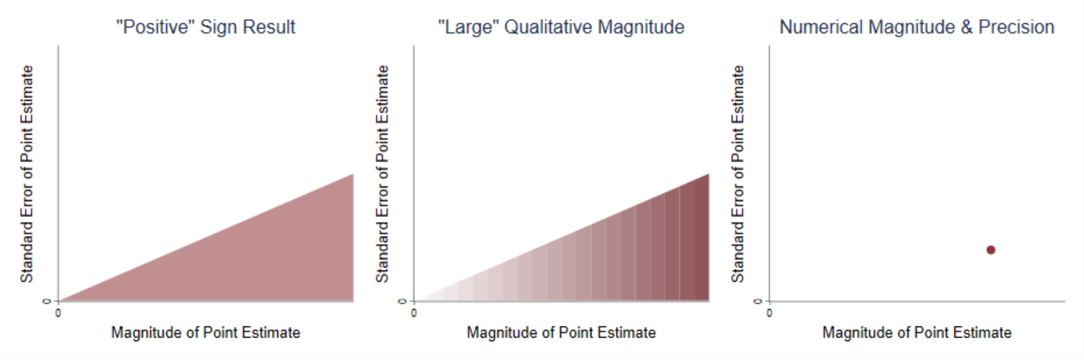


Figure 3: Information Conveyed by Headline Results

The remainder of the article is organized as follows. Section 2 provides a little more motivation for our study. Section 3 discusses our methodology. Section 4 presents results. Section 5 concludes. We also include in the Appendix a model style guide for social science journals that come to believe that both magnitude and precision are the vital outcomes of studies that authors should be reporting as headline results.

2 Motivation

2.1 The Value of More Informative Headline Results

Headline results can shape any reader's understanding of a paper's import, but this is especially true for those outside the specialty or outside academia who may choose to read only the headline results or may only have access to headline results. [10] Yet despite these high stakes, as Figure 3 illustrates, authors who give only sign or qualitative headline results convey vague and limited information to readers.

Figure 3 demonstrates the vagueness of most social science headline results by illustrating what information they convey to the reader. The left panel represents the information conveyed by a typical sign headline result, such as “The tax cut increased hours worked.” Between the positive direction and (we will assume) the statistical significance of the finding, the reader knows only that the estimated effect is somewhere in the shaded region—a very imprecise conclusion. The center panel presents a headline

[10] Many readers, especially those in the developing world, only have access to abstracts. See Hopewell, Clarke, Moher, Wager, Middleton, Altman, and Schulz (2008): “In 2006, Arthur Amman, President of Global Strategies for HIV Prevention, made a disquieting remark: 'I recently met a physician from southern Africa, engaged in perinatal HIV prevention, whose primary access to information was abstracts posted on the internet. Based on a single abstract, they had altered their perinatal HIV prevention program from an effective therapy to one with lesser efficacy. Had they read the full text article they would have undoubtedly realized that the study results were based on short-term follow-up, a small pivotal group, incomplete data, and unlikely to be applicable to their country situation. Their decision to alter treatment based solely on the abstract's conclusions may have resulted in increased perinatal HIV transmission”'

result with a qualitative magnitude, such as "The tax cut substantially increased hours worked," also presumably statistically significant. This qualitative formulation is only marginally more helpful: the reader has a vague notion that the effect is large (reflected in the gradient shading) and is presumably statistically significant, but has no certainty about either the estimated magnitude or precision. In contrast, the right panel presents the information conveyed from simply reporting a numerical magnitude with accompanying precision. In this case, the reader need not guess—the main facts about the estimate are conveyed explicitly.

Given the apparent appeal of the clarity in the third panel, it is surprising that almost no social science papers (fewer than $0.3\% \pm 0.4\%$ in economics and even fewer in political science and sociology) provide this information as part of their headline results.[11] We are disturbed that many papers do not even provide qualitative headline results and only provide sign headlines, essentially hiding the ball as illustrated in the left-most panel. The prevalence of sign headlines probably reflects the long tradition of null hypothesis significance testing in empirical work in the the social sciences and possibly a tradition of comparative statics in theory work.

The most common defense of sign results that we hear is a desire to be terse, but this does not adequately justify the scarcity of numbers in abstracts. For example it would take only 6 additional characters to rewrite the abstract:

> It is well documented that voter turnout is lower among persons who grow up in families from a low socioeconomic status compared with persons from high-status families. . . . We move past previous studies, however, and show that the reform nevertheless contributed to narrowing the voting gap between individuals of different social backgrounds **by raising turnout among those from low socioeconomic status households**. . . . (emphasis added) [12]

as:

> It is well documented that voter turnout is lower among persons who grow up in families from a low socioeconomic status compared with persons from high-status families. . . . We move past previous studies, however, and show that the reform nevertheless contributed to narrowing the voting gap between individuals of different social backgrounds **by raising turnout 3.1pp among those from low socioeconomic status households**. . . .

[11] For ease of notation, we report confidence intervals as $\pm X$ representing a 90% confidence interval assuming a t-distribution, rather than the asymmetric Clopper-Pearson exact confidence intervals for a binomial distribution. As a result, sometimes the reported interval erroneously extends below 0% or above 100%.

[12] Lindgren, Oskarsson, and Persson (2019).

Likewise, adding precision typically will require a little extra space but not much. In the above abstract, for example, we would suggest replacing 3.1pp with 3.1pp ± 2.2pp, or putting a confidence interval 3.1pp [CI: 0.9pp,5.3pp].

