# Opinion on the replication debate over Heyes and Saberian (2019)

David Roodman[1]

May 2025

**Abstract**

Heyes and Saberian (2019) finds that U.S. immigration judges are less likely to grant asylum in cases heard on warmer days. Spamann (2022) corrects errors in that paper, enlarges the sample, proposes additional revisions, and strongly challenges the conclusion. In a rejoinder, Heyes and Saberian (2022) incorporates many of these comments, yet maintains that "results…are qualitatively unchanged." Experimenting with a new academic-literary form, I review the case as a judge might, to offer a take that is more independent and legible than the partisans can offer. I agree with Spamann (2022): the only viable explanation for the combined evidence is publication bias or other forms of result filtration.

[1] I thank Matt Clancy, Anthony Heyes, Oliver Kim, Jeremy Klemin, Rossa O'Keeffe-O'Donovan, Rafael Proenca, and Holger Spamann for comments. The data and code archive is at https://doi.org/10.7910/DVN/EPDXVX.

## 0 Summary

This document opines on a replication debate. Heyes and Saberian (2019, 2022) find that on warmer days immigration judges are less likely to grant asylum. Spamann (2022) does not. This opinion is thorough, in part because it explores new territory in rhetorical form. But the ultimate assessment, and the reasoning behind it, can be stated rather succinctly:

- The parties concur that the original finding, for cases heard between January 1, 2000, and September 30, 2004, does not withstand correction of certain coding errors, and does not generalize to a much longer period, 1990–2019.
- The defense of the original result rests on a novel specification change. Heyes and Saberian show that excluding Chinese asylum applicants restores the original headline result.
- It is hard to defend this exclusion. It appears motivated by the argument that a special law reduced judges' discretion over most Chinese applicants, and therefore left little scope for the play of temperature. But there is no sign of reduced discretion in the data. Cross-judge variance in asylum grant rates was equally wide for Chinese and non-Chinese applicants.
- A separate, secondary analysis in Heyes and Saberian finds warmer weather depressing grants of parole in California in 2012–15. I show that this result is not robust to shifting the time frame to 2016–19.
- A secondary analysis in Spamann (2022) does not find an effect on the sentencing decisions of federal judges. Heyes and Saberian (2022) does not challenge this finding.
- Heyes and Saberian (2019, 2002) possess traits plausibly correlated with publication bias and other forms of result filtration, including the introduction of a discretionary specification change (the Chinese exclusion).
- No theory is offered for why temperature would affect decision-making in the non-Chinese, 2000–04 cases, but not in larger samples.
- The only viable explanation for the results before us is that the main Heyes and Saberian (2019, 2002) findings are products of filtration.

In sum, the texts at hand do not provide compelling evidence that outdoor temperature affected decision-making inside tribunals on asylum, sentencing, or parole. Still, Heyes and Saberian's construction of a novel econometric laboratory for studying human behavior stands as a contribution.

> A judicial opinion…works in differing but related ways. Like a novel, it portrays a human conflict. Like a letter, it intervenes in the conflict it portrays. Like a treatise, it gives a systematic analysis meant to be applicable to many situations. Like a work of history or criticism, it compares disputes that have occurred over the years and analyzes what past authors have proposed. Like a dialogue, it embraces clashing approaches to the conflict before the court. Like a script or computer program, it gives instructions to those who act and decide. Like an oration, it seeks to persuade.
>
> – John Leubsdorf, "The Structure of Judicial Opinions," 2001.

## 1 Introduction

The abstract of Heyes and Saberian (2019) reads, in substantial part,

> We analyze the impact of outdoor temperature on high-stakes decisions…made by professional decision-makers (US immigration judges). In our preferred specification…a 10°F degree increase in case-day temperature reduces decisions favorable to the applicant by 6.55 percent. This is despite judgements being made indoors, "protected" by climate control.

This estimate comes from a sample consisting of asylum cases dated to 2000–04. Heyes and Saberian (2019)—henceforth HS19—also finds a depressing effect of higher temperature on the willingness of a California agency to release prisoners on parole.

Spamann (2022; henceforth S22) finds errors in HS19 and challenges the validity and generalizability of the asylum finding:

> [The HS19] estimate is the result of coding and data errors and of sample selection. Correcting the errors reduces the point estimate by two-thirds, with a wide 95 percent confidence interval straddling zero. Enlarging the sample to 1990–2019 flips the point estimate's sign and rules out the effect size reported by Heyes and Saberian with very high confidence.

S22 too performs a secondary analysis, not of parole decisions in California, but of decisions by federal judges about whether and how long to sentence convicted offenders to prison. Here too, the impact estimates are, if anything, positive.

The rejoinder by Heyes and Saberian (2022; HS22) concedes some of the criticisms, yet finds the core HS19 result preserved:

> The results from both the main linear specifications are qualitatively unchanged, with estimated treatment effects similar in size to the original and retaining statistical significance at conventional levels. Some secondary results lose significance with the erosion of sample size. We also acknowledge the additional finding by Spamann (2022) with respect to external validity.

From HS22's thesis and S22's antithesis, this document attempts to forge a synthesis—one that is more objective and credible than anything either party could write, and one that is perhaps more legible, since it need only cover the most relevant technical details. As will be explained, the goal is to judge the following: *to the extent that a reasonable observer updated their priors after reading the*

*original paper, how much should the subsequent debate reverse or strengthen that update?* Along the way, I develop rubrics that might be useful in future opinions like this one.

In drafting this opinion, I have sought to minimize my own exercise of discretion, preferring to take cues from the existing texts. If the parties agree on a specification change, then I stipulate the old way as an error. Where they disagree, that ideally defines the scope for the opinion. I believe that to minimize the risk of being seen as a party to the debate, I should impose a high burden of justification on myself before introducing novel estimates, especially since I have not pre-registered any. In a few instances I do take the liberty of applying the authors' specifications to new *samples*, judging, that is, that the gain in information justifies the arguably modest exercise of discretion. For example, I transfer an HS19/HS22 specification originally run on 2012–15 data to 2016–19 data.

I find strongly for the S22 comment. The comment and rejoinder agree that the original result does not survive S22's corrections, and that it does not generalize from 2000–04 to 1990–2019. HS22's response relies on a novel sample restriction, excluding Chinese asylum applicants. HS19 and HS22 articulate no theory for why outdoor temperature would influence indoor decision-making in the smaller sample but not larger ones. The only viable theory for the full suite of results generated through the debate involves mechanisms of "filtration" such as publication bias.

That said, a paper's value should depend on the credibility of its strategy for measuring causal effects or other quantities of interest, not merely on whether the best estimates flowing from this strategy are "statistically significant." While S22's null results are more convincing, HS19's construction of a novel laboratory for studying human behavior stands as a contribution.

## 2 Process history

The "docket" for this case—the public record—is short by legal standards but typical by academic ones. The original paper appeared in *American Economic Journal: Applied Economics* (*AEJ Applied*) in 2019. The comment and rejoinder came in 2022, the latter labeled a "correction." To my knowledge, no pre-registered analysis plans are publicly available. The authors evidently exchanged drafts, for the comment responds to a novelty in the rejoinder, the exclusion of Chinese applicants. The American Economic Association data editor validated replication packages for the 2022 papers before they were publicly posted on the Harvard Dataverse.[1] As with most journals, referee reports and editor instructions are not in the public record.[2]

A draft of this opinion was sent to Heyes, Saberian, and Spamann. Heyes and Spamann supplied

[1] The data editor began checking the reproducibility of all AEA publications in July 2019, after HS19 (Vilhuber 2023).

[2] One journal with open review is *Economics* (https://degruyterbrill.com/journal/key/econ/html).

comments, many of which have been incorporated.

I intend to maintain this document as a *living* opinion, incorporating any additional feedback from the original parties or any other commentator. I will preserve the revision history on arXiv.

## 3 Frameworks

To organize the discussion of the comment and the rejoinder, I marshal two conceptual frameworks: Clemens's (2017) typology of replication, and the catalog by Shadish, Cook, and Campbell (2002) of varieties of validity in the empirical study of causal mechanisms.

