over-65 Medicare population has opted out of TM in the last decade and enrolled in private insurance plans through Medicare Advantage (MA). In MA, private insurers receive capitated payments from the government for providing Medicare beneficiaries with health insurance that roughly mimics commercial health insurance for the under-65 population. Today, almost a third of Medicare beneficiaries are enrolled in MA.

Empirical comparisons of MA and TM face two primary challenges. First, differences in health care utilization between patients in MA and TM may partly or entirely reflect differences in the patient mix, rather than a "treatment effect" of MA per se. Second, historically, data availability has been asymmetric: administrative claim-level data from TM are widely available to researchers, but detailed claim-level data from MA insurers have been more elusive. The primary contribution of this paper lies in our analysis of new, claim-level data from MA insurers. Specifically, we take advantage of newly available claims data from MA plans in 2010 provided by the Health Care Cost Institute (HCCI). The data consist of claims paid by three MA insurers (Aetna, Humana, and UnitedHealthcare) that cover almost 40 percent of MA enrollees. The key advantage of these data is that they contain claim-level data in MA—i.e., health care utilization and payments to providers—that is analogous to the existing and commonly used claims data for TM.

A simple tabulation of the MA and TM claims points to a large difference in public and private health care spending levels. We calculate that MA spending per enrollee-month in 2010 totaled \$642, of which \$590 was paid by MA insurers and the rest by enrollees out of pocket. In contrast, average spending per enrollee-month in TM was \$911, of which \$771 was paid directly by the Medicare program to providers. Capitated payments to the MA plans roughly track the latter amount; the MA plans in the HCCI data received on average \$767 per enrollee-month. In other words, the revenue of the MA plans we observe is 30 percent higher than the payments they make for their enrollees' health care. If this applied to the entire MA population in 2010 (including those outside our sample), it would imply \$21 billion in annual (2010) revenue for MA insurers in excess of their spending on health care claims.

The bulk of our analysis compares health care spending and utilization for enrollees in MA and TM. To proxy for what an MA enrollee's health care experience would have been like if she were (counterfactually) in TM, we construct a "comparable" group of TM enrollees. We present results from two main approaches. First, we adjust for key observables—comparing outcomes for MA and TM enrollees in the same county and with the same risk score. Medicare risk scores are based on a predictive model of health care spending that accounts for demographics and detailed information on prior health conditions. The county and risk score adjustment also captures the spirit in which Medicare sets reimbursement rates for MA insurers; these are the two dimensions that enter the formula by which capitation rates are computed. Second, we include an additional adjustment for unobserved health not captured by the risk score (Brown et al. 2014), which is based on mortality differences between MA and TM enrollees in the same county and with the same risk score. Without either adjustment, MA spending per enrollee-month is 30 percent lower than TM spending per enrollee-month. Holding county and risk score fixed, the spending difference becomes 25 percent, and adjusting for mortality data, prior studies have used a variety of approaches to infer health care utilization and spending differences between MA and TM. These include comparing MA plans' (mandatory) self reports of enrollee utilization to utilization measures in TM claims data (Landon et al. 2012), analyzing beneficiaries' self reports of care received in TM and in MA (Ayanian et al. 2013), analyzing hospital discharge data from New York counties experiencing MA exit (Duggan, Gruber, and Vabson 2018), and inferring cost differences from estimates of demand for MA plans and a supply-side model of the market (Curto et al. 2014). These papers have tended to find lower health care utilization in MA—with estimates ranging from 10 percent to 60 percent.

Our finding of similar pricing in MA and TM echoes a recent finding by Baker et al. (2016) and contrasts with the conventional wisdom that MA prices will be higher than TM prices due to the greater bargaining power enjoyed by the larger public sector (e.g., Philipson et al. 2010). It also differs from prior findings that TM prices are substantially lower than prices in the private, under-65 market both on the inpatient side (Cooper et al. 2015) and the outpatient side (Clemens and Gottlieb 2017). It seems plausible that the lower prices that private insurers pay for over-65 enrollees relative to under-65 enrollees is the consequence of regulation that is specific to the over-65 population, and requires hospitals to accept TM rates if an alternative payment rate was not negotiated (Berenson et al. 2015).