Sometimes, of course, adding numerical magnitudes can take up noticeably more space if those numbers require an explanation or context. It may also require authors to choose their preferred specification if there are many in the paper or to provide a range of estimates. But we still think allocating space to this task is worthwhile to avoid the ambiguity inherent in sign results. Consider the sign headline result in a leading paper:

> The conventional wisdom for the health care sector is that idiosyncratic features leave little scope for market forces to allocate consumers to higher performance producers. However, we find robust evidence across several different conditions and performance measures that higher quality hospitals have higher market shares and grow more over time.[13]

However robust the evidence, if the size of the effects is small and precisely estimated, then the finding actually supports the conventional wisdom of "little scope" rather than rejecting it. The following sentence could usefully be added: "We find that patient flows shifting to higher quality hospitals increased 30-day survival rates for heart attacks, heart failure and pneumonia by 1%, 0.2%, 0.2% respectively." Adding these numerical magnitudes would take some space, of course, but allows the reader to judge whether the results indeed tend to reject the conventional wisdom or support it.

To choose an example from sociology, where sign results are even more common than in economics, consider the following: "Because landlords operating in poor communities face more risks, they hedge their position by raising rents on all tenants, carrying the weight of social structure into price."[14] Such a sign result eliminates the possibility that landlords substantially lower rents, but allows the possibility that rents were raised by so little that a more informative reporting would be that rents were largely unchanged. How much better it would be to know by what percentage rents rose so we could understand the importance of the finding and better evaluate the subsequent claim that "market strategies of landlords contribute to high rent burdens in low-income neighborhoods."

Including magnitudes and precision is also valuable for "no effect" findings. Consider the headline result that "The treatment has no effect on self-reported voting behavior after the election. . . ."[15] Does "no effect" mean that the result is not statistically significant, so that it is possible the effect is large but imprecisely estimated? Or does it mean that the estimates suggest the effect is small? Certainly we cannot take the report at face value as showing that the treatment has literally no estimated effect—that is impossible. All

[13] Chandra, Finkelstein, Sacarny, and Syverson (2016).

[14] Desmond and Wilmers (2019).

[15] Boas, Hidalgo, and Melo (2019).

ambiguity would be resolved with a numerical magnitude and standard error. Of course sufficiently careful writing could avoid some of these difficulties, but the simplest and most consistent solution is to include numbers in headline results.

We recognize that space is not the only plausible defense for the paucity of numerical headline results in the social sciences. There is also a cost in style and narrative storytelling, as numbers may interrupt the flow of the abstract. Still we think that social scientists should pay that cost (we are scientists after all). And, even in cases where authors find it challenging to distill their findings down to single numerical results, perhaps some movement toward more precision is warranted, such as moving from sign headline results to qualitative, as one of the style guides in our appendix suggests.

Another possible justification not to have numerical headline results in social science is that specification uncertainty could make results less reliable than they appear. A social scientist, perhaps more so than a medical researcher conducting a randomized trial, must make many judgments about the choice of model and methodology to test a hypothesis, and a numerical headline result could give the reader a misleading impression of the definitiveness of the finding. That critique, however, could equally justify not providing a sign result. It also seems overly paternalistic with regard to readers who will probably take all results with a grain of salt but would still like to know the estimated magnitude and the size of the estimated standard error.

Some may minimize the importance of numbers being absent from headline results because an interested reader can find them in the paper. But saying, "go search for the bottom line" reveals little respect for the readers's time.[16] It also turns out that the numbers may not be in the paper at all. Generally magnitudes will be in the paper somewhere, but the precision of the headline estimates, surprisingly, may not. [17]

While we have not been able to figure out any good reason for papers to have only sign headline results, sometimes presenting headline results with qualitative magnitudes could be preferable to numerical ones. Qualitative magnitudes may be superior if the reader needs the authors' help to know if a numerical estimate is large or small when the estimate itself might have little meaning to the typical reader. But even in cases where a numerical result is not ideal, qualitative magnitudes are substantially more informative

[16] It is worth emphasizing here, as reported Figure 6, that the majority of articles with only sign results in the abstract do not report numerical results in the introduction either.

[17] Consider for example the headline result of Aitken and Harrison (1999), that "The net impact of foreign investment, taking into account these two offsetting effects, is quite small." As is nearly universal, the headline result does not include a precision. But it is worth noting that even an interested reader can't simply dig through the paper to find a table, identify the right coefficient and spy the standard error. The problem is that net impact is a derived quantity. The net impact of "DFI is calculated by multiplying the coefficients in the first five rows [of estimated coefficients] by their actual values and then adding them together for each plant," and then summing "this net effect across all plants, weighted by each plant's share of employment," and "averaging across all years." While standard errors for regression coefficients are reported for the main variables, Aitken and Harrison do not report covariances and a variety of other data that would be necessary to calculate standard errors of this complex derived quantity. In any event, it is surely the authors' job, not the readers', to do such calculations.

than sign results, and so our biggest worry is the more than 1/4 of papers in economics and roughly 1/2 in political science and sociology that provide only sign headline results.

2.2 A Point of Comparison: Headline Results in Medical Journals

Abstracts in medical journals, in contrast with those of the social sciences, are jam-packed with numerical point estimates and precision. Medical abstracts in the journals we study typically begin with a brief discussion of the context or methodology of a study before providing rich, detailed numerical results. Consider the following excerpt from an abstract published in the New England Journal of Medicine:

> Infections after placement of cardiac implantable electronic devices (CIEDs) are associated with substantial morbidity and mortality. . . . We conducted a randomized, controlled clinical trial to assess the safety and efficacy of an absorbable, antibiotic-eluting envelope in reducing the incidence of infection associated with CIED implantations. . . . A total of 6983 patients underwent randomization: 3495 to the envelope group and 3488 to the control group. [A serious infection] occurred in 25 patients in the envelope group and 42 patients in the control group (12-month Kaplan–Meier estimated event rate, 0.7% and 1.2%, respectively; hazard ratio, 0.60; 95% confidence interval [CI], 0.36 to 0.98; P=0.04). . . . an antibacterial envelope resulted in a significantly lower incidence of major CIED infections than standard-of-care infection-prevention strategies alone, without a higher incidence of complications.[18]

Discussions such as this one give the reader full view of numerous relevant outcomes and findings. There is no vagueness about the size of effects or the precision of the estimates. All the information is presented front-and-center for the reader.

