American Economic Journal: Applied Economics 2022, 14(4): 519–528
https://doi.org/10.1257/app.20200118

Comment on "Temperature and Decisions: Evidence from 207,000 Court Cases"†

By Holger Spamann*

Heyes and Saberian (2019b) estimate from 2000–2004 data that outdoor temperature reduces US immigration judges' propensity to grant asylum. This 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. An analysis of all criminal sentencing decisions by US federal district judges from 1992 to 2003 yields no evidence of temperature or other weather effects either. (*JEL* K37, K41, Q54)

Heyes and Saberian (2019b, hereafter AHSS) estimate from immigration court cases in January 2000 through August 2004 that outdoor temperature reduces US immigration judges' propensity to grant asylum. Regressing grant decisions on weather (temperature, dew point, precipitation, pressure, wind, cloud cover) and various controls including pollution (O_3, CO, $PM_{2.5}$) and spatial, temporal, and judge fixed effects, AHSS find that every 10ºF increase in temperature causes a 1.1 percentage point reduction in grants, which is 6.55 percent of the baseline grant rate. AHSS draw the obvious troublesome conclusion for the ideal of justice.

This comment revisits this result from three angles. First, the comment shows that AHSS's estimate is the result of coding and data entry errors—in particular, the mismatching of weather measurements to location and time, the omission of 2001 data, and the use of case completions other than judicial decisions to grant or deny asylum. Correcting these errors reduces the point estimate by two-thirds, to 0.4 percentage points, with a standard error of 0.4. The contrary result in Heyes and Saberian (2022, hereafter "[the] erratum") is driven entirely by its unwarranted exclusion of Chinese applicants.

* Harvard Law School (email: hspamann@law.harvard.edu). Benjamin Olken was coeditor for this article. Thanks to Colby Wilkinson for downloading and cleaning the weather and pollution data; to Alex Perrault for first drawing my attention to the Philadelphia, San Antonio, and San Diego errors; to Anthony Heyes for discussing their data with me; and to three anonymous reviewers for comments. The discussion of coding errors, especially Section I, responds to a request by the editor and reviewers. This work was mostly done while I was a fellow at the Wissenschaftskolleg zu Berlin; I am very grateful for their hospitality. I have no relevant or material financial interests that relate to the research described in this paper.

† Go to https://doi.org/10.1257/app.20200118 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

Second, this comment extends the sample to the years 1990–2019, six times as long as AHSS's sample period. In this larger sample, the point estimate is a 0.3 percentage point increase in grants for every 10ºF increase in temperature. The 95 percent confidence interval may or may not include zero, depending on how it is calculated, but it certainly excludes an effect of the absolute magnitude reported in AHSS, let alone with the same sign. Excluding Chinese applicants (as in the erratum) has no effect on these estimates.

Third, to probe external validity, this comment analyzes all criminal sentencing decisions by US federal district judges from 1992 to 2003. (AHSS use California parole board decisions, but these are not made by judges.) This yields no evidence of an effect from temperature or other weather, either, and can again rule out an effect of the absolute size reported by AHSS with very high confidence. (In parallel work, Evans and Siminski [2021] obtain similarly precise null estimates of temperature effects on criminal adjudication in New South Wales, Australia.)

All data and scripts used in this paper are available at doi.org/10.7910/DVN/3LOR3R.

I. Errors in AHSS

AHSS's data and code are publicly available on the journal's website.[1] Running AHSS's "regression.do" on their final dataset "matched.dta" reproduces the point estimates and other statistics of AHSS's "preferred specification."[2] This is shown here in Table 1, model 1 for only the temperature coefficient and the joint F-statistic for weather, which will be the sole focus of discussion.[3]

Inspection of AHSS's data, regression code, and data assembly code reveals errors, however, that considerably inflate these results. First, weather and pollution measurements are mismatched to locations and times. Seven courts have data from faraway cities (e.g., the immigration court in Arlington, Virginia has data from Arlington, Texas), and weather measurements are not adjusted from GMT to local time nor restricted to the 20-mile radius around the courthouse mentioned in AHSS (p. 247). Fixing these errors reduces the estimated coefficient by one-third (Table 1, model 2).