Our findings of similar geographic variation in spending and pricing in MA and TM also contrast with recent findings that geographic variation in spending in commercial (i.e., under-65) insurance is similar to TM, but stems from much larger pricing variation and lower quantity variation in commercial insurance relative to TM (Philipson et al. 2010, Institute of Medicine 2013, and Cooper et al. 2015). This contrast between TM and commercial insurance has been interpreted as reflecting the lower powered incentives in the public sector relative to the private sector in constraining utilization, and monopsony power in the public sector to constrain prices relative to what the private sector can achieve (Philipson et al. 2010). Of course, there are other reasons why patterns of health care provision for those under 65 may differ from the patterns for the over 65. We consider this same set of facts in the context of Medicare Advantage, which arguably provides a cleaner comparison group to TM for understanding variation under private and public regimes since MA and TM are provided to the same broad population.

Our finding that MA appears to reduce both "high-value" and "low-value" care in similar magnitude contributes to what we believe is an emerging, cautionary tale on the bluntness of policy instruments in the health care sector. Our evidence here speaks to the blunt nature of supply-side restrictions on care. Likewise, on the demand side, recent evidence suggests that high deductible plans reduce "high-value" and "low-value" care in equal measure (Brot-Goldberg et al. 2017), and that even targeted increases in the price of some types of care can depress care use across the board, including free preventive care services (Cabral and Cullen 2017).

Most broadly, our work is part of the large literature on the relative consequences of public and private ownership. This literature has spanned a range of disparate industries, including education, pensions, electricity, and transportation. In the specific context of health care, recent empirical work has emphasized that the private

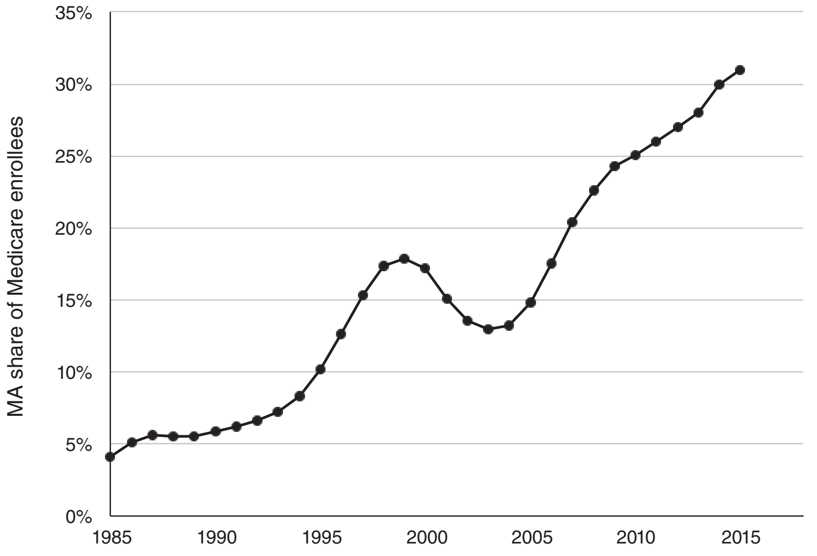

Figure 1. MA Penetration over Time

Notes: The figure shows the share of Medicare beneficiaries enrolled in Medicare Advantage plans, year by year. All data are from December of the year indicated.

Source: CMS' Medicare Managed Care Contract Plans Monthly Summary Reports

Every year, plans enter into a bidding process, which dictates the benefits and premium associated with each plan that is offered to beneficiaries. While the precise rules by which plan bids translate to plan premiums and benefits are somewhat complicated, we summarize the key features here (see Curto et al. 2014 for a more detailed description). Each plan submits a bid b, which should be interpreted as the monthly compensation required by the plan to provide "standard" monthly coverage in the local area in which the plan is offered to an "average" Medicare beneficiary. By "standard" coverage, we refer to the standard Part A and Part B financial coverage offered by TM; MA plans typically offer more comprehensive coverage, but they obtain a separate compensation for it (known as the "rebate") on top of their bid b. As will be clearer later, by "average" beneficiary we refer to a beneficiary with an average health risk.

This bid b is then assessed against its local benchmark B, which is set administratively by CMS. In principle, the benchmark B is supposed to approximate the counterfactual cost to CMS from covering an "average" beneficiary in that county through TM. In practice, the variation in benchmarks across locations departs somewhat from this principle, presumably reflecting various political economy considerations. On average in our observation period (2010), benchmark rates are higher than corresponding TM costs, and more so in some areas than in others; subsequent to our time period of analysis, the Affordable Care Act has reduced the level of these MA benchmark rates. Overall in our data (again, before applying the sample restrictions described below), the average benchmark across counties (weighted by the number of Medicare beneficiaries) is \$836 per enrollee-month, compared to an average TM cost of \$798, and this difference is lower in urban counties (benchmark

HCCI data also contain monthly enrollment indicators and some limited enrollee demographics (age bins, gender, and zip code).