The prominence of numbers in medical abstracts is no accident. Rather, it is a reflection of coordinated effort over time within the medical academic community. In contrast with the social sciences, researchers and journal editors in medicine have been engaged in a rich discussion about the presentation of headline results for more than two decades. In 1996, a group of doctors, journal editors, and statisticians met in Chicago to produce the "CONSORT Statement" (i.e., Consolidated Standards on Reporting Trials), calling for standardization and transparency in reporting of randomized trials (Begg et al., 1996). The statement, which has been updated regularly since its initial publication, calls for authors reporting an empirical result to include the "estimated effect size and its precision (such as a 95% confidence interval)."[19]

[18] Tarakji et al. (2019).

[19] See Moher et al. (2001), Hopewell et al. (2008), and Moher et al. (2012).

Many medical journals have adopted the abstract reporting guidelines of the CONSORT Statement. The Lancet requires that "[f]or randomised trials, the abstract should adhere to CONSORT[.]"[20] The Journal of the American Medical Association requires authors to "[w]hen possible, present numerical results (e.g., absolute numbers and/or rates) with appropriate indicators of uncertainty, such as confidence intervals."[21] The New England Journal of Medicine requires that "[s]ignificance tests should be accompanied by confidence intervals for estimated effect sizes, measures of association, or other parameters of interest."[22]

3 Data and Methodology

To compile the data used in this investigation, we created a stratified random sample of 2370 articles from top journals in Economics, Political Science, Sociology, and Medicine published between 1999 and 2019. We read the abstracts and then categorized the articles based upon the headline results in the abstract.

3.1 Sampling of Articles

We began by identifying three top journals in each of the four disciplines. We first included the flagship journal of the principal American academic association of each discipline: the American Economic Review, the American Political Science Review, the American Sociological Review, and the Journal of the American Medical Association. We then included the top two journals in each discipline according to the latest SCImago Journal Rankings that met the following two criteria: (1) the journal publishes empirical articles of general interest,[23] and (2) the journal has published continuously since 1999.[24] These criteria identified the following eight additional journals: the Quarterly Journal of Economics and the Journal of Political Economy for Economics, the American Journal of Political Science and the Journal of Politics for Political Science, the Annual Review of Sociology and the Journal of Sociology for Sociology, and the New England Journal of Medicine and the Lancet for Medicine.

To ensure even representation across journals and time, we randomized the order of articles within each journal by year. We then selected the first 10 articles in each journal-

[20] See https://thelancet.com/pb/assets/raw/Lancet/authors/tl-info-for-authors.pdf, accessed 2021-02-01.

[21] See https://jamanetwork.com/journals/jama-health-forum/pages/instructions-for-authors, accessed 2021-02-01.

[22] See https://www.nejm.org/author-center/new-manuscripts, accessed 2021-02-01.

[23] That is, the journal is not a field-specific journal nor tends to present a certain type of article that only applies to a subset of academics in the discipline.

[24] SCImago calculates journal rankings from information in the Elsevier Scopus database using an eigenvector centrality measure that accounts for both the number of citations and the prestige of the citing journal. Rankings were tabulated on 2020-01-19.

year that were "in scope" for the investigation (or until we ran out of articles in that journal year). Altogether, we reviewed 3895 articles yielding a random sample of 2370 in-scope articles.

> **Determination of Scope**: An article is in scope if it has at least one empirical finding among its headline results (i.e., results reported in the abstract) that either is or could be quantified with a numerical magnitude and precision. Otherwise, it is out of scope.

Purely theoretical articles without empirical findings are out of scope, as are case studies, and models parameterized with values to generate magnitude estimates but no precision. Experimental papers are squarely in scope.

This procedure resulted in 630 in-scope articles in Economics, Political Science, and Medicine, and 480 in Sociology (as the Annual Review of Sociology frequently has fewer than 10 in-scope articles in a given year).

3.2 Categorization of Magnitude and Precision

For each in-scope article, we evaluated how the authors report empirical findings in their abstract ("headline results"). The article is categorized as follows:

> **Numerical Magnitude**: The magnitude of at least one empirical finding in the abstract is presented as a number (using words or numerals).
>
> *Example*: "The tax cut increased hours worked by 1 percent."
> *Example*: "The unemployment rate fell by one third."
>
> **Qualitative Magnitude**: The article does not qualify as "numerical" as described above, but the magnitude of at least one empirical finding in the abstract is described in qualitative language such as "big," "small," "substantial," "sizable," "no effect" or using another qualitative description of magnitude.
>
> *Example*: "The tax cut increased hours worked by a small amount."
> *Example*: "We find a substantial decrease in the unemployment rate."
>
> **Sign Result**: The article does not qualify as either "numerical" or "qualitative" as described above, but at least one empirical finding in the abstract is presented as (a) being positive or negative, (b) one variable increasing or decreasing in another variable, or (c) one variable being larger or smaller than another, or something similar.
>
> *Example*: "The tax cut increased the number of hours worked."
> *Example*: "The tax cut had a smaller effect than the direct subsidy."