Second, a large amount of pollution data is entirely missing or drawn from further than 20 miles from the courthouse. AHSS have complete pollution data from within 20 miles for only 75,835 of 269,756 asylum cases. Carbon monoxide (CO) data is missing for all 47,555 cases decided in 2001; AHSS drop these from the sample (complete case method). Of the remaining cases, 146,239 have pollution measurements from greater than 20 miles away; AHSS keep these in the sample.[4] Neither dropping these observations nor using the distant measurements is satisfying. The latter yields a noisy measurement, while the former sacrifices most of the sample

[1] https://www.aeaweb.org/doi/10.1257/app.20170223.data (Heyes and Saberian 2019a).

[2] AHSS's "preferred specification" is model 1 of AHSS's Table 2.

[3] The F-statistic of 3.75 produced by AHSS's publicly posted data and code and reported here differs slightly from the value that AHSS report (3.41).

[4] The problem is mostly with particulate matter less than 2.5 microns in width ($PM_{2.5}$), which is measured at a median of 73 miles from the courthouse.

with no concomitant benefit: there is no evidence that the pollution controls have an effect on the outcome or, more to the point, that their exclusion leads to omitted variable bias for the weather estimates.[5] Model 3 and subsequent models restrict pollution measurements to 20 miles but do not drop observations without such measurements from the sample; rather, they replace missing pollution values with zeroes and add dummies indicating missingness (cf. Jones 1996). Model 3 and subsequent models also use the one station with the most complete coverage for each combination of city, year, and variables (CO, O_3, $PM_{2.5}$) or set of variables (weather), eliminating noise from local variation when the closest active measurement station changes from day to day. As expected, model 3's improvements increase the sample size and reduce the standard error relative to model 2. But they reduce the point estimate even more.

Third, and finally, AHSS's data are not limited to judicial decisions on the merits (i.e., to grant or deny asylum; AHSS, p. 244) but contain all completions of asylum proceedings.[6] In particular, the data contain 20 percent withdrawals and 7 percent abandonments. These are decisions of the applicant, not of the judge, and should thus not be included in an analysis of judicial decision making. This issue cannot be analyzed or fixed using AHSS's publicly posted data because they do not contain a variable distinguishing the different types of completion; however, the requisite data are available directly from the Executive Office for Immigration Review (EOIR). AHSS's data come from the defunct website asylumlaw.org (AHSS, p. 244), but asylumlaw.org, in turn, obtained the data from the EOIR.[7] The EOIR provided these data in response to individual Freedom of Information Act requests (e.g., Chen, Moskowitz, and Shue 2016) and, since June 2018, makes them openly available online.[8] The data obtained directly from EOIR will henceforth be referred to as EOIR data.

To show that the change in data source does not drive the subsequent results, model 4 first reestimates model 3 with exactly the same sample and specification but substituting the EOIR data for AHSS's data. The number of observations and cities is somewhat larger because model 4 also replaces AHSS's weather and pollution data with fresh downloads from the NOAA ISD Lite and the EPA AQS databases, respectively. The point estimate stays virtually unchanged from model 3, while the R^2 and standard error increase slightly. The only change from model 4 to model 5 is to limit the sample to judicial grants or denials of asylum. Now the point estimate shrinks by another fifth, while the standard error increases to almost the same size. The 95 percent confidence interval still includes AHSS's point estimate of -1.075, but it also contains a large swath on the other side of zero. The F-statistic of the

[5] In model 2 the joint Wald test statistic for the three pollution coefficients is $\chi^2(3) = 5.96$, $p = 0.11$, and the Wald test statistic for a change in the weather coefficients due to omitting the pollution controls is $\chi^2(6) = 5.58$, $p = 0.47$. The statistics are even lower, and the p-values higher, when restricting the sample to the 70,843 complete observations with measurements taken at 20 miles or less.

[6] This can be inferred by comparison of AHSS's number of cases to the official statistics (Executive Office for Immigration Review 2005, pp. K2, K4) and was confirmed by AHSS in private communication. US immigration judges only decided 149,164 asylum applications on the merits during fiscal years 2000–2004.