The CMS data serve multiple roles. One role is to provide parallel claim-level data for Medicare beneficiaries enrolled in Traditional Medicare (TM). Because TM offers fee-for service coverage, we essentially observe every health care claim made by TM enrollees during 2010. The TM claims data allow us to form a "benchmark" comparison of health care spending and utilization against which we can compare the measures obtained from HCCI.

The CMS data have a second, equally important role: providing enrollment, demographic, health, and mortality data for all enrollees (TM and MA). For the universe of Medicare enrollees, we can observe monthly enrollment information in TM (Parts A and/or B) or MA, risk score, demographics (zip code, age, and gender), dual eligibility status (in Medicaid and Medicare), detailed health conditions from the prior year, and mortality. The detailed CMS data on MA enrollees allow us to validate the completeness of our baseline sample in HCCI, and to adjust our comparison to TM spending for the differential demographics, health conditions, and mortality among MA enrollees compared to TM enrollees.

Finally, the CMS data contain detailed information on payments to MA insurers by CMS. This allows us to construct payments to MA plans per enrollee-month, as well as payment components.

B. *Baseline Sample*

The HCCI data include most, but not all, MA enrollees covered by the three HCCI insurers. Based on the qualitative information that HCCI obtained from the three participating insurers, it appears that inclusion in the HCCI data was made on a plan-by-plan basis, with "highly capitated plans" left out. That is, insurance plans that pay providers on a capitated basis are omitted from the HCCI data. The HCCI data also indicate that they exclude special needs plans (SNPs), which are MA plans for individuals with specific diseases (such as end-stage liver disease, chronic heart failure, or HIV-AIDS) or certain characteristics (such as residence in a nursing home).

Ideally, we would have plan identifiers in the HCCI data, which would allow us to match this information to the plan identifiers in the CMS data, and thus know which MA plans are excluded. This would allow us to adjust for the demographics and health conditions of MA enrollees specifically enrolled in HCCI plans. However, with the exception of SNPs that are not in the HCCI data and can be identified in the CMS enrollment data, plan and insurer identifiers are omitted from the HCCI data. Instead, we rely on the fact that the MA market is localized and the use of provider capitation is most common in particular regions such as California and construct our baseline sample by focusing on states where the HCCI data coverage appears to be approximately complete.

We judge the completeness of the HCCI data by comparing enrollment statistics for the HCCI insurers in the HCCI and CMS data. In the CMS data, we know for each MA enrollee whether he or she was enrolled in an MA plan offered by one of the HCCI insurers. This allows us to generate a pseudo HCCI enrollment dataset in the CMS data, which covers all enrollees who "should" have been in the

Table 1—Baseline Sample

Data source / sample	All CMS[a]			Baseline CMS[b]			All HCCI[a]	Baseline HCCI[b]
	TM	MA (all insurers)	MA (HCCI insurers)	TM	MA (all insurers)	MA (HCCI insurers)	MA (HCCI insurers)	MA (HCCI insurers)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Enrollee-level summary[c]								
Number of enrollees (000s)	26,420	10,475	3,911	15,641	5,291	2,270	2,941	2,290
Female	0.575	0.574	0.574	0.576	0.567	0.568	0.569	0.571
Age	75.4	74.6	74.5	75.4	74.3	74.1	—	—
Coarse age:[d]								
65–74	0.520	0.555	0.560	0.516	0.568	0.581	0.592	0.590
75–84	0.330	0.328	0.325	0.333	0.323	0.315	0.306	0.308
85+	0.150	0.117	0.115	0.151	0.109	0.104	0.102	0.102
Dual eligible	0.143	0.123	0.111	0.129	0.072	0.073	—	—
SNP enrollees	—	0.081	0.065	—	0.000	0.000	0.000	0.000
Risk score	1.089	1.031	1.032	1.085	0.986	0.994	—	—
Died in 2010	0.050	0.039	0.039	0.052	0.036	0.036	—	—
Panel B. Spending per enrollee-month[e]								
Number of enrollee-months (000s)	304,908	118,737	44,371	180,608	60,273	25,867	32,506	25,394
Total spending ($/month)	938	—	—	911	—	—	639	642
Insurer spending ($/month)	798	—	—	771	—	—	586	590
OOP spending ($/month)[f]	140	—	—	140	—	—	53	52
Panel C. Payments to insurers per enrollee-month[e]								
Overall CMS expenditure ($)[g]	—	820	819	—	767	778	—	—
Actuarial value of incremental consumer benefits ($)[h]	—	63	53	—	56	51	—	—
Plan payments for organic MA services ($)[i]	—	800	806	—	751	767	—	—

Notes: The table presents summary statistics for various sample definitions. Columns 6 and 8, highlighted in gray, are comparable and are used to validate our sample construction.