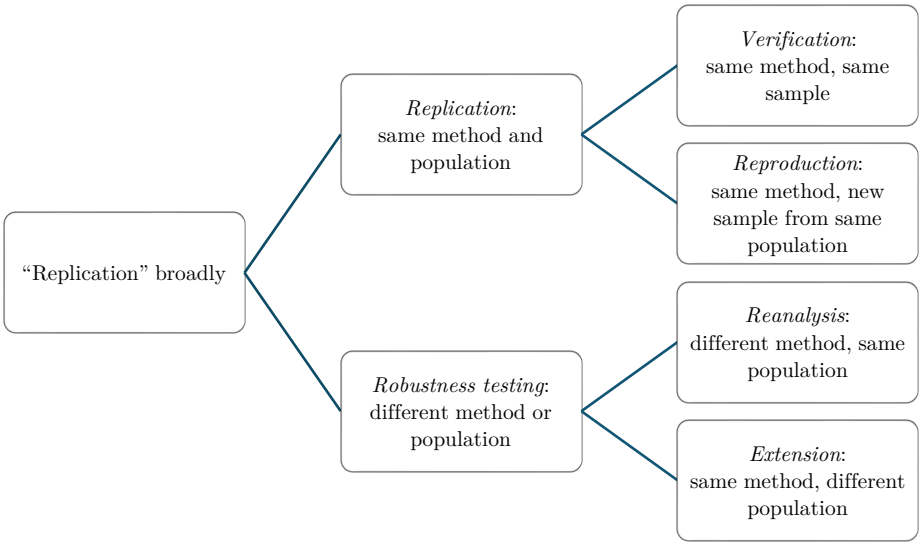
Clemens distinguishes between *replication* of an estimate, which applies the same methods to the same population, and *robustness testing*, which deviates in one or the other. Subdividing, *replication* is *verification* if it applies the same methods not merely to the same population but the same sample, and is otherwise *reproduction*. Meanwhile, *robustness testing* is *extension* if it changes only the population and *reanalysis* if it changes only the methods. These ideas are diagrammed in Figure 1. In fact, many instances of what is commonly understood as robustness testing change both the population and the methods, and so fit in neither subcategory. For this reason, I will sometimes use "replication" in its informal sense too, for any analyses that is proximate to an original study in sample and specification.[3]

With reference to causal estimation, Shadish, Cook, and Campbell (2002, Table 2.1) frames a typology of varieties of validity. *Construct validity* is the fidelity of a constructed variable to the concept it is interpreted to represent. *Statistical conclusion validity* is the quality of judgments about the statistical significance of measured correlations. *Internal validity* is the quality of inference from correlation to causation. *External validity* is "the validity of inferences about whether the cause-effect relationship holds over variation in settings, subjects, treatment variables, and outcome variables." Notably, external validity encompasses generalizability to *subsamples*, "as when an advanced breast cancer patient asks whether a newly-developed treatment that improves survival in general would improve her survival in particular" (Shadish, Cook, and Campbell 2002, p. 83; see also Cronbach 1982, p. 118).

As will be seen, at least in the present context, the different kinds of tests in the Clemens typology tend to generate different kinds of implications for the validity of the original results. For example, extensions (changes in population) usually speak to external validity.

[3] To cite an authority, most of what the Institute for Replication publishes is "replication" only in this broad sense.

**Figure 1. Clemens (2017) typology of replication**




## 4 Arguments and evidence

### 4.1 Related work

With coauthors, Anthony Heyes and Soodeh Saberian have conducted several studies of how local environmental conditions affect human decision-making. Heyes, Neidell, and Saberian (2016) assesses how returns to the S&P 500 index depend on weather and air pollution in New York City. The pollution variables in the regressions are three: ozone ($O_3$), carbon monoxide (CO), and sub-2.5-micron particulate matter (PM2.5). The weather controls are six: temperature, wind speed, atmospheric pressure, dew point (a measure of humidity), precipitation, and cloud cover. Among the variables, PM2.5 is found to be significantly linked to stock prices. Archsmith, Heyes, and Saberian (2018) does something similar for the precision of baseball umpires' decisions, with CO and PM2.5 appearing as significant predictors. Other studies examine the impact of air pollution on the quality of speeches given by members of parliament in Ottawa (Heyes, Rivers, and Schaufele 2019), sleeplessness in China (Heyes and Zhu 2019), and crime in Chicago (Herrnstadt, Heyes, Muehlegger, and Saberian 2021). These too find impacts of air pollution while controlling for temperature and other weather indicators.

HS19—in which temperature rather than air pollution is spotlighted as the treatment—appears to have inspired at least three other studies in the same vein. Evans and Siminski (2021) checks for an impact of temperature on the disposition of criminal cases in New South Wales, Australia, and finds essentially none. The other two studies, however, corroborate HS19. Behrer and Bolotnyy (2024) assesses effects on the actions of police, prosecutors, and judges in Texas; with regard to the

last, it finds fewer case dismissals on days above 90°F. Among cases not dismissed, convictions are only slightly more common; conditional on conviction, however, sentences are longer and fines higher. In India, higher temperature is associated with a higher conviction probability (Craigie, Taraz, and Zapryanova 2023).

### 4.2 The original paper: Heyes and Saberian (2019)

Among the studies just mentioned, HS19 most resembles the one of umpires, for it too is a multi-location analysis set in the U.S.

he HS19 analysis covers all asylum cases completed between January 1, 2000, and September 30, 2004, a time frame determined by a data set that the authors obtained from asylumlaw.org, which in turn acquired it through a Freedom of Information Act (FOIA) request.[4] For each case, the data set provides the nationality of the applicant, the type of case (whether begun with the applicant requesting asylum or the government seeking deportation), the judge who oversaw the case, the judge's decision, and the city and date. Linking to separate databases, HS19 estimates for each city-date pair the daytime average of temperature and five other weather variables, as well as of airborne concentrations of $O_3$, CO, and PM2.5. HS19's preferred specification is a linear model for the binary outcome, grant of asylum. Alongside temperature, the control set consists of the other weather and pollution variables, as well as fixed-effect dummies for nationality, case type, judge, day of week, calendar year, and city/month-of-year combination. Weather indicators are averaged over 6am–4pm each day. HS19 (Table 2, col. 1) finds that for each 10°F increase in outdoor temperature, a judge is 1.075 percentage points less likely to grant asylum. That is, the point estimate is −1.075. The standard error, clustered by city/month-of-year combination, is 0.274. This impact is 6.55 percent of the average grant rate, as reported in the HS19 abstract.

The regressions for California parole decisions follow the same structure. The preferred one yields −1.560 (standard error 0.468; HS19, Table 7, col. 1).

HS19 performs several placebo tests of the asylum finding, all of which reassure. Shifting the weather observations a day forward or back—or 100 days forward or back—essentially erases the result. Similarly, when temperature averages are simultaneously included for each of the days in the weeklong span centered on the case date, only that for same-day temperature shows much significance. Taking weather data from the *farthest* observatory instead of the nearest also vitiates the result.

HS19 also performs also many robustness tests. Varying the set of fixed effects in eight ways does

[4] See bottom of web.archive.org/web/20050429204003/http://www.asylumlaw.org/legal_tools/index.cfm?fuseaction=showJudges2004.

not call the result into question. Neither does taking 24-hour weather averages instead of 6am-to-4pm ones, nor allowing city-specific effects, nor dropping winter cases, nor controlling for the product of temperature and precipitation, nor dropping the pollution controls, nor dropping California cases, nor restricting to clear days or to days with no precipitation, nor dropping cases heard by judges in the top or bottom decile or quartile in overall leniency.

## 4.3 The comment: Spamann (2022)

Focusing on HS19's preferred asylum specification, S22 introduces some 10 changes to data and code. Each embodies a critique, and each raises a question about the meaning of the original result. The comments are summarized in Table 1 and discussed next. I organize them through the Clemens typology: verification, reanalysis, extension.

### 4.3.1 Verification

Nearly half the comments arise from S22's verification exercise. To be clear, what was verified was not that the publicly posted code and data produce the published results, but that the analysis was carried out as described.

One comment goes to the construct validity of the outcome variable. S22 points out that 27 percent of asylum applications were withdrawn or abandoned before being decided, and that HS19 codes them as if they were denied on the merits. "These are decisions of the applicant, not of the judge, and should thus not be included in an analysis of judicial decision making." The point is sensible. That said, the inclusion should bias the results toward zero, making the HS19 headline result conservative. At any rate, the comment and the rejoinder both drop cases not decided on the merits, in effect stipulating that their inclusion in HS19 is a mistake.

Several S22 comments pertain to the assignment of pollution and weather values to cases. The HS19 weather indicators were erroneously averaged over 6am–4pm Greenwich time rather than local time. Some courts were assigned data from the wrong observatories. The Arlington, VA, court, for instance, was mapped to Arlington, TX.

The last clear error surfaced in S22 is that HS19 misinterprets the date field in its data set as "date of hearing" (HS19, p. 244). It is in fact date of *completion*, meaning when the judge files a decision. S22 (note 11) draws out the implications for construct and internal validity:

> In half of the cases decided on the merits, the latest hearing dates and completion dates coincide because the judge decided the case orally at the end of the hearing and formally completed the case. However, in 40 percent of the cases, the completion date is after the latest hearing. This can happen because formal completion of the case is delayed by formalities after a decision on the merits, in which case the completion date is noisier than the hearing date. Alternatively, the judge can take the case into consideration and make a decision in writing after the hearing, in which case the judge chooses the completion date and the weather on that date is no longer plausibly exogenous.

In correspondence, Holger Spamann suggested the following story to illustrate how using date of completion opens the door to endogenous causation. Denying asylum to someone's face is deeply uncomfortable for some judges. When they are leaning toward denial, they may defer the decision till after the hearing, to both avoid that discomfort and, in response to it, reflect. But decisions must be written and cases closed. When it is unpleasantly hot outside, judges more often stay indoors and tackle the work. One can reasonably doubt, *a priori*, that this story exercises much influence in the data. But the same holds for HS19's preferred story that outdoor temperature affects decision-making inside climate-controlled buildings.

Using a more complete data set (discussed below), S22 links to weather and pollution readings using date of hearing instead of date of completion. When the two differ, we cannot be certain which better corresponds to the date of *decision*, which may precede formal closing of the case. But the switch addresses the endogeneity concern and, precisely because of the ambiguity, constitutes an appropriate robustness test.