[7] See the data note at the bottom of https://web.archive.org/web/20050429204003/http://www.asylumlaw.org/legal_tools/index.cfm?fuseaction=showJudges2004.

[8] https://fileshare.eoir.justice.gov/FOIA-TRAC-Report.zip. The data are updated monthly. This comment uses the July 2019 release, which contains data through July 2019.

joint null hypothesis for all six weather variables, which was large in AHSS, is now very small. The bottom line is that after correcting the coding and data entry errors in AHSS, AHSS's preferred specification yields no evidence of an effect of outside temperature or other weather on judicial decisions in AHSS's sample period, January 2000 through August 2004.

After these issues were brought to their attention, Heyes and Saberian issued the erratum to "correct errors" but found "estimated treatment effects similar in size to the original and retaining statistical significance at conventional levels." The driver of this discrepant finding is that unlike their original article, this comment, and other literature analyzing asylum data (e.g., Ramji-Nogales, Schoenholtz, and Schrag 2007; Chen, Moskowitz, and Shue 2016), the erratum excludes Chinese applicants, who constitute a quarter of its cases.[9] To show this, model 6 reproduces the erratum's "preferred specification" using its data and code but including Chinese applicants. The estimates in model 6 are virtually the same as model 5: a small temperature coefficient of -0.37 with a standard error of 0.38. (For comparison, without Chinese applicants the coefficient would be -0.97 and the standard error 0.48, similar to the numbers reported in the erratum.) The erratum argues that Chinese applicants were different because during its sample period two-thirds of Chinese asylum grants were based on a claim of coercive population control (CPC), which was de facto (if not legally) limited to Chinese applicants. However, there is no reason why weather would not affect CPC claims if it affected other asylum claims. In the extended sample using 1990–2019 (Section II, below), there is no evidence for weather effects even when Chinese applicants are excluded (model 10). The erratum's discrepant result when excluding Chinese applicants and years outside 2000–2004 thus appears to be spurious.

II. Extended Sample Period, 1990–2019

The EOIR data allow us to enlarge the sample period to the years 1990–2019.[10] The larger sample yields no evidence of a temperature effect or other weather effect, either, and confidently rules out an effect of the absolute magnitude estimated in AHSS.

Models 7 and 8 are extensions of models 4 and 5, respectively, to the larger sample. When the sample is all case completions as (erroneously) in AHSS, the estimate is now a precisely estimated zero: the 95 percent confidence interval in model 7 is $[-0.2, 0.2]$ when clustering on city-month as in AHSS; i.e., the upper 95 percent confidence bound on the absolute effect size in model 7 is one-fifth of AHSS's absolute point estimate. When the sample is only judicial grants and denials (model 8), the point estimate is larger in absolute value than in model 7 but still only one-fifth of the size estimated in AHSS and with the opposite sign. The upper 95 percent

[9] The treatment of Chinese applicants is not the only difference—excluding Chinese applicants from model 5 yields a temperature coefficient of only 0.59—but tracking down the source of the other differences in the data assembly and sources would be burdensome and, in light of the main text, unnecessary.

[10] The EOIR data begin around 1987, but coverage is incomplete until 1990.

confidence bound of 0.5—clustering on city-month—is half of the absolute effect size estimated in AHSS.

Model 9 makes two final improvements to the specification. First, it uses a fiscal-year instead of a calendar-year fixed effect to better account for the possibility of changing reporting conventions. Second, it uses the exogenous latest hearing date rather than the endogenous case completion date, excluding from the sample the 10 percent of cases with a hearing after the completion date.[11] These changes increase the point estimate, shrink the standard error, and increase R^2 by 10 percent each. Model 10 shows that excluding Chinese applicants makes no difference in this larger sample (cf. supra, discussion of erratum).