No Reported Headline Result: The article does not qualify for any category above, but the abstract nevertheless discusses empirical findings.

Example: "We use a new dataset to estimate the effect of the tax cut on hours worked."
Example: "The estimated effect of the policy change on the unemployment rate challenges the findings in prior literature."

We categorized the article based on the precision in headline results as follows:

Numerical Precision: The precision of at least one empirical finding in the abstract is presented as a number (using words or numerals).

Example: "We estimate that the unemployment rate fell by 30% (with a 95% confidence interval from 25% to 35%)."
Example: "We estimate that the unemployment rate fell by 30% (with a standard error of 2.5 percentage points)."
Example: "The tax cut increased hours worked by 1 percent (p=0.015)."[25]

Qualitative Precision: The article does not qualify as "numerical" above, but the abstract describes the degree of precision of at least one empirical finding.[26]

Example: "We estimate a 30% decrease, but with large standard errors."

No Precision Reported: The abstract does not have either a numerical or qualitative discussion of the precision of any empirical finding.[27]

Our methodology is intentionally conservative because we rate an article as numerical if it has any numerical headline result, even if there are other results reported that would fall into the qualitative or sign categories. Thus, our estimate that 63% of economics articles report no numerical headline results serves as a lower bound on the frequency with which no results are actually reported numerically.

3.3 Citation Counts

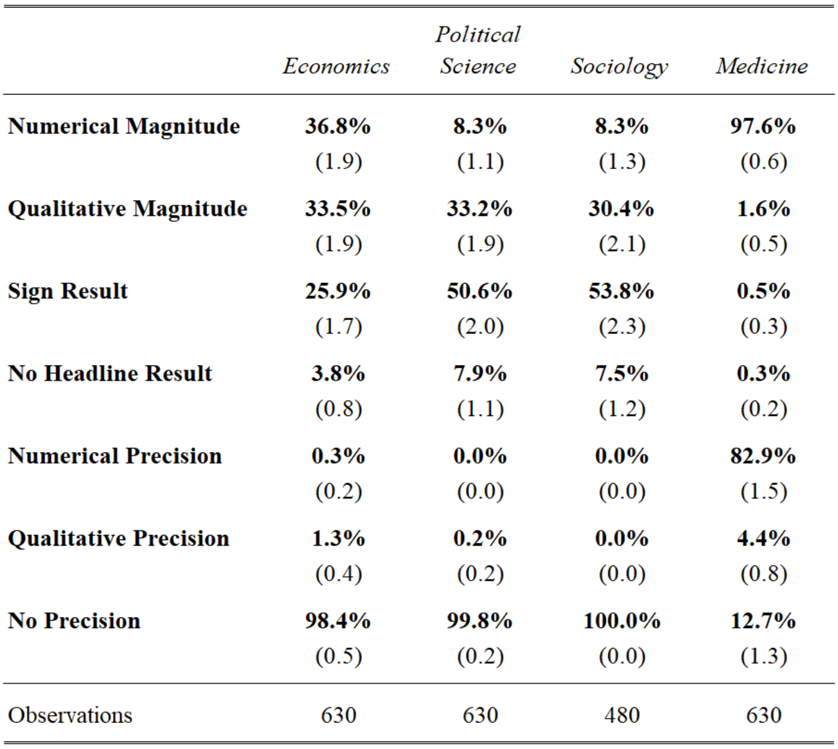
The citation counts data were collected by hand from Google Scholar. To minimize the effect of collection timing on the counts, the data for all citation counts were collected during the same week. More recent articles will have had less time to accumulate citations

[25] Although p=.015 is not a precision per se, it does provide enough information when combined with the magnitude to calculate precision.

[26] We categorize articles that offer numerical magnitudes along with a statement bounding significance (e.g. $p<0.05$) as qualitative, because they are not providing an exact precision nor means to calculate it.

[27] Note that abstracts mentioning "statistically significant" results are categorized as "none," because the statement alone neither describes the precision explicitly nor gives the reader sufficient information to understand the precision. Regardless, this is a rare statement in abstracts.

Table 1: Shares of Articles by Category, by Discipline



	Economics	*Political Science*	*Sociology*	*Medicine*
Numerical Magnitude	**36.8%**	**8.3%**	**8.3%**	**97.6%**
	(1.9)	(1.1)	(1.3)	(0.6)
Qualitative Magnitude	**33.5%**	**33.2%**	**30.4%**	**1.6%**
	(1.9)	(1.9)	(2.1)	(0.5)
Sign Result	**25.9%**	**50.6%**	**53.8%**	**0.5%**
	(1.7)	(2.0)	(2.3)	(0.3)
No Headline Result	**3.8%**	**7.9%**	**7.5%**	**0.3%**
	(0.8)	(1.1)	(1.2)	(0.2)
Numerical Precision	**0.3%**	**0.0%**	**0.0%**	**82.9%**
	(0.2)	(0.0)	(0.0)	(1.5)
Qualitative Precision	**1.3%**	**0.2%**	**0.0%**	**4.4%**
	(0.4)	(0.2)	(0.0)	(0.8)
No Precision	**98.4%**	**99.8%**	**100.0%**	**12.7%**
	(0.5)	(0.2)	(0.0)	(1.3)
Observations	630	630	480	630