[a] Sample includes all Medicare enrollees who are 65 or older by the end of 2010.

[b] Baseline sample excludes SNP enrollees, and enrollees in the 15 states in which the number of enrollee-months in HCCI is not within 10 percent of that in CMS.

[c] At the enrollee level, we define an individual as enrolled in TM if she is never enrolled in MA during the sample year and is enrolled in TM for at least one month of the sample year; we define her as enrolled in MA if she is enrolled in MA in any month of the year, and we assign her to an HCCI insurer if she is covered by one of them in her first month in MA. Age, dual eligibility, and SNP enrollment is likewise defined based on the first month in which an enrollee is observed during the sample year.

[d] In HCCI, we only have information about age in 3 bins: 65–74, 75–84, and 85+.

[e] We count an enrollee-month in TM if she is enrolled in TM that month and never enrolled in MA during the sample year; any enrollee-months in MA (or in HCCI insurers) are counted as such.

[f] Out-of-pocket (OOP) spending denotes amount owed by enrollee. For TM enrollees, OOP spending may be partially covered by supplemental (Medigap or employer-sponsored) coverage.

[g] This includes all payments made from CMS to the MA plans, including risk-adjusted payments and rebates.

[h] This is also known as the "rebate."

[i] The variable "Plan payments for organic MA services ($)" is equal to "Overall CMS expenditure ($)" plus additional premiums paid by the beneficiaries minus the non-cost-sharing component of the rebate.

(column 6) and the actual HCCI data (column 8). This is what we would expect given our construction of a baseline sample for which the HCCI data should include all relevant MA enrollees.[2]

[2] We have about 1 percent more enrollees in our HCCI sample (column 8) than the pseudo HCCI sample in the CMS data (column 6). This is to be expected, given that plan assignment is missing for about 1 percent of MA enrollees in the CMS data.

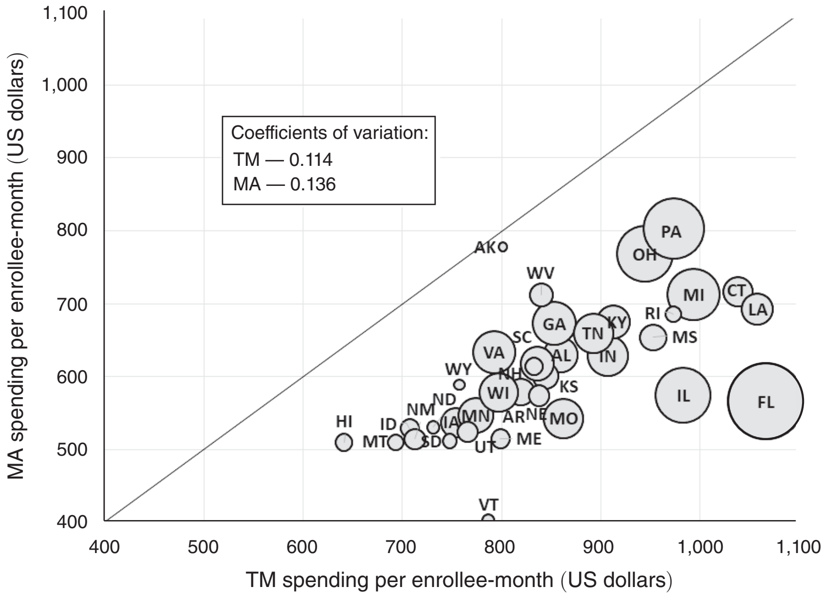

Figure 2. State-by-State Comparison of TM and MA Spending

Notes: The figure plots MA spending per enrollee-month against TM spending per enrollee-month for each of the 36 states in our baseline sample. Coefficients of variation across states in spending are computed using total Medicare enrollees in the state as a weight. The size of each bubble is proportional to the number of total Medicare enrollees in the state.

for Medicare (including the federal government's costs of administering Medicare Advantage) were about 1.7 percent of Medicare TM claims in 2010 (Boards of Trustees 2011).