Before discussing the results of what is arguably the best specification, model 9, a note on inference. Like AHSS, Table 1 displays conventional "sandwich" standard errors clustered at the city-month level in parentheses. However, there are two concerns about the validity of these standard errors and the resulting inference. First, the clusters are of extremely unequal size because immigration courts have vastly differing case loads: one-fourth of all cases are heard in New York City, and another tenth each in Los Angeles and Miami. For example, in model 9, the 12 largest of the 661 city-month clusters—all 12 from New York City—contain one-fourth of all observations, and the largest 5 percent of the clusters contain almost half of all observations. With such unequal cluster sizes, conventional inference can be grossly misleading (Carter, Schnepel, and Steigerwald 2017; MacKinnon and Webb 2017). The wild bootstrap-*t* (Cameron, Gelbach, and Miller 2008; Roodman et al. 2019) provides a superior alternative (MacKinnon and Webb 2017). Table 1 shows wild bootstrap-*t* 95 percent confidence intervals clustered on city-month in square brackets. Second, clustering on city-month is insufficient because weather is serially correlated; hence, treatment assignment in one month is correlated with treatment assignment in adjoining months (cf. Abadie et al. 2017). Table 1 therefore also shows wild bootstrap-*t* 95 percent confidence intervals clustered on city in braces.[12]

Regardless of how the 95 percent confidence interval for temperature is constructed in model 9, it always excludes the point estimate reported in AHSS (-1.075) by a large margin and even in absolute value. The temperature effect and effect size reported in AHSS can thus be ruled out with high confidence. In fact, model 9 suggests that there is no temperature or other weather effect at all. The most credible, city-clustered bootstrap 95 percent confidence interval for temperature includes

[11] A reviewer advises that the date variable on asylumlaw.org, and hence AHSS, is the completion date. 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 either scenario, it is preferable to use the hearing date, which is set a long time in advance (AHSS, p. 245). The only problem with the hearing date is that after completion of the asylum case, a new hearing date can be set upon a motion to reopen or reconsider the case (8 C.F.R. 1003.23) or in some other circumstances, in which case the new hearing date overwrites the hearing date at which the asylum claim was decided. This is the reason to exclude the 10 percent of cases with a latest hearing that is after the completion date.

[12] "City" refers to the EOIR's BASE_CITY variable, which might be better described as "Immigration Court" because there are two of them in some cities (like New York City, Los Angeles, and Miami).

Table 1—Asylum Decisions

	Panel A. Models 1–5				
	(1)	(2)	(3)	(4)	(5)
	Probability of grant (%)				
Temperature (10ºF)	−1.07	−0.67	−0.48	−0.48	−0.39
	(0.27)	(0.27)	(0.24)	(0.29)	(0.36)
	[−1.63, −0.52]	[−1.22, −0.10]	[−0.97, −0.00]	[−1.08, 0.11]	[−1.12, 0.34]
	{−1.88, −0.31}	{−1.45, 0.20}	{−1.14, 0.04}	{−1.29, 0.23}	{−1.50, 0.66}
R^2	0.17	0.17	0.17	0.20	0.25
N	206,924	202,963	248,587	267,222	138,938
City-months	514	344	348	558	548
Cities	43	29	29	49	47
F (all weather)	3.75	2.32	1.82	0.89	0.31
F (temp., clouds, rain)	5.74	2.98	2.39	1.09	0.53
Data source	AHSS final	AHSS raw	AHSS raw	Agencies	Agencies
Sample years	2000–2004	2000–2004	2000–2004	2000–2004	2000–2004
Cases (completions)	All	All	All	All	Merits
Court location	AHSS	Actual	Actual	Actual	Actual
Weather >20 mi.	Yes	No	No	No	No
Weather 6am–4pm	GMT	Local	Local	Local	Local
Pollution >20 mi.	Yes	Yes	No	No	No
Pollution missingness	CC	CC	MD	MD	MD
Stable stations	No	No	Yes	Yes	Yes
Chinese excluded	No	No	No	No	No