#### 4.3.2 Reanalysis

S22 introduces several methodological changes motivated by concerns about construct, statistical conclusion, and internal validity. While they arguably improve on the original, the issues they raise are not so incontrovertibly *errors*.

HS19 emphasizes that all weather readings come from stations within 20 miles of the courthouses to which they are matched. But PM2.5 measurements often come from farther away according to S22—a median 73 miles away. That raises doubt about the construct validity of the pollution controls. Relatedly, the need for pollution data shrinks the HS19 sample. For example, lack of CO readings causes all 2001 cases to be dropped. S22 proposes imposing a stricter criterion for pollution data availability—coming from an observatory within 20 miles—and handling missingness in a different way. Instead of dropping cases for which not all three pollution variables are observed, missing values are replaced with zero and companion missingness dummies are entered as controls. I will critique both approaches to missingness in section 5.1.

S22 also revisits the construction of confidence intervals. Because HS19 clusters by combination of city and month of year, it does not fully adjust for serial correlation, for example between January and February of 2000 in Los Angeles. S22 proposes simply clustering by city, which should adjust for all serial correlation. This issue too I will discuss below. Relatedly, S22 points out that the clusters have highly unequal sizes, with New York's by far the largest. In this context, classical clustered standard errors can be biased. Wild-bootstrapped confidence intervals are more reliable (MacKinnon and Webb 2017; Roodman et al. 2019). S22 reports them too.

Separately, S22 inserts fiscal year fixed effects in the place of calendar year fixed effects, since it is more common for immigration policy to change at the start of a federal fiscal year.[5]

#### 4.3.3 Extension

S22's most important step is to access a much larger snapshot of the government's asylum case records. The vast expansion in sample, from 2000–04 to 1990–2019, enables a powerful test of external validity.

#### 4.3.4 S22 headline result

When S22 incorporates all these comments except for clustering by city, the main asylum impact estimate flips in sign and shrinks, yet remains statistically distant from zero. The original's estimate of $-1.075$ percentage points per 10°F (standard error 0.274) is replaced by 0.30 (standard error 0.13; S22, Table 1, col. 9). Switching the clustering basis from city-month to city widens the standard error modestly, to 0.16.[6] The corresponding wild-bootstrapped 95% confidence intervals are $[0.05, 0.55]$ and $[-0.01, 0.82]$. Despite the statistical strength of the new findings, S22 does not conclude that warmer weather makes immigration judges *more* likely to grant asylum.

[5] S22 always defines the calendar and fiscal year dummies with respect to the date of case completion, even when hearing date is used for obtaining weather and pollution values. This choice may be an error or, at least when using fiscal year dummies, may be motivated by the idea that federal policy applies as of the date of filing.

[6] Obtained by editing the S22 code. S22 reports city-month-clustered standard errors and wild-bootstrapped city- and city-month-clustered confidence intervals.

**Table 1. Comments in Spamann (2022) and Heyes and Saberian (2022)**

| Issue | Test type | Validity type | Response |
|---|---|---|---|
| *Raised in comment* | | | |
| Applications not decided on the merits coded as denied. | Verification | Construct | Incorporated |
| Weather indicators averaged over 6am–4pm Greenwich Mean Time, not local time. | Verification | Construct | Incorporated |
| Seven courts erroneously matched to weather or pollution data from distant observatories, such as Arlington, VA, to Arlington, TX. | Verification | Construct | Incorporated |
| Much pollution data from >20 miles away. Median distance for PM2.5 is 73 miles. | Verification | Construct | Not addressed |
| Case dates are for completion, not hearing; the two differ about half the time, sometimes by months. | Verification | Construct, internal | Not addressed |
| The pollution controls have little joint significance; yet including them shrinks samples because of missingness. E.g., all 2001 cases dropped. | Reanalysis | Statistical conclusion | Not addressed |
| Fiscal year arguably a better basis for fixed effects than calendar year. | Reanalysis | Internal | Not addressed |
| *Contra* claim in text, clustering by city–month-of-year pair does not fully address serial and cross-sectional correlation. Clustering by city at least addresses the first. | Reanalysis | Statistical conclusion | Not addressed |
| Clusters have highly unequal sizes, with New York's the largest. Here, wild bootstrapping improves on classical confidence intervals. | Reanalysis | Statistical conclusion | Not addressed |
| A much larger data set was available, and now covers 1990–2019. Original results are for January 2000–September 2004. | Extension | External | Acknowledged |
| *Raised in rejoinder* | | | |
| Applicants from China "adjudicated using a quite different set of criteria and more evincible applicant circumstances," and so are best dropped, unlike in original. | Extension | External | Challenged |

Note: See section 3 for definitions of the varieties of robust testing and validity.

## 4.4 The rejoinder: Heyes and Saberian (2022)

HS22 does not directly challenge the S22 comments. Each is incorporated or left undiscussed. In revised regressions, cases not decided on the merits are dropped, 6am–4pm weather averages are recalculated in local time, and corrections are made in the mapping of courthouses to weather and pollution monitoring stations. HS22 does not dispute that these corrections lead to something closer to a null result.

HS22 does not engage with S22's comments about the handling of distance and missingness in the pollution variables, the construction of standard errors and confidence intervals, the use of fiscal rather than calendar years, and the misinterpretation of the case dates. Most items in that list do not rise to the level of an error. But the last one seems to.

HS22's most consequential responses are about extensions. First, regarding the expansion of the asylum analysis to 1990–2019, HS22 says that S22 "certainly does provide persuasive evidence of the lack of generalizability to the decade(s) before and after that period."

Second, while staying within the original 2000–04 data set, HS22 introduces a novel extension,

which *narrows* the sample. HS22 argues that HS19 would have done better to exclude Chinese asylum applicants. Why? Effective October 1, 1996, Congress created a special application channel for applicants fleeing persecution from coercive population control (CPC) in their home country. In practice, the program supported women fearing forced sterilization or abortion under China's one-child policy. But Congress capped grants through this channel at 1,000 a year. To implement the cap, the Executive Office for Immigration Review placed applicants who received *conditional* grants of asylum in a process queue and drew it down by 1,000 people each fiscal year. HS22 argues that idiosyncrasies of the CPC program make it best to drop all Chinese cases—26,928 out of 109,800 observations.

HS22 motivates the exclusion on two bases, but does not fully lay out the reasoning from either one. First, CPC cases were "adjudicated using a quite different set of criteria and more evincible applicant circumstances." The argument I infer here is that when it is easier for applicants to prove they meet the criteria for asylum, judges have less discretion, which leaves less scope for the play of heat. Second, "the cap meant that the initial resolution of a subset of CPC cases within a fiscal year…differed from number that were ultimately resolved." I will opine on these arguments in section 6.1.

HS22's new, preferred estimate is that for each 10°F temperature rise, the probability of an asylum grant fell by −0.974 percentage points (standard error 0.482; Table 1, col. 1). That is close to the original estimate of −1.075 (0.274), if less precise. I find that this new result is also close to what the original would have been if it too had been computed while excluding Chinese, at −1.072 (0.48).

S22 engages with the idea of excluding Chinese applicants. S22's preferred specification in the much larger 1990–2019 sample yields 0.29 (0.15) instead of 0.30 (0.13) if Chinese are excluded cases during October 1, 1996–May 10, 2005, when the CPC channel was capped (S22, Table 1, col. 10).[7]

### 4.5 Summary: asylum grants

Each of the three texts emphasizes results from a different sample. And not all samples are studied in all texts. This makes it harder to know which disagreements owe to specification changes and error corrections and which to sample changes. Figure 2 lays out a more coherent landscape of results. To the degree possible, it runs the preferred specification in each paper on the various samples in play—with or without Chinese applicants, cases of 2000–04 or 1990–2019.[8] Chinese-only samples

[7] Running the posted S22 code on the data does not quite replicate this result: it returns 0.32 (0.20) and a sample size of 487,431 instead of the reported 561,132. In correspondence, Holger Spamann stated that he has a log to document the published estimate, but cannot now reproduce it, for reasons he has not determined. All other results are validated.

[8] For the S22-based results, standard errors and confidence intervals are clustered by city, as S22 prefers. Only when reporting bootstrapped confidence intervals does S22 cluster by city rather than city-month.

are also introduced. (The reader may want to discount these estimates since this sample definition is novel and not pre-registered. I exercised my own discretion in producing and reporting the results.) A major gap in the figure is that the HS19/HS22 specifications are not extended to the 1990–2019 sample.[9] In each panel of the figure, point estimates are depicted along with their non-bootstrapped 95% confidence intervals.

The upper left panel of Figure 2 confirms that while the original paper finds a significant negative effect of temperature on asylum grants within the January 2000–September 2004 sample, the comment and rejoinder produce coefficients for this sample that are not in great tension with each other, and not statistically distant from zero.[10] In the upper middle panel, we see that dropping Chinese applicants generates sharp disagreement. Now the rejoinder finds a significant impact while the comment does not. Since HS22 finds a weakly negative effect in the full sample and a strongly negative one when excluding Chinese, it is perhaps not surprising that their specification produces a strongly positive effect in the Chinese-only subsample (upper-right of figure).