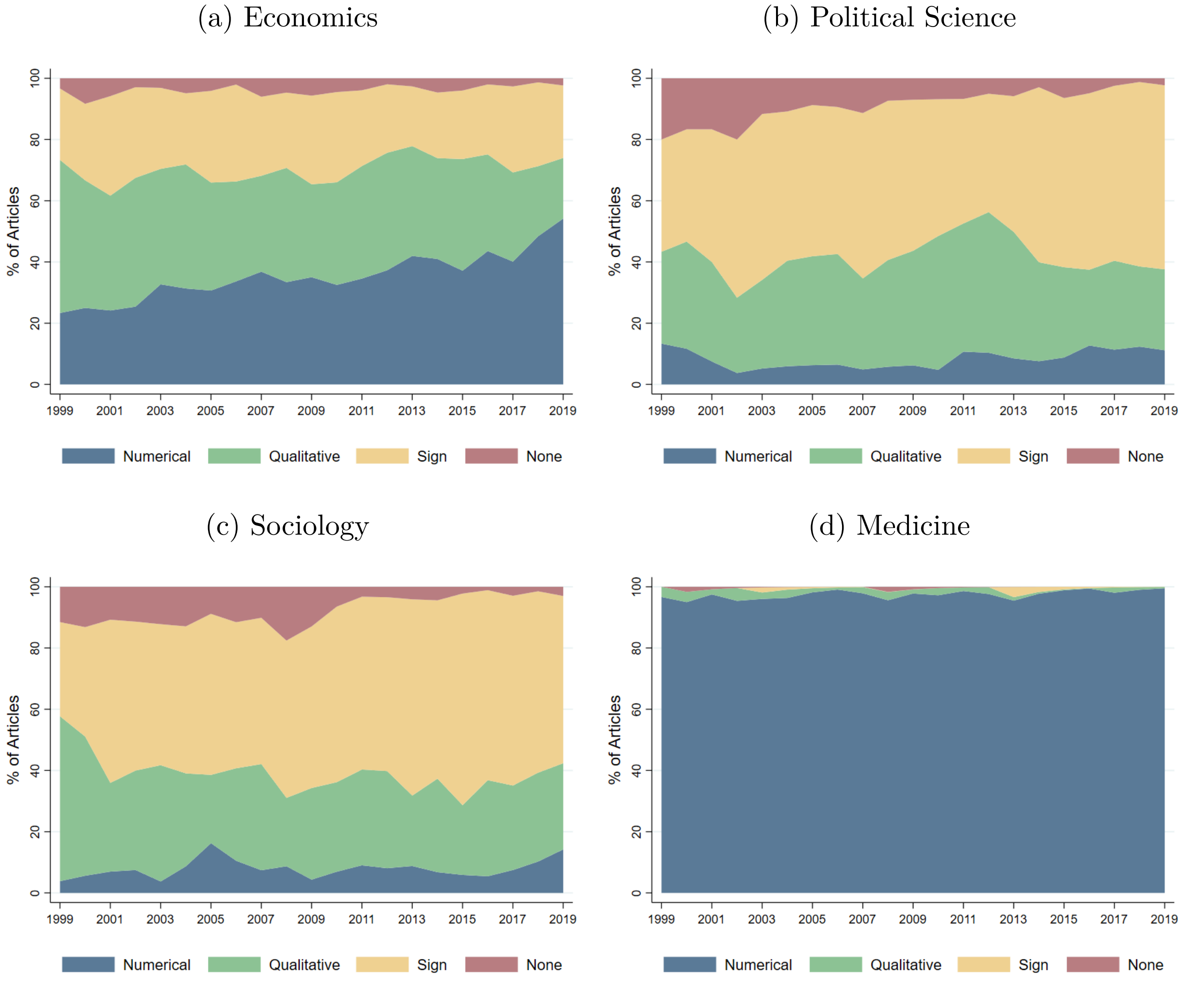
Notes: Robust standard errors reported in parentheses.

so comparisons can only reasonably be made between articles published in the same year, which we take into account when conducting our regression analysis.

3.4 Process and Quality Control

To compile the dataset as described above, a team of research assistants (RAs) under our supervision downloaded the abstracts from journal websites, collected citation counts, and categorized the abstracts using the methodology above. We spent several weeks training the research assistants in categorization, only beginning categorization once we were confident of their consistency. All citation counts and all categorizations were systematically double-checked by the RAs. Whenever a discrepancy occurred between a primary and double-checking RA, an author reviewed the discrepancy. Moreover, all collection and checking tasks were randomly assigned by journal-years to eliminate any RA-specific confounding effects.

Figure 4: The Composition of Headline Results Over Time, by Discipline


(a) Economics

(b) Political Science

(c) Sociology

(d) Medicine

Notes: Since only 30 empirical articles are sampled in each discipline in each year, these charts present two-year moving averages of the shares in each category.

4 Results

4.1 Analysis of Magnitude and Precision

We find that the vast majority of economists, political scientists, and sociologists do not report numerical magnitudes or precision when presenting their headline results. Only 8%±1% of empirical political science or sociology papers have numerical magnitude headline results, as reported in Table 1. Even in economics, arguably a more number-focused discipline, only 36% ± 3% of articles have any numerical headline results. Essentially no one reports numerical precision in these social sciences: 0.3%±0.4% of economics articles, and 0.0% of political science and sociology articles (i.e., not a single political science or sociology article in our sample). These findings are in stark contrast to medicine, where 98% ± 1% of articles present numerical magnitudes and 83% ± 2% present numerical precision in their headline results.

Despite the many disadvantages of sign headline results, Table 1 shows that they are extremely common in the social sciences. Roughly one quarter of economics papers and one half of political science and sociology papers have only sign results (26% ± 3%, 50% ± 4%, and 54% ± 4%, respectively). In contrast, almost no articles in medicine have only sign results (0.5% ± 0.5%).