	Panel B. Models 6–10				
	(6)	(7)	(8)	(9)	(10)
	Probability of grant (%)				
Temperature (10ºF)	−0.37	−0.02	0.27	0.30	0.29
	(0.38)	(0.10)	(0.14)	(0.13)	(0.15)
	[−1.16,0.40]	[−0.23,0.19]	[−0.01,0.56]	[0.05,0.55]	[−0.01,0.60]
	{−2.22,0.37}	{−0.25,0.43}	{−0.24,0.79}	{−0.01,0.82}	{−0.10,0.74}
R^2	0.25	0.23	0.33	0.37	0.31
N	109,798	1,487,308	695,084	606,754	561,132
City-months	397	667	663	661	661
Cities	36	58	57	56	56
F (all weather)	1.32	0.11	0.86	1.41	0.93
F (temp., clouds, rain)	1.85	0.05	1.53	2.17	1.62
Data source	AHHS err.	Agencies	Agencies	Agencies	Agencies
Sample years	2000–2004	1990–2019	1990–2019	1990–2019	1990–2019
Cases (completions)	Merits	All	Merits	Merits	Merits
Court location	Actual	Actual	Actual	Actual	Actual
Weather >20 mi.	No	No	No	No	No
Weather 6am–4pm	GMT	Local	Local	Local	Local
Pollution >20 mi.	No	No	No	No	No
Pollution missingness	CC	MD	MD	MD	MD
Stable stations	No	Yes	Yes	Yes	Yes
Chinese excluded	No	No	No	No	Yes

Notes: Table lists linear regressions with fixed effects for judge, applicant nationality, day of the week, year, city-month, and defensive application and controls for other weather (dew point, rain, cloud cover, pressure, wind) and pollution (O_3, CO, $PM_{2.5}$). Standard errors clustered on city-month are in parentheses; wild bootstrap-t (99,999 replications) 95 percent confidence intervals clustered on city-month (city) are in brackets (braces). Court locations (for matching courts to weather and pollution data) are as in AHSS (including errors, as discussed in the main text) or the actual locations, as indicated. Weather variables are hourly measurements averaged over 6am–4pm GMT or local time, as indicated; pollution variables are measured daily. Weather and pollution measurement stations are or are not restricted to a 20-mile radius around the courthouse and stable within city-year (for a given variable), as indicated. Missingness in pollution variables is handled either using CC (complete case method: dropping observations with missing values) or MD (missing dummies are added while missing values are replaced with zero). Model 1 is an exact reproduction of AHSS's "preferred specification," i.e., AHSS Table 2, model 1.

Sources: "AHSS final" and "AHSS raw" refer to AHSS's publicly posted "matched.dta" and raw data, respectively (Heyes and Saberian 2019a). "AHSS err." refers to the erratum dataset (Heyes and Saberian 2022b). "Agencies" refers to fresh data downloaded from NOAA ISD Lite (National Oceanic and Atmospheric Administration 2019), EPA AQS (US Environmental Protection Agency 2019), and EOIR (Executive Office for Immigration Review 2020). The sample contains data from January 1, 2000 to August 31, 2004 or January 1, 1990 to July 31, 2019, with all completions (including abandonments, withdrawals, and "other" completions) or only judicial decisions on the merits (grant/deny) and applicants from all countries or excluding China, as indicated.

zero, and the joint F-statistics for weather are small. Perhaps more importantly, the positive sign of the temperature point estimate is the opposite of what AHSS had found and had argued one should expect if an effect existed (pp. 238, 262). Finally, the existence of weather effects in climate-controlled courthouses seemed rather improbable a priori, and the small point estimates are much more consistent with this skeptical view than with weather effects of a meaningful size. (For calibration, the interquartile range between individual judges is over 20 percentage points even after adjusting for all the covariates of table 1, and the interquartile range for different nationalities is even larger. See Ramji-Nogales, Schoenholtz, and Schrag 2007; Fischman 2013.) Unreported specifications with city instead of city-month fixed effects (i.e., looking at weather levels) or controlling for the seven-day cross-year average weather yield similar nonresults. To address remaining doubts, the next section turns to another dataset for a fresh look at the same phenomenon.