Spending in MA and TM: *Raw Comparisons*.—The raw summary statistics also show dramatic differences in total health care spending between the TM and MA populations. In our baseline sample, the average TM enrollee spends \$911 per month (Table 1, column 4), while the average MA enrollee spends 30 percent less, \$642 (Table 1, column 8).

Figure 2 shows raw spending in MA and TM separately for each of the 36 states in our baseline sample. Spending is lower in MA in all states, but the differences range from about 3 percent lower MA spending in Alaska to over 45 percent lower MA spending in Florida and Vermont.

Geographic variation in spending within TM has attracted a great deal of attention. The "Dartmouth Atlas" finding of large differences across areas in TM spending and utilization without corresponding differences in mortality is widely viewed as indicative of the inefficiencies of the public Medicare system (Fisher et al. 2003a, b; Skinner 2011; and Institute of Medicine 2013). Our analysis suggests that, if

of insurance). Using these data, we focused on the three HCCI insurers in the 36 states that constitute our baseline sample. Our estimate of 8 percent is the overall ratio in this sample between the sum of "general and administrative costs" to the sum of "net incurred claims after reinsurance."

B. *Selection on Observables*

Selection on "Priced" Observables.—In our first empirical strategy to correct for selection, we re-weight the TM population to match the MA population in terms of county and risk score. Within the above framework, it can be viewed as assuming that, conditional on county and risk score, MA_i is as good as random assignment. The risk score is a summary statistic based on an extremely rich set of demographic and health measures. These health measures reflect both patient health and propensity to receive health care—since diagnoses are only recorded if care is received (Song et al. 2010; Finkelstein, Gentzkow, and Williams 2016)—both of which may differ between TM and MA enrollees.

Specifically, consider a Medicare enrollee in county z_i with (continuous) risk score r_i, and an outcome y_i^{TM} in TM. We map r_i to a discrete risk score bin r_i', so that all Medicare beneficiaries are partitioned into a set of discrete groups, defined by their county and risk score bin $g_i = (z_i, r_i')$. Using the sample of beneficiaries in the CMS data who are enrolled with the HCCI insurers (Table 1, column 6), we assign each group g a weight $w_g = N_g/N$, where N_g is the number of enrollees that belong to group g and $N = \sum_g N_g$.[8] Each unweighted TM outcome

$$\overline{y}^{TM}_{unweighted} = \frac{1}{N_{TM}} \sum_{i \in TM} y_i^{TM} \tag{4}$$

is then replaced with a re-weighted TM outcome

$$\overline{y}^{TM}_{re\text{-}weighted} = \frac{1}{\sum_{i \in TM} w_{g_i}} \sum_{i \in TM} w_{g_i} y_i^{TM}, \tag{5}$$

which we compare to the corresponding MA outcome

$$\overline{y}^{MA} = \frac{1}{N_{MA}} \sum_{i \in MA} y_i^{MA}. \tag{6}$$

In addition to the transparency and simplicity of this re-weighting approach, it has the added attraction that it captures the spirit by which MA insurers are being paid by CMS. As described in Section I, CMS payments to MA insurers are based on a county-specific benchmark, and multiplied by the enrollee's risk score r_i. Our baseline approach, which re-weights on precisely these two dimensions—county and risk score—can therefore be viewed as correcting for selection concerns associated with the two dimensions by which CMS varies its payments. As mentioned above, following CMS' payment policy for MA insurers during our 2010 study year, we use risk scores for MA enrollees that are deflated by 3.41 percent.

[8] A slight complication of this procedure arises when an MA enrollee belongs to a group for which there are no TM enrollees, which may happen in small counties and high (i.e., less common) risk scores. This applies to only 0.07 percent of enrollee-months. In such a case, we amend this procedure with an extra step, where we reclassify to such "empty" TM groups the TM enrollee in the same county whose risk score is the closest to the corresponding unmatched MA enrollee.

in spending between MA and TM reflects a treatment effect of MA as opposed to selection into MA by individuals who—conditional on risk score and county—have lower predicted spending due to unmeasured differences in health or preferences for health care. The relative importance of selection or treatment is particularly important in the context of assessing the cost implications of any expansion of the MA program to cover those currently enrolled in TM.

If we want to condition on a richer set of variables, it gets more difficult to