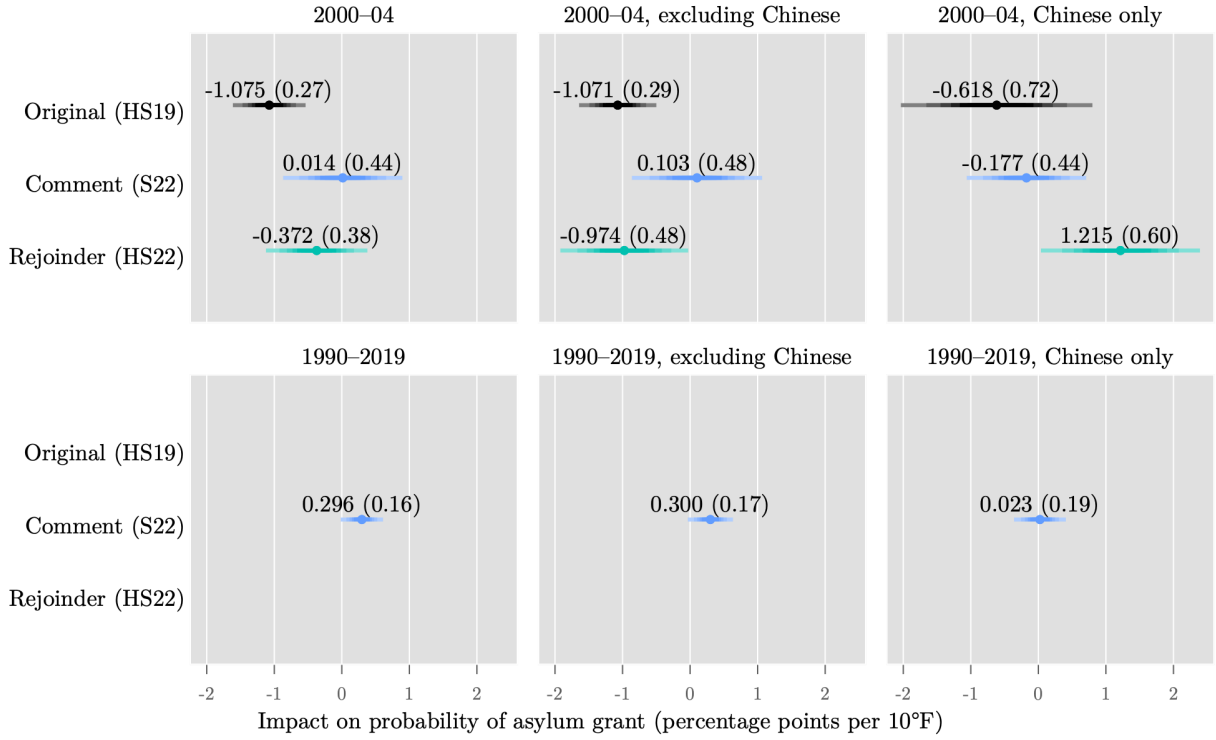
Moving to the 1990–2019 data set, the comment's finding of a *positive* association is shown, with or without Chinese applicants.

In sum, one fundamental disagreement emerges between the comment and the rejoinder, in the 2000–04 sample excluding Chinese. Its statistical mirror image appears in the 2000–04 Chinese-only sample.

[9] It may be possible to fill this gap, but it is not easy. The asylum.org data set used in HS19/HS22 lacks the unique case identifiers, so when comparing regressions run on the two data sets it is difficult to pin down the differences in sample. S22 only matches HS19 exactly when starting from asylum.org's 2000–04 data set.

[10] Only one of these three estimates comes directly from the published documents. The comment result reported here differs slightly from S22's (Table, col. 5), for it incorporates all of S22's modifications aside from sample expansion—notably, basing year fixed effects on fiscal years and dating cases by last hearing. The rejoinder always excludes Chinese applicants, so here one of the rejoinder's regressions is modified to bring them back in.

**Figure 2. Key estimates of impact of outdoor temperature on asylum grants**




Notes: Point estimates and 95% confidence intervals shown. Standard errors in parentheses, clustered by city-month (HS19 and HS22) or city (S22). Most displayed results do not appear exactly in HS19, S22, or HS22, but are obtained by modifying public code to restrict the sample as indicated, or, for S22, to cluster in that paper's preferred way. Chinese applications excluded only for October 1, 1996–May 10, 2005.

## 5 Secondary analyses

HS19 and S22 each check the impacts of outdoor temperature in another kind of tribunal—HS19 in the deliberations of the California Board of Parole Hearings on whether to release prisoners on parole, S22 in federal courts when judges decide whether and how long to sentence convicted offenders to prison. Each paper winds up corroborating its conclusion on asylum grants. HS19 finds a significant, negative impact of temperature on grants of parole. S22's estimates for sentencing are positive, yet hard to distinguish from zero.

## 6 Narrow points

Before reaching an overall assessment, I will discuss narrower matters.

### 6.1 The case for excluding Chinese cases is weak.

As noted in section 4.4, HS22 argues that Chinese asylum applicants are best left out of the analysis because many of them applied through the Coercive Population Control program. HS22 points out that CPC applicants' cases are typically more clear-cut, which (I read to imply) affords less discretion to judges—and that in a study of judges' exercise of discretion. HS22 also observes that the annual cap on grants created a wedge between when a grant was (conditionally) made, and when it

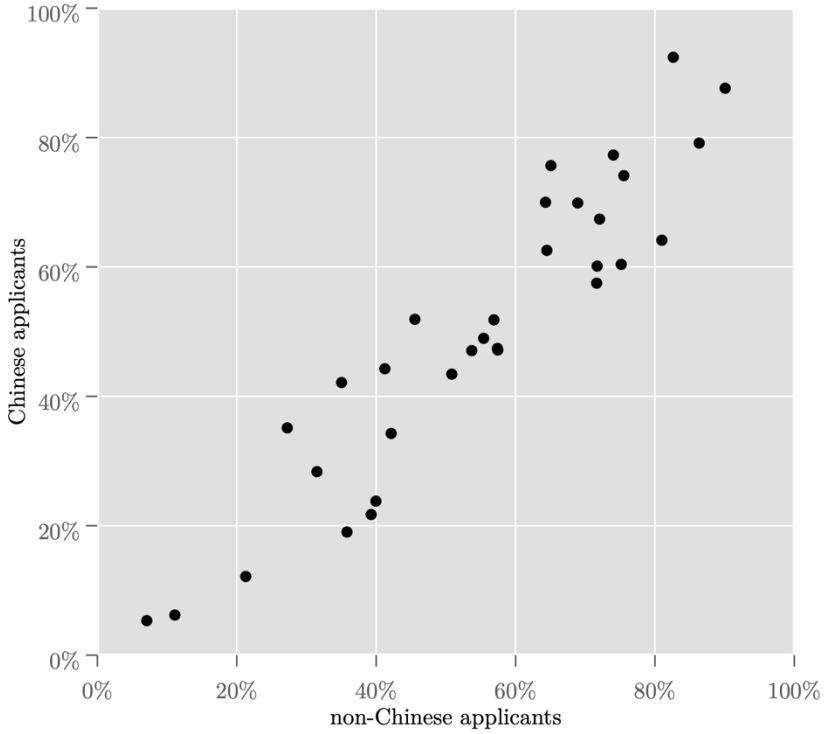
was finalized.

For this reader, the arguments are not convincing. The first makes superficial sense: including cases over which judges can exercise less discretion should bias estimates of impacts on that discretion toward zero. But lack of discretion is inconsistent with the strong positive result when HS22's specification is applied to the Chinese-only sample (Figure 2, upper-right). And it is well-known that judges differed greatly in their rates of asylum grant in this period, even among Chinese applicants. Ramji-Nogales, Schoenholtz, and Schrag (2007), which HS19 cites, extensively documents how different judges, even within the same courthouse, granted asylum at dramatically different rates. The first single-country example (Figure 6) is of Chinese applicants in a period nearly the same as HS19's, fiscal years 1999–2005. One judge never granted asylum to Chinese applicants, another did so 90% of the time, and the rest were evenly distributed in between. Fischman (2014, Figure 8) shows a similar pattern after restricting the data to New York in 2003 for homogeneity. In Figure 3 below, I recast Fischman's descriptive analysis into a scatter plot of grant rates for Chinese versus non-Chinese applicants in the HS19/HS22 data. Copying the papers just cited, I restrict to judges with at least 50 decisions in the data. The correlation is strong, meaning that the probability of winning asylum varies much more across these New York judges than across the Chinese/non-Chinese split. If the cross-judge variability in non-Chinese cases reflects exercise of discretion, it almost certainly does in Chinese cases too. Seemingly both components of variation would be equally subject to the weather.

If I have misread HS22's invocation of "quite different set of criteria and more evincible applicant circumstances" for CPC applicants—if the motivation is not that the CPC criteria afforded judges less discretion—then the case for excluding Chinese on this basis is even weaker. The Chinese subsample is large, and harbors significant variation in the outcome, which temperature might help explain. Excluding it because of heterogeneity begs the question of what other dimensions of heterogeneity in the data—in the circumstances of judges and applicants—ought to drive further exclusions.

The second HS22 argument I do not understand. If a judge issues a conditional asylum grant only to see its ripening deferred to another fiscal year, the judge has still exercised discretion, potentially under the influence of the weather.

**Figure 3. Asylum grant rates by immigration judge, New York, HS22 data**




Notes: Each dot represents a judge. Copying Fischman (2014, FIgure 8), only New York judges with at least 50 decisions are included. The HS22 data span January 1, 2000—September 30, 2004.