III. External Validation: Sentencing

AHSS explicitly consider asylum adjudication merely a "test-bed" to "investigate the link from outdoor temperature to decisions made by experienced professional decision-makers." If such a link existed, one should be able to observe it also with other "experienced professional decision-makers"—above all, other judges. Criminal sentencing decisions by US federal district judges fulfill AHSS's (pp. 239–40) criteria for an "ideal test-bed": (i) high-stakes decisions (ii) made by experienced professionals (iii) operating in a climate-controlled indoor environment (iv) generating high-frequency data on prescheduled dates with a rich set of covariates (to make cases roughly comparable).[13]

The United States Sentencing Commission (USSC) makes sentencing decisions available through the present but provides decision dates only for fiscal years 1992 through 2003, comprising 610,687 cases. The data contain a rich set of case and defendant characteristics including the offense level and criminal history scores that determine the USSC's sentencing grid. (See Cohen and Yang 2019 for a description of the data and their institutional background.) One or more of these covariates are missing for 15 percent of these cases and weather data is missing for some others, such that the usable number of observations from the USSC is 464,518.

The main outcome variable is whether the judge sentenced the defendant to prison time and, if so, for how long. To capture these two dimensions, Table 2 shows regression results for the binary imprisonment decision (model 1) and, for those defendants who did receive prison time, the natural logarithm of the sentence length (model 2). Weather and pollution regressors enter and standard errors and confidence intervals are calculated as in models 3–5 and 7–10 of Table 1. The other specification choices follow Yang (2015) and Spamann (2018).

[13] Sentencing decisions are partially constrained by sentencing guidelines, but judges retain plenty of discretion. The lower bound of the recommended range is at least 20 percent below the upper bound; moreover, departures from this range are frequent. Economists have studied how judges use this discretion (e.g., Fischman and Schanzenbach 2012; Cohen and Yang 2019).

Table 2—Sentencing Decisions

	Probability of imprisonment (%) (1)	Ln(length of prison sentence) × 100 (2)
Temperature (10ºF)	0.12	0.01
	(0.10)	(0.26)
	[−0.07, 0.31]	[−0.51, 0.53]
	{−0.16, 0.35}	{−0.72, 0.61}
R^2	0.49	0.75
N	471,897	391,341
City-months	1,054	1,054
Cities	88	88
F (all weather)	0.81	0.11
F (temperature, clouds, rain)	1.35	0.16

Notes: Table lists linear regressions with fixed effects for sentencing grid cell, type of offense, whether a statutory minimum applies, number of convictions, trial versus plea bargain, race, Hispanic, US citizen, gender, education, day of the week, fiscal year, and city-month and controls for offender age, offender age squared, other weather (dew point, atmospheric pressure, wind speed, precipitation, sky cover), and pollution (O_3, CO, $PM_{2.5}$). Weather variables are hourly measurements averaged over 6am–4pm local time, pollution variables are measured daily, and both are restricted to 20 miles around the courthouse. Missing values for pollution variables have been replaced with zero, and dummies for missingness have been added. Standard errors clustered on city-month are in parentheses; wild bootstrap-*t* (99,999 replications) 95 percent confidence intervals clustered on city-month (city) are in brackets (braces).

Sources: NOAA ISD Lite (weather, National Oceanic and Atmospheric Administration 2019), EPA AQS (pollution, US Environmental Protection Agency 2019), and USSC (other variables, Spamann 2017). The sample includes all federal sentencing decisions in fiscal years 1992–2003 with complete data (other than pollution).

For the probability of imprisonment, the effect of temperature is fairly precisely estimated to be zero: 0.12 percentage points per 10°F, with a 95 percent bootstrap confidence interval of only $[-0.16, 0.35]$ percentage points even when clustering on city. For sentence length conditional on imprisonment, the point estimate is almost exactly zero (0.01 percentage points per 10ºF), with a bootstrap confidence interval of $[-0.72, 0.61]$ when clustering on city. All weather variables are jointly statistically insignificant. In parallel work, Evans and Siminski (2021) estimate similarly precise zeroes for temperature effects on criminal adjudication in New South Wales, Australia. In sum, there is even less evidence for a temperature or other weather effect in sentencing than in asylum adjudication.

IV. Conclusion

This comment shows that the effects of temperature and other weather on judicial decision making are at most small and probably nonexistent in both asylum and sentencing. Contrary findings by AHSS resulted from errors in coding and data entry. Like other ostensible extraneous influences on judging (cf. Weinshall-Margel and Shapard 2011; Spamann 2018), the weather effect turns out to be spurious. Reports of such influences should be read with circumspection.