Sign headline results rarely explicitly say they are statistically significant, but we take that to be implied, assuming that editors and referees won't allow them to be reported if not statistically significant. We confirm the suspicion that statistical significance is implied by reviewing the text of fifty random economics articles that present only sign headline results: we find that 98% ± 3% of these articles base their headline results on empirical findings in the paper that are statistically significant, commonly at the 5% confidence level.

A silver lining for economists is that the reporting of numerical magnitudes appears to be steadily increasing over time. In fact, the share of economics articles reporting numerical magnitudes has roughly doubled over the past 20 years, as seen in blue in Figure 4. But this improvement is apparently unique to economics. In political science and sociology, the reporting of numerical magnitudes has remained scant, while the prevalence of sign results (yellow in the figure) appears to have slightly increased.

We test for linear time trends in the reporting of numerical magnitudes by running a simple OLS regression for each discipline, including journal fixed effects. Our results are reported in Figure 5. We estimate that the share of economics articles with numerical magnitudes in their headline results has increased 1.3 ± 0.5 percentage points per year on average over our sample period. Unfortunately, this increase in numerical reporting appears to be predominantly displacing qualitative headline results (which fell on average by 0.9 ± 0.5 percentage points per year over the same period) rather than replacing sign results, whose trend is less clear. Despite the observed trends in the reporting of magnitudes, the reporting of precision in headline results has not changed in the social science disciplines, remaining at essentially zero.

Lack of space does not appear to be the cause of either the prevalence or persistence of sign headline results, nor for the utter absence of precision. We studied the introductions of two random samples of economics papers and found that most papers with sign headline results do not put magnitudes in the introduction either as depicted on the left of Figure 6. The right panel of the figure shows the pittance of papers–even among those with numerical headline results–that report precision in the introduction. Given that precision does not rise to either abstracts nor introductions, it appears that authors simply place quite low importance on reporting precision.

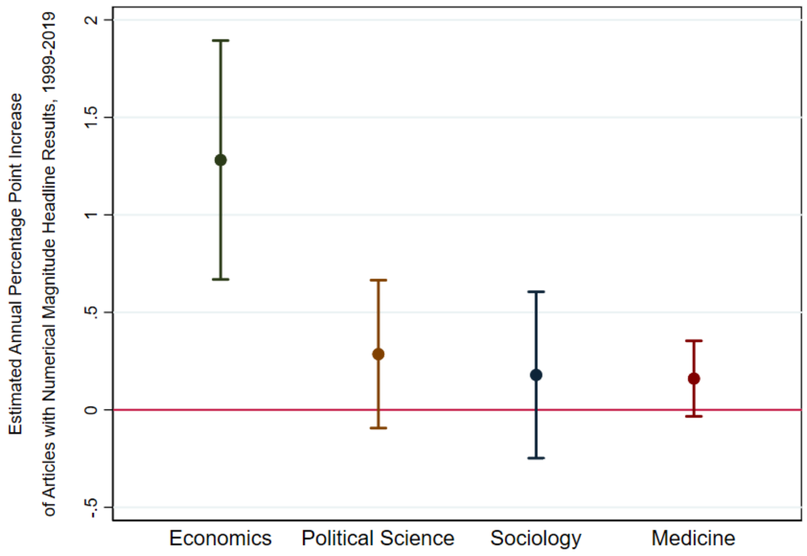


Figure 5: Annual Change in Share of Articles with Numerical Magnitudes Headline Results, by Discipline

Notes: To arrive at these estimates, we run a simple linear OLS regression within each discipline, regressing a binary indicator for the reporting of numerical magnitude on the year, controlling for journal and including robust standard errors. Error bars reflect 90% confidence intervals.

4.2 Citation Counts

We find that economics articles reporting numerical magnitudes for their headline results get more citations than those that do not. Across the social sciences more generally, articles with numerical magnitudes tend to enjoy more (or at least no fewer) citations. We estimate the following log-linear OLS regression separately for each discipline:

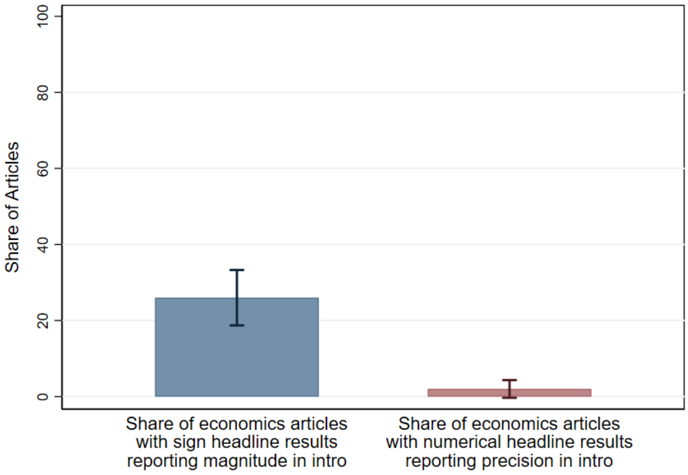
$$\ln c_i = \alpha + \beta n_i + \delta_{jy} + \varepsilon_i \tag{1}$$

where c_i is the citation count for article i, n_i is a dummy indicating whether the article has a numerical headline result, and δ_{jy} are fixed effects for journal j in year y. The coefficient β is thus the average percent extra citation counts of articles with numerical headline results compared to those without, after controlling for journal-by-year fixed effects.

In economics, we estimate that empirical articles with numerical headline results enjoy 19% ± 11% more citations than empirical articles without numerical headline results. We find positive point estimates for both sociology (+24% ± 30%) and political science as well (+1% ± 29%), but both are very noisy.