## 6.2 Associations in the HS19/HS22 secondary analysis fade in later data. I could not check the S22 secondary analysis in the same way.

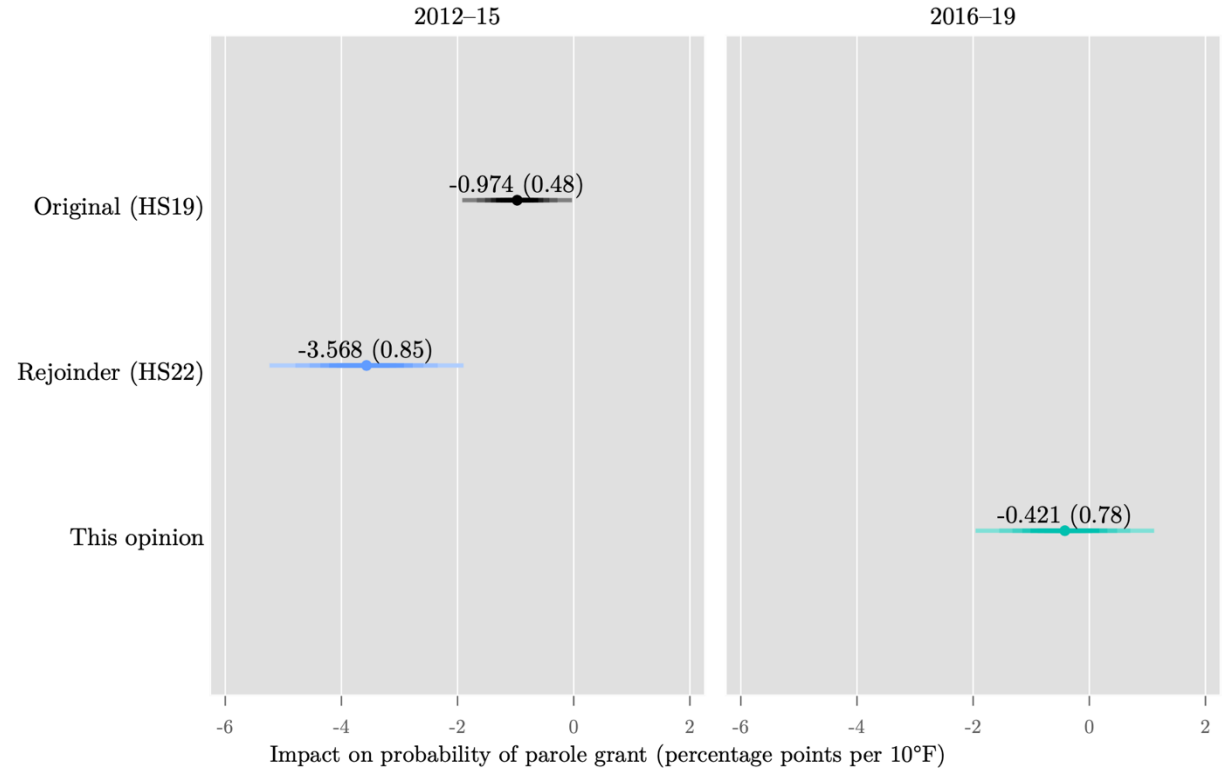
The two sides hardly engage with each other's secondary analyses—of grants of parole in California in HS19/HS22, of federal sentencing in S22. The lack of discussion poses a problem for me. The lack of debate to adjudicate forces me to either pass over these pieces of relevant evidence, or go beyond adjudication. I choose the latter, to an extent: I perform a novel analysis in a way that is minimally arbitrary, though again not pre-registered. The idea is to check external validity by transferring each side's secondary analysis to a new time frame. Unfortunately, this is impractical for S22's sentencing regressions; those are confined to fiscal years 1992–2003 because only in this period do the public data disclose exact dates, which are needed to link to outdoor temperature at the moment of decision. Not so the HS19 and HS22 parole results: where they are computed from 2012–15 data, 2016–19 data are now available.[11]

[11] 2016–19 data were obtained by request from the California Board of Parole Hearings. In fact, even newer data are available. But the COVID pandemic forced the Board to conduct hearings virtually, complicating the meaning of "weather at the hearing location."

Recall that HS19 finds that each 10°F increase in outdoor local temperature changes the probability of a parole grant by −0.974 percentage points (standard error 0.48). While S22 does not directly comment on this estimate, HS22 does update it to reflect several S22 comments on the asylum regressions. For example, parole hearings that do not result in a clear decision to grant or deny are now dropped. In an appendix, HS22 shows that incorporating this and perhaps other corrections leads to a much larger point estimate, −3.568 (standard error 0.85; Table A.7, col. 1). This suggests that including non-decisive hearings makes the original parole estimate conservative.

The HS19, HS22, and new parole results are gathered in Figure 4. I find that transferring the HS22 parole specification to the next quadrennium quite changes the result. The point estimate is now −0.421 (0.78).

**Figure 4. Key estimates of impact of outdoor temperature on parole grants in California, from original, rejoinder, and this opinion**




Notes: Point estimates and 95% confidence intervals shown. Standard errors clustered by venue-month in parentheses.

## 6.3 Federal judges do exercise discretion in sentencing.

HS22 describes the null result in S22's secondary analysis, for prison sentencing, as not very surprising since "it is well understood that immigration adjudication is an area of law where the space for discretion by judges, and therefore for 'mood' to matter, is particularly wide." Immigration

judges may indeed have a freer hand. Yet it bears emphasizing that when sentencing, a federal judge bears the burden of discretion too, as Judge Robert Pratt of the Southern District of Iowa describes:

> Of the many complicated tasks federal district court judges must undertake, sentencing is indisputably one of the most important and, in my opinion, the most difficult. In each case, the law requires the judge to consider a variety of factors in an effort to arrive at a sentence that is ''sufficient but not greater than necessary'' to reflect the seriousness of the offense, to promote respect for the law, to provide just punishment, to afford adequate deterrence, to protect the public, and to provide the defendant with needed educational training, medical care, and correctional treatment. Factors a judge must consider in fashioning this ''sufficient but not greater than necessary'' sentence include the nature and circumstances of the offense, the history and characteristics of the offender, the need for the sentence imposed, the types of sentences available, the sentencing range established by the United States Sentencing Guidelines, pertinent policy statements, the need to avoid unwarranted sentencing disparities among similarly situated offenders, and the need to provide restitution to victims.…Despite having a checklist of factors to consider, there is not necessarily a ''correct'' sentence in a given case. Instead, the entire sentencing process relies on the district court judge's ''discretion'' to impose a sentence that is appropriate under the circumstances. Indeed, most sentencing decisions are reviewed by appellate courts using an ''abuse of discretion'' standard. (Pratt 2016)

If, nevertheless, federal judges exercise less discretion than immigration judges, the loss of outcome variation *per case* may be offset in S22's sentencing analysis by the greater number of cases—471,897 versus the 82,872 in HS22's headline regression. Greater statistical power might help detect a smaller effect. For S22's preferred estimate for the binary outcome of whether to sentence to any prison time, the standard error is 0.10 (S22, Table 2, col. 1). That is about a fifth of the 0.482 standard error in HS22's preferred estimate for immigration judges (HS22, Table 1, col. 1). S22's null findings for sentencing therefore deserve some weight.

### 6.4 The HS19/HS22 standard error calculations contain errors and misconceptions, but this proves to matter little.

HS19 describes its clustering strategy as follows:

> In our preferred specification, standard errors are clustered by city-month, which serves two purposes: to account for spatial correlation across cities and to allow for autocorrelation in decisions in each month.

This passage can leave the impression that standard errors are two-way clustered by city and month, for that would simultaneously address dependence in the time series and cross-section (Thompson 2011; Cameron, Gelbach, and Miller 2011). Instead, standard errors are clustered by the combination of city and month of year (not month). This clustering does not address spatial (cross-city) correlation. And it only addresses temporal correlation at a yearly cadence—for example, correlation within January 2000 and between it and January 2001, but not between January and February of 2000.

The appendices of HS19 and HS22 report tests of robustness to changing the standard error calculation. These results are themselves erroneous in several ways. In HS19, Table A.4, the table's

notes describe column 5 in two contradictory ways. In column 1, the regression clustering by year-week is labeled as clustered by city-week.[12] Column 1 of HS22 Table A.11 is mislabeled in the same way.[13] Two other regressions in the HS22 table, one with Eicker-Huber-White heteroskedasticity-robust standard errors and one with two-way-city-and-week-clustered errors, have their labels swapped, while the "Newey-West" column actually again displays Eicker-Huber-White errors—identical to the previous ones except without a small-sample correction.[14]

The latter error matters less than it might seem, for Newey-West standard errors are not defined in this setting. Newey and West (1987) define them for time series. Roodman (2002), among others, adapts them for panels. They adjust for error correlations within observational units over time, up to a researcher-specified lag distance. None of the data sets of interest here is structure as a time series or panel. The way the HS19 asylum data set is shoehorned into a pseudo-panel structure is confusing and non-deterministic.[15] Consider the asylum case of a Somalian that was heard in Arlington, VA, and concluded on January 6, 2000. What are the predecessor and successor to this observation? In the HS19 analysis, the answer is random and depends on the starting random number seed. In one run, the predecessor is the case of a Salvadoran decided the day before, and the successor the case of an Indonesian decided the day after, both in Arlington. The case of a UK national concluded the same day in Arlington is coded as happening 1,238 time steps later.