Judicial decision making is not perfect. There is evidence of biases and inconsistencies in judicial decisions in general (e.g., Rachlinski and Wistrich 2017; Spamann and Klöhn 2016) and in US asylum adjudication (Ramji-Nogales, Schoenholtz, and Schrag 2007; Fischman 2013) and federal sentencing (e.g., Yang 2015; Cohen and Yang 2019), in particular. It is essential, however, to gauge the extent of the problem accurately. It is dangerous to paint the justice system as worse than it actually is.

REFERENCES

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" NBER Working Paper No. 24003.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–27.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity." *Review of Economics and Statistics* 99 (4): 698–709.

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.

Cohen, Alma, and Crystal S. Yang. 2019. "Judicial Politics and Sentencing Decisions." *American Economic Journal: Economic Policy* 11 (1): 160–91.

Evans, Sally, and Peter Siminski. 2021. "The Effect of Outside Temperature on Criminal Court Sentencing Decisions." *Series of Unsurprising Results in Economics* 21 (1): 1–23.

Executive Office for Immigration Review. 2005. *FY 2004 Statistical Year Book*. Falls Church, VA: US Department of Justice.

Executive Office for Immigration Review. 2020. "EOIR Case Data (July 2019)." U.S. Department of Justice: Executive Office for Immigration Review [publisher], Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E120222V1.

Fischman, Joshua B. 2013. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." *American Law and Economics Review* 16 (1): 40–85.

Fischman, Joshua B., and Max M. Schanzenbach. 2012. "Racial Disparities under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums." *Journal of Empirical Legal Studies* 9 (4): 729–64.

Heyes, Anthony, and Soodeh Saberian. 2019a. "Replication data for: Temperature and Decisions: Evidence from 207,000 Court Cases." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E113722V1.

Heyes, Anthony, and Soodeh Saberian. 2019b. "Temperature and Decisions: Evidence from 207,000 Court Cases." *American Economic Journal: Applied Economics* 11 (2): 238–65.

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

Heyes, Anthony, and Soodeh Saberian. 2022b. "Replication Data for: "Correction to 'Temperature and Decisions: Evidence from 207,000 Court Cases' and Reply to Spamann." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E127263V1.

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.

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

National Oceanic and Atmospheric Administration. 2019. "Integrated Surface Data Lite." ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/ (accessed February 16, 2020).

Rachlinski, Jeffrey J., and Andrew J. Wistrich. 2017. "Judging the Judiciary by the Numbers: Empirical Research on Judges." *Annual Review of Law and Social Science* 13 (1): 203–29.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60 (2): 295–412.

Roodman, David, Morten Ørreaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata using Boottest." *Stata Journal* 19 (1): 4–60.

Spamann, Holger. 2017. "Replication Data for: Are Sleepy Punishers Really Harsh Punishers?: Comment." Harvard Dataverse V2. https://doi.org/10.7910/DVN/TZRNKD.

Spamann, Holger. 2018. "Are Sleepy Punishers Really Harsh Punishers? Comment on Cho, Barnes, and Guanara (2017)." *Psychological Science* 29 (6): 1006–09.

Spamann, Holger. 2022. "Replication Data for: Comment on 'Temperature and Decisions: Evidence from 207,000 Court Cases.'" Harvard Dataverse. https://doi.org/10.7910/DVN/3LOR3R.

Spamann, Holger, and Lars Klöhn. 2016. "Justice Is Less Blind, and Less Legalistic, than We Thought: Evidence from an Experiment with Real Judges." *Journal of Legal Studies* 45 (2): 255–80.
US Environmental Protection Agency. 2019. "Air Quality System (AQS)." https://www.epa.gov/aqs (accessed February 5, 2020).
Weinshall-Margel, Keren, and John Shapard. 2011. "Overlooked Factors in the Analysis of Parole Decisions." *Proceedings of the National Academy of Sciences* 108 (42): E833.
Yang, Crystal S. 2015. "Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing." *Journal of Legal Studies* 44 (1): 75–111.

This article has been cited by:

1. Anthony Heyes, 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, 529-533. [Abstract] [View PDF article] [PDF with links]