In medicine, we find a surprising negative relationship, albeit quite noisy (-29%±50%). We suspect this curious finding is driven by the fact that in medical journals, articles with no numerical headline results are exceptional: the finding may therefore be due to a

Figure 6: Numerical Reporting in Economics Introductions



Notes: We select 100 random economics articles whose abstracts report sign results (left) and 100 that report numerical results (right). We then read the introductions to see what share report numerical point estimates (left) or numerical precision (right) for these headline results. Error bars represent 90% confidence intervals.

small sample or these articles may have special qualities that get them cited even if their headlines have limited information.

To ensure robustness, we also top- and bottom-code (winsorize) our observations to ensure that results are not driven by outliers with unusually high or low citation counts. Employing various winsorization schemes, we find results are essentially the same as our un-winsorized sample reported in Table 2.

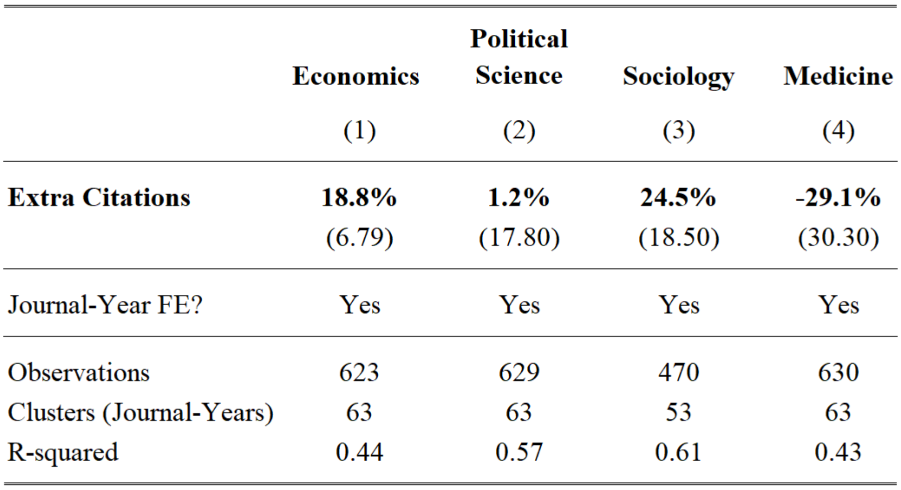
Our finding of extra citations may provide some additional motivation for authors to report numerical headline results. To be sure, we cannot draw causal inference: it is of course possible that unobservable characteristics between articles correlated with numerical headline results can account for the difference in citations, or perhaps that more important results are more likely to be reported numerically. But if authors are interested in broadening the audience of their research, why take the risk by leaving numbers out?

5 Conclusion

Romer (2020) has recently documented and lamented the scarcity of standard errors and confidence intervals in the text of economics papers. We find that for headline results, this lack of numerical reporting extends even to point estimates of magnitudes and also that it is even more extreme in other social sciences, namely political science and sociology. The sharp contrast with medicine points out that there is another way.

Our view is that headline results convey what authors see as most important. We are

Table 2: Extra Citations for Articles with Numerical Magnitudes for Headline Results



	Economics	Political Science	Sociology	Medicine
	(1)	(2)	(3)	(4)
Extra Citations	**18.8%**	**1.2%**	**24.5%**	**-29.1%**
	(6.79)	(17.80)	(18.50)	(30.30)
Journal-Year FE?	Yes	Yes	Yes	Yes
Observations	623	629	470	630
Clusters (Journal-Years)	63	63	53	63
R-squared	0.44	0.57	0.61	0.43

Notes: To arrive at these estimates, we run a log-linear OLS regression within each discipline, with a binary predictor variable for articles with numerical headline results. Standard errors in parentheses, clustered by journal-year.

pleased to find how many more numerical magnitudes there are today in economics than there were two decades ago, but sad that sociology and political science do not exhibit a similar trend. At the same time, the distance these social sciences remain behind medicine is disappointing, especially for political science and sociology. So too is the utter absence of precision in headline results.

We find the lack of progress in getting rid of sign results in social science papers particularly perplexing. Apparently the forces that increase numerical reporting (at least in economics) do not diminish sign reporting even though we find little to recommend sign reporting. We suspect that the absence of numerical reporting reflects a culture of empirical analysis that is still largely wedded to hypothesis testing, where satisfying a statistical threshold and rejecting the null hypothesis that $\beta = 0$ is considered more important than understanding the magnitude and precision of the estimate.

For those who like us crave social science research that emphasize estimation over hypothesis testing and put magnitudes and precision front and center there are two choices. One is to wait until the trends in economics take hold in sociology and political science and hope they spill over to precision reporting. However, the pace of progress is very slow and there is no evidence of increased emphasis on precision.

The second option is to take a proactive approach. The fact that medicine has made great strides in moving toward a culture of estimation and precision and away from significance and hypothesis testing is not happenstance. Decades ago, scientists and statisticians in medicine undertook the project of articulating best practices for reporting empirical headline results that prioritized transparency and informativeness. Perhaps it is high time for social scientists to consider a similar undertaking, or at the very least for journals to consider adopting guidance for authors along these lines. We provide two sample style guides in the Appendix, one relatively modest and the other more ambitious.

Finally, we apologize for any places in this article (and our other work) where we have only sign results. We have tried to practice what we preach, but we are deeply embedded in the very culture we hope to nudge toward a numeric focus.

6 Appendix

For editors convinced that social science articles should prominently include magnitude and precision of empirical findings and move toward what medicine does—or at least avoid sign results—we offer the following suggested style guidelines.