Errors and misconceptions aside, none of the tests fully lives up to the HS19 textual motivation to address spatial and temporal dependence. The one that comes closest two-way clusters by city and week of year (HS22, Table A.11, col. 8). It produces the largest standard error. However, the increase is modest, from the 0.482 in HS22's preferred specification to 0.532. And I find that two-way clustering by city and month (not month of year), in order to live up to the textual motivation, only increases the standard error to 0.543. HS19 and HS22 are therefore correct that their preferred results are robust to reasonable revisions of the standard error estimation.

---

[12] The line in the Stata command file "regression.do" is "qui reg res  $weather6t4 $pollutants $dummies , vce (cluster yw)" where "yw" is a unique identifier for year-week combination.

[13] The line in HS22's "regression.do" is "qui reg res  $weather6t4 $pollutants $dummies if ch==0 , vce (cluster yw)".

[14] "qui ivreg2 res  $weather6t4 $pollutants $dummies  if ch==0, robust" in HS22's "regression.do".

[15] The HS19 file "regression.do" first sorts and groups the sample by hearing date and hearing city and assigns a unique identifier within each group to each of its (randomly ordered) cases ("bys date city: gen id=_n"). Then it sorts by that identifier and date to define a time variable ("egen idtime=group(id date)"). In the example, the Somalian, Salvadoran, and Indonesian cases all happen to get identifier 1 within their city-date combinations, which are for Arlington on successive days. So they are placed in sequence in HS19's panel structure.

### 6.5 The parties' distinct methods for handling missing pollution values could both cause bias, but this proves to matter little.

As mentioned in section 3.3.2, many cases that could be included in the asylum regressions are missing a nearby observation for at least one air pollution variable. HS19 and HS22 respond to missingness using the "complete case method," which is to say, they drop observations missing any pollution values. This method could bias the results if the pollution readings are not missing at random (Allison 2002). And shrinking the sample reduces power. In contrast, S22 recodes missing pollution readings as zero and adds corresponding missingness dummies to the control set. S22 cites Jones (1996) for this method. However, Jones concludes that "the missing-indicators methods show unacceptably large biases in practical simulations and are not advisable in general." The bias is absent only if temperature is uncorrelated with the pollution variables, which it is not. The effect is akin to omitted variable bias. The more a pollution control is censored, the more its explanatory power will be shifted onto the treatment of interest, to the extent they covary.

Could differences in the handling of missingness, and the resulting biases, explain the principal disagreement between comment and rejoinder? To check, I modify regressions on both sides by simply dropping the pollution controls. In fact, HS22 (Table A.5, col. 1) and S22 (note 5) agree that this simplification hardly affects results,[16] and I find the same. The estimate from HS22's preferred specification increases in magnitude, from −0.974 (0.48) to −1.169 (0.47). S22's moves the opposite way, from 0.103 (0.48) to 0.114 (0.47). Thus, eliminating the difference in the handling of missingness slightly *increases* the disagreement.

## 7 Principles of judgment

In the ideal, a judicial opinion applies relevant principles and standards to the facts of a particular case to reach a judgment. The facts have been reviewed. Left to be adjudicated are certain disagreements over how to interpret those facts. In this section, I will first think through what a "replication opinion" can usefully opine on. Then I will descend one step toward concreteness. Often, commentary on non-randomized studies devotes great energy to the question of whether the treatment or instrument for treatment is exogenous. Here, the main debate is over robustness—that is, whether a result is "spurious" or tells us something that is, or at least was, true about the world. I will lay out a simple framework for thinking about whether a published result is filtered randomness.

[16] Even when the pollution controls are not entered, HS19 still restricts to observations for which they are all observed.

### 7.1 The scientific stakes: how much the debate affects the update

Amid the back and forth over how results shift as one expands or shrinks the sample or modifies the mapping of courthouses to weather stations, it is easy to lose track of the broader stakes in this effort at adjudication. What can an opinion like this one realistically hope to judge?

By way of illustrating the complexity here, one can easily list claims one should *not*—or need not—try to adjudicate. The disagreement is not over *facts* if by "facts" is meant that certain code run against certain data sets will produce certain results. Nor is moderate discordance among verified estimates, such as in the 2000–04 sample including Chinese, of primary concern. Rarely will independent research teams make the same decisions, and rough agreement reassures as to robustness. Zooming out, any resolution achieved regarding this debate will not decide the general question of how temperature affects human decision-making. Nor, technically, should we expect it to decide the proposition even in the sample where it is in dispute, non-Chinese cases completed between January 1, 2000, and September 30, 2004. By Bayes's theorem, conclusions depend on priors as well as evidence. The range of *prima facie* reasonable priors is wide, and single pieces of statistical evidence are often not dispositive. And consider this final complication: a reader might bring a deeply skeptical prior to *any* non-experimental, non-null empirical finding published in an economics journal. If the views of a cynic cannot be changed, then they too are not at stake.

What is at stake, I suggest, is this: *to the extent that a reasonable observer updated their priors after reading the original paper, how much should the subsequent debate reverse or strengthen that update?* To borrow from Clemens (2017), this framing "offers a standard rather than a definition." Like the "reasonable doubt standard" in criminal law, it will be operationalized differently in each case. Unlike under that standard in law, however, the judgment here need not be binary.

### 7.2 Observable correlates of publication bias and other forms of filtration

That articulation of the stakes begs a question: what information should the priors of a reasonable—but not completely cynical—observer contain? While naïve to the evidence adduced by HS19, the prior should contain knowledge of many kinds. Readers can surmise from experience, or from literature cited in HS19, that environmental stressors influence human decision-making. Common sense says that climate control in buildings substantially decouples the inside environment from the outside one. Readers should also know something about the psychology and sociology of economics research. It is human nature to find null results unexciting. As foragers, we'd rather know where the wild berries are than where they are not, perhaps because the first leads more clearly to an action plan. And as pack hunters, we crave the approval of our peers, for when the next woolly mammoth is felled, status will help assure that one's family gets a piece. These base preferences get spun up in

the academic system into a web of incentives that distort the results that reach readers' eyes. Peers disproportionately cite significant findings. Editors prefer articles they expect to be cited more. Researchers prefer to write articles they expect editors to accept and peers to admire. Academic departments want highly cited faculty. Hiring and tenure committees make awards on the basis of publication.

The resulting distortions have long been recognized, among them a bias toward reporting significant results. Cole (1957) made 120 sequential, hourly observations of the metabolic rate of a "unicorn," which were in fact an $I(1)$ series integrated from a random normal variate. After averaging the five days of hourly data into one, he found no clear diurnal pattern. But then "it occurred to me that in summer at 40° north latitude the hour of rise of the moon may be retarded by approximately 1 hour each night." Averaging over a 23-hour cycle to "eliminat[e] the effect of lunar periodicity"—"a standard sort of procedure for analyzing such data"—produced a stronger oscillation, especially after being passed through a 3-hour moving-average filter. Metabolic activity in the unicorn was then seen to peak at 3am. Cole observed that "like other 'biological clocks,' this rhythm is independent of the temperature at which the observations were made." But with further trial and error, the rhythm "could easily be shown to be highly correlated with environmental fluctuations."

Today the well-intended searching that Cole described is often invoked with the metaphor of Gelman and Loken (2014), "the garden of forking paths."[17] Unlike "fishing," "mining," "file-drawering" (Rosenthal 1979), and "$p$-hacking" (Simonsohn, Nelson, and Simmons 2013), not to mention "data manipulation" and "fraud," "the garden of forking paths" intimates that when researchers preferentially select for statistical significance or other desirable attributes, they need not do so consciously. In the same spirit, it bears stressing that filtration can take place all along the chain from research assistant to researcher to referee to editor. Denton (1985) models data mining as an industry in which individual producers—investigators—engage in no mining, yet the industry as a whole does, because of selection by journals. To embrace all these mechanisms in a minimally pejorative way, I will use the term "filtration." One participant in the debate at hand, Anthony Heyes, has done important work documenting filtration in economics (Brodeur, Cook, and Heyes 2020a, 2022; Brodeur, Cook, Hartley, and Heyes 2024).

Filtration undermines classical inference by upending the assumption that a statistic of interest is drawn at random from its supposed sampling distribution (Cole 1957; Sterling 1959; Tullock 1959). It corrodes statistical conclusion validity. But while filtration is evidently not unusual, how much it

[17] The phrase comes from a short story by Jorge Luis Borges, "El Jardin de Senderos que se Bifurcan."

affects any given paper is hard to assess. "It is increasingly acknowledged that *p*-hacking is an insidious problem," note Brodeur, Cook, and Heyes (2022). "However,…it is difficult or impossible to detect or quantify in any individual study." As a result, *a rational reader must imagine its probabilistic presence everywhere*. In this way, filtration pollutes the entire published corpus.

The pollution does not spread evenly. The rational reader will appraise the risk of filtration in any given study on the basis of observable traits. Relevant factors include:[18]

- *Study type*. Brodeur, Cook, and Heyes (2020a) finds less inflation of *z* statistics in randomized (RCT) and regression discontinuity (RDD) studies and more in one that use instrumental variables (IV) or difference-in-differences (DID).
- *Status and experience of the authors*. Brodeur et al. (2016) finds evidence consistent with the hypothesis that "well-established researchers facing less intense selection should have less incentives to inflate" *z* statistics. It reports negative correlations between *z* inflation and whether an author earned a PhD many years ago, or serves on a journal editorial board (a mark of career advancement), or has tenure.
- *Pre-registration of an analysis plan*. According to Brodeur, Cook, Hartley, and Heyes (2024) pre-registration reduces *z* inflation *if* the plan is thorough.
- *Minimizing apparent exercise of discretion during analysis*. "Deciding whether to exclude data after looking at the impact of doing so" is item 7 on the list of "questionable research practices" of John, Loewenstien, and Prelec (2012).[19]
- For the same reason, *avoiding (the appearance of) multiple hypothesis testing, especially when inferences are not adjusted for it.*
- *Quality of execution.* Implementing afresh the methods described in a paper *always* surfaces some errors. Nevertheless, some code bases contain more errors than others. And the incentives that reward filtration tend, once the preferred sort of result is found, to penalize further quality checking, for that puts a result at risk. We can therefore expect filtration to correlate to a degree with prevalence of implementation errors.
- *Robustness tests, as distinct from placebo tests*. Robustness testing can demonstrate that preferred results do not depend on certain specification and sampling choices—choices that are often arbitrary at the margin. In contrast, successful placebo tests, which show the result fading under certain modifications, do not discriminate between a true relationship and filtered

[18] Reed (2018) also inventories some traits associated with credibility.

[19] As a positive example, Roodman (2007) tests the robustness of a set of studies of the impact of foreign aid on economic growth by transferring between them such specification choices as the measure of aid receipts and the control variables. As a group, the studies at once generate and confine the variation in specification.

randomness.

- *External validity.* If a result holds in other contexts, or in resamplings from the same context, that tends to undercut the ability of filtration to explain it. In this way, *external validity speaks to internal validity.* (On the other hand, if heterodoxy is harder to publish, then the publication process may filter for consistency with prior evidence, reducing the reassurance from demonstrations of external validity.)
- *(In)significance of the debated result.* Filtration usually favors "significant" results. Ergo, if a debated result is insignificant, that enhances its credibility (Ioannidis 2005). The opposite holds for comments: probably S22 owes its publication in no small part to its null results.
- *Theoretical support.* A result that coheres with an accepted causal mechanism is more credible. (Of course, results that do not speak to an established theory, or contradict one, are potentially important.)

The main controversy in the HS19 replication debate is over whether the headline result in HS19 is best seen as valid or "spurious" (S22, p. 522). A judgment between these competing theories might be more objective if a joint prior for the two could be formally stated and formally updated after observing traits such as those listed above. As is often the case, however, formal Bayesian reasoning is as correct as it is impractical. Here, a prior would need to be formulated to apply to a conceptual space with dimensions for each of the listed traits, and link those traits to risk of filtration. The only practical way forward is to start, transparently, with an informal prior whose essence is a judicious suspicion that published claims arise in part through filtration—from what Leamer (1983) calls "whimsy." "The attempt to form a prior distribution from scratch involves an untold number of partly arbitrary decisions," writes Leamer (1983). "The public is rightfully resistant to the whimsical inferences which result, but at the same time is receptive to the use of priors in ways that control the whimsy."

## 8 Opinion

As in any empirical study, the implications of HS19 consist in two propositions:

1. An effect is measured with reasonable fidelity in a particular setting.
2. The finding plausibly generalizes to other settings.

Without the second proposition, the first is perhaps only of historical interest. With the second, the first hints at something more universal in human nature. I will opine on the propositions in turn.

### 8.1 Filtration is the only viable explanation for all the results.

To recap, HS19 reports that "a 10°F degree increase in case-day temperature reduces decisions

favorable to the applicant by" 1.075 percentage points (HS19, Table 2, col. 1). The finding pertains to the population of applicants whose cases were completed between January 1, 2000, and September 30, 2004. Through the back-and-forth, the parties stipulate certain errors in the original. Rectifying them cuts the point estimate by two-thirds, to 0.37, and increases the standard error from 0.27 to 0.38 (upper left of Figure 2, above). S22's further modifications produce a largely compatible result, 0.014 (0.44).

That is, the comment and the rejoinder concur that the original headline lacks construct and external validity. Under a strict reading, the following statement in the rejoinder is therefore incorrect:

> The results from both the main linear specifications are qualitatively unchanged, with estimated treatment effects similar in size to the original and retaining statistical significance at conventional levels.

Of course, there is more to the controversy. HS22 preserves a version of the original finding through a novel exclusion of Chinese applicants. How much should this argument sway the rational reader?

As I read it, HS22 does not contend that HS19's inclusion of Chinese is an *error*, meaning something that almost no reasonable researcher would intentionally do. Nor does S22. HS22 describes the inclusion only as "an additional concern" that creates "ambiguities." There is therefore a contradiction between the content of HS22 and its title: it is labeled a "correction" (perhaps by the editors) but leans heavily on a change that is not a correction.

Still, whether or not correction of error is at stake, if the argument for the exclusion of Chinese is compelling, HS22's finding that the HS19 result persists in a large subsample retains some face validity. How much weight this reading deserves depends on the viability of a competing explanation: filtration. To assess the competition between these two explanations, I evaluate HS22's preferred specification using section 6.2's catalog of traits:

- *Study type.* HS19's treatment is taken to be as good as random, conditional on fixed effects and other controls. It does not fit neatly into any of the four categories for which Brodeur, Cook, and Heyes (2020a) provides evidence about $z$ statistic inflation (RCT, RDD, IV, DID). With its rich set of fixed effects, the specification is perhaps closest to DID, which is more susceptible to inflation.
- *Status and experience of the authors.* Anthony Heyes is well established, having received his PhD in 1993. Soodeh Saberian is more junior: HS19 is based on a chapter of Saberian's (2018) dissertation.
- *Pre-registration of an analysis plan.* To my knowledge, no plan was pre-registered.
- *Minimizing apparent exercise of discretion during analysis.* HS19 reduces the appearance of

discretion by largely copying its specification from other studies, notably Archsmith, Heyes, and Saberian (2018) on baseball umpires; and by analyzing a sample defined by forces beyond the authors' control (a FOIA request they did not file). On the other hand, the larger database exploited by S22 was already available, and was used in Chen, Moskowitz, and Shue (2016), which HS19 cites. And the rationale for excluding Chinese is not only wholly new in the rejoinder, but weak at the levels of principle and evidence.

- *Avoiding (the appearance of) multiple hypothesis testing, especially when not adjusted for.* HS19's specifications include six weather variables and three air pollution variables. Related studies by the same authors have foregrounded certain pollution variables as treatments while casting the weather variables as controls. HS19 does the opposite. The study therefore bears the appearance of multiple hypothesis testing. Brodeur, Cook, and Heyes (2020b) proposes a demanding standard for clearing suspicion of $p$-hacking, based on rerunning a published specification with all possible subsets of the actual control set. That standard seems harsh here: HS19 probably did not test $2^{6+3}$ hypotheses by subsetting the nine weather and pollution indicators. But, plausibly, they tested nine, by including them all at once, then framing a narrative around whichever was most robustly significant.
- *Quality of execution.* The HS19/HS22 implementation is less careful than one might hope. Errors and questionable steps include matching two Arlingtons 1,000 miles apart, not noticing (or else mentioning) that all 2001 cases are dropped for lack of CO readings, coding undecided cases as denials, mistaking dates of hearing for dates of conclusion, clustering standard errors in a way that does not correspond to the motivation in text, and running a method adapted for panel data to a dataset that is not a panel.
- *Robustness tests, as distinct from placebo tests.* As summarized in section 3.2, HS19 demonstrates the robustness of the original headline finding to many specification changes. Still, HS22 does not dispute S22's conclusion that the original finding is fragile to the correction of certain errors. For comparison, S22's preferred specification proves more robust to key changes in sample—adding or dropping Chinese, covering 2000–04 or 1990–2017. The six S22-based results in Figure 2 are fairly concordant. Not so the HS19/HS22 ones.
- *External validity.* HS22 concurs with S22 that the latter "certainly does provide persuasive evidence of the lack of generalizability to the decade(s) before and after" 2000–04. The HS22 result also does not generalize to the full 2000–04 sample, including the Chinese cases. Looking farther afield, the result does not generalize to the sentencing decisions of federal judges.

While it may generalize to California parole decisions in 2012–15 (HS22, Table A.7), this corroboration fades in the next quadrennium (see Figure 4 above). Among the three HS19-inspired studies cited in section 4.1, which have not been subjected to the same scrutiny, the two in Texas and India corroborate HS19 while the one in New South Wales does not. Overall, the general—though not complete—lack of external validity reduces confidence in internal validity.

- *Insignificance of the debated result.* Here, the contested result is statistically significant.
- *Theoretical support.* HS19 and HS22 largely avoid theorizing about how outdoor temperature affects indoor decision-making. "One area in which we have been agnostic throughout the paper is channels" (HS19, p. 262). HS19 does cite papers on how higher temperatures make people more intemperate and risk-averse. But these cannot easily explain why the effect would penetrate the barrier of indoor climate control, nor why it would be present in one sample—in 2000–04, excluding Chinese—and not others.