We begin with Style Guide 1. It is modest but could dramatically change the writing of many social science articles (half the articles in political science and sociology and a quarter of those in economics) by discouraging sign results in favor of qualitative results or numerical ones. As we have mentioned, sign results seem without redeeming virtue.

Style Guide 2 goes further. It insists upon presenting numerical findings with both magnitude and precision in abstracts, introductions, and the text of articles more generally. This seems to us the wiser course for disciplines that aspire to policy relevance and to being sciences.

Obviously many other guides are possible and some journals may focus on numerical magnitude reporting alone, or may focus on abstracts, or introductions, or any combination.

6.1 Style Guide 1—Eliminating Sign Results

When reporting an empirical result in the abstract or in text, authors are encouraged to report a numerical point estimate or at the least a qualitative description of the estimate's magnitude.

Don't write: "Wages increased." Instead write: "Wages increased substantially" or "modestly," as appropriate, or better yet report the percentage increase.

6.2 Style Guide 2—Requiring numerical magnitudes and precision

The suggested style guide below is inspired in part by style guides in medical journals like the New England Journal of Medicine.

1. *General.* When reporting an empirical result in the abstract or in the text, authors should report both a numerical point estimate and measure of precision, such as a confidence interval, whenever possible.

   Don't write: "Wages increased." Instead write: "Wages increased by 15%±3%" or "wages increased by 15% (90% confidence interval: 12% to 18%)."

2. *Special rules for abstracts.* If there are multiple empirical findings reported in the abstract, authors may use discretion to choose which (one or several) are the most important when choosing to report numerical point estimates and precision, while describing other results verbally.

3. *Verbal Descriptions.* Verbal descriptions of estimates in the text are often helpful, but not if they simply report the existence or direction of an effect. Reporting that “wages increased” has limited value. Instead write: “wages increased substantially” or “wages increased very little,” as appropriate.

4. *Significant and Insignificant.* Avoid using the word “significant” in a way that creates ambiguity between statistical significance and the substantiality of a result. Never write: “The wage increase was significant.” If you mean “substantial” write that. If you mean “statistically significant” write that or instead give the point estimate and confidence interval. Same with “insignificant.”

5. *No Impact.* Never write: “We find *no effect* of minimum wage laws on wages.” Instead be clear whether you found a small and precise effect, a large but noisy effect where the confidence interval includes 0, or something else entirely.

References

Aitken, B. J. and A. E. Harrison (1999). Do domestic firms benefit from direct foreign investment? evidence from venezuela. *American economic review 89*(3), 605–618.

Begg, C., M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, et al. (1996). Improving the quality of reporting of randomized controlled trials: the consort statement. *Jama 276*(8), 637–639.

Boas, T. C., F. D. Hidalgo, and M. A. Melo (2019). Norms versus action: Why voters fail to sanction malfeasance in brazil. *American Journal of Political Science 63*(2), 385–400.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review 106*(8), 2110–44.

Desmond, M. and N. Wilmers (2019). Do the poor pay more for housing? exploitation, profit, and risk in rental markets. *American Journal of Sociology 124*(4), 1090–1124.

Edlin, A., C. Roux, A. Schmutzler, and C. Thöni (2019). Hunting unicorns? experimental evidence on exclusionary pricing policies. *The Journal of Law and Economics 62*(3), 457–484.

Fieldhouse, A. J., K. Mertens, and M. O. Ravn (2018). The macroeconomic effects of government asset purchases: Evidence from postwar us housing credit policy. *The Quarterly Journal of Economics 133*(3), 1503–1560.

Gelman, A. and H. Stern (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician 60*(4), 328–331.

Giorcelli, M. (2019). The long-term effects of management and technology transfers. *American Economic Review 109*(1), 121–52.

Hopewell, S., M. Clarke, D. Moher, E. Wager, P. Middleton, D. G. Altman, and K. F. Schulz (2008). Consort for reporting randomised trials in journal and conference abstracts. *The Lancet 371*(9609), 281–283.

Im, S.-A., Y.-S. Lu, A. Bardia, N. Harbeck, M. Colleoni, F. Franke, L. Chow, J. Sohn, K.-S. Lee, S. Campos-Gomez, et al. (2019). Overall survival with ribociclib plus endocrine therapy in breast cancer. *New England Journal of Medicine 381*(4), 307–316.

Lerner, J. and U. Malmendier (2010). Contractibility and the design of research agreements. *American Economic Review 100*(1), 214–46.

Lindgren, K.-O., S. Oskarsson, and M. Persson (2019). Enhancing electoral equality: can education compensate for family background differences in voting participation? *American Political Science Review 113*(1), 108–122.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2012). Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International journal of surgery 10*(1), 28–55.

Moher, D., K. F. Schulz, D. Altman, C. Group, et al. (2001). The consort statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Jama 285*(15), 1987–1991.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods 5*(2), 241.

Pope, D. G. and M. E. Schweitzer (2011). Is tiger woods loss averse? persistent bias in the face of experience, competition, and high stakes. *American Economic Review 101*(1), 129–57.

Romer, D. (2020). In praise of confidence intervals. In *AEA Papers and Proceedings*, Volume 110, pp. 55–60.

Tarakji, K. G., S. Mittal, C. Kennergren, R. Corey, J. E. Poole, E. Schloss, J. Gallastegui, R. A. Pickett, R. Evonich, F. Philippon, et al. (2019). Antibacterial envelope to prevent cardiac implantable device infection. *New England Journal of Medicine 380*(20), 1895–1905.

Wasserstein, R. L. and N. A. Lazar (2016). The asa statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p<0.05".