The strongest factor militating *against* the filtration theory is that HS22's revised specification survives most of the robustness testing defined in HS19, as reported in the HS22 appendix. On the other hand, the HS19 specification did too, only to be substantially eroded by error corrections. Most of the other traits listed above—discretionary removal of a subsample previously embraced, an appearance consistent with multiple hypothesis testing, minimal external validity, minimal supporting theory—nudge a rational observer toward the filtration theory.

Not only is filtration a credible theory; it is the *only* theory put forward to explain all the key results emerging from the debate. HS22 does not offer a mechanism by which temperature would affect decision-making in some samples and not others.

In sum, with regard to the construct, statistical conclusion, and internal validity of the headline findings in HS19 and HS22, *to the extent that a reasonable observer updated from HS19's headline finding, the debate should overturn the update.*

## 8.2 External validity is lacking.

Even if the above conclusion is wrong and the HS22 headline result should be embraced, there remains the question of whether the debate should cause us to reverse any update regarding the impact of temperature on decision-making in other settings. It should. The parties concur that the result does not persist in the full 2000–04 sample, let alone the 1990–2019 one. HS22 does not question S22's null result for federal sentencing. HS19's result for parole grants in California disappears after shifting to an adjacent period.

### 8.3 Conclusion

The authors of HS19 should be congratulated for conceiving of and implementing this natural experiment in the effect of outdoor conditions on indoor behavior. That contribution stands regardless of whether the best impact estimate in this setting is a null result. In my view, S22 gets us much closer to the best estimate. Taken together, the texts examined here do not provide compelling evidence that outdoor temperature affects granting of asylum, sentencing to prison, or release on parole.

## References

Allison, Paul D. 2002. *Missing Data*. Quantitative Applications in the Social Sciences. Sage. https://doi.org/10.4135/9781412985079.

Behrer, A. Patrick, and Valentin Bolotnyy. 2024. "Heat and Law Enforcement." *PNAS Nexus* 3 (5). https://doi.org/10.1093/pnasnexus/pgad425.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32. https://doi.org/10.1257/app.20150044.

Archsmith, James, Anthony Heyes, and Soodeh Saberian. 2018. "Air Quality and Error Quantity: Pollution and Performance in a High-Skilled, Quality-Focused Occupation." *Journal of the Association of Environmental and Resource Economists* 698728 (May). https://doi.org/10.1086/698728.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020a. "Methods Matter: *p*-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60. https://doi.org/10.1257/aer.20190687.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020b. "A Proposed Specification Check for *p*-Hacking." AEA Papers and Proceedings 110 (May): 66–69. https://doi.org/10.1257/pandp.20201078.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2022. "We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and *p*-Hacking in Online Experiments." IZA Discussion Paper 15478. https://docs.iza.org/dp15478.pdf.

Brodeur, Abel, Nikolai M. Cook, Jonathan S. Hartley, and Anthony Heyes. 2024. "Do Preregistration and Preanalysis Plans Reduce *p*-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement." *Journal of Political Economy: Microeconomics* 2 (3): 527–61. https://doi.org/10.1086/730455.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2): 238–49.

https://doi.org/10.1198/jbes.2010.07136.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. 2016. "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Quarterly Journal of Economics* 131 (3): 1181–1242. https://doi.org/10.1093/qje/qjw017.

Clemens, Michael. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31 (1): 326–42. https://doi.org/10.1111/joes.12139.

Cole, LaMont C. 1957. "Biological Clock in the Unicorn." *Science* 125 (3253): 874–76. https://doi.org/10.1126/science.125.3253.874.

Craigie, Terry-Ann, Vis Taraz, and Mariyana Zapryanova. 2023. "Temperature and Convictions: Evidence from India." *Environment and Development Economics* 28 (6): 538–58. https://doi.org/10.1017/s1355770x23000050.

Cronbach, Lee J. 1982. *Deigning Evaluations of Education and Social Programs.* Jossey-Bass Publishers.

Denton, Frank T. 1985. "Data Mining as an Industry." *Review of Economics and Statistics* 67 (1): 124–27. https://doi.org/10.2307/1928442.

Evans, Sally, and Peter Siminski. 2021. "The Effect of Outside Temperature on Criminal Court Sentencing Decisions." *Series of Unsurprising Results in Economics*. http://doi.org/10.26021/10832.

Fischman, Joshua B. 2014. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." *American Law and Economics Review* 16 (1): 40–85. https://doi.org/10.1093/ALER/AHT011.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6). https://doi.org/10.1511/2014.111.460.

Herrnstadt, Evan, Anthony Heyes, Erich Muehlegger, and Soodeh Saberian. 2021. "Air Pollution and Criminal Activity: Microgeographic Evidence from Chicago." *American Economic Journal: Applied Economics* 13 (4): 70–100. https://doi.org/10.1257/app.20190091.

Heyes, Anthony, Matthew Neidell, and Soodeh Saberian. 2016. "The Effect of Air Pollution on Investor Behavior: Evidence from the S&P 500." Working Paper 22753. National Bureau of Economic Research. https://doi.org/10.3386/w22753.

Heyes, Anthony, Nicholas Rivers, and Brandon Schaufele. 2019. "Pollution and Politician Productivity: The Effect of PM on MPs." *Land Economics* 95 (2): 157–73. https://doi.org/10.3368/le.95.2.157.

Heyes, Anthony, and Soodeh Saberian. 2019. "Temperature and Decisions: Evidence from 207,000 Court Cases." *American Economic Journal: Applied Economics* 11 (2): 238–65. https://doi.org/10.1257/app.20170223.

Heyes, Anthony, and Soodeh Saberian. 2022. "Correction to 'Temperature and Decisions: Evidence from 207,000 Court Cases' and Reply to Spamann." *American Economic Journal: Applied Economics* 14 (4): 529–33. https://doi.org/10.1257/app.20200068.

Heyes, Anthony, and Mingying Zhu. 2019. "Air Pollution as a Cause of Sleeplessness: Social Media Evidence from a Panel of Chinese Cities." *Journal of Environmental Economics and Management* 98 (102247). https://doi.org/10.1016/j.jeem.2019.07.002.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Jones, Michael P. 1996. "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association* 91 (433): 222–30. https://doi.org/10.1080/01621459.1996.10476680.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43. http://www.jstor.org/stable/1803924.

Leubsdorf, John. 2001. "The Structure of Judicial Opinions." *Minnesota Law Review.* 86 (447): 447–96. https://scholarship.law.umn.edu/mlr/1677.

MacKinnon, James G., and Matthew D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2): 233–54. https://doi.org/10.1002/jae.2508.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703. https://doi.org/10.2307/1913610.

Pratt, Robert. 2016. "The Discretion to Sentence." *Federal Sentencing Reporter* 28 (3): 161–64. https://doi.org/10.1525/fsr.2016.28.3.161.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2010. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60 (2): 295–411. https://stanfordlawreview.org/wp-content/uploads/sites/3/2010/04/RefugeeRoulette.pdf.

Reed, W. Robert. 2018. "A Primer on the 'Reproducibility Crisis' and Ways to Fix It." *Australian Economic Review* 51 (2): 286–300. https://doi.org/10.1111/1467-8462.12262.

Roodman, David. 2002. "NEWEY2: Stata Module to Extend Newey (HAC Covariance Estimation)." Statistical Software Components, June. https://ideas.repec.org/c/boc/bocode/s428901.html.

Roodman, David. 2007. "The Anarchy of Numbers: Aid, Development, and Cross-Country Empirics." *World Bank Economic Review* 21 (2): 255–77. https://doi.org/10.1093/wber/lhm004.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb.

2019. “Fast and Wild: Bootstrap Inference in Stata Using boottest.” *Stata Journal* 19 (1): 4–60. https://doi.org/10.1177/1536867X19830877.

Rosenthal, Robert. 1979. “The File Drawer Problem and Tolerance for Null Results.” *Psychological Bulletin* 86 (3): 638–41. https://doi.org/10.1037/0033-2909.86.3.638.

Saberian, Soodeh. 2018. *Essays on Environmental Economics*. University of Ottawa. http://doi.org/10.20381/ruor-21843.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2013. “*P*-Curve: A Key to the File Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534. https://doi.org/10.1037/a0033242.

Spamann, Holger. 2022. “Comment on ‘Temperature and Decisions: Evidence from 207,000 Court Cases.’” *American Economic Journal: Applied Economics* 14 (4): 519–28. https://doi.org/10.1257/app.20200118.

Sterling, Theodore D. 1959. “Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa.” *Journal of the American Statistical Association* 54 (285): 30–34. https://doi.org/10.1080/01621459.1959.10501497.

Thompson, Samuel B. 2011. “Simple Formulas for Standard Errors That Cluster by Both Firm and Time.” *Journal of Financial Economics* 99 (1): 1–10. https://doi.org/10.1016/j.jfineco.2010.08.016.

Tullock, Gordon. 1959. “Publication Decisions and Tests of Significance—A Comment.” *Journal of the American Statistical Association* 54 (287): 593–593. https://doi.org/10.1080/01621459.1959.10501522.

Vilhuber, Lars. 2023. “Report of the AEA Data Editor.” *AEA Papers and Proceedings* 113: 850–63. https://doi.org/10.1257/pandp.113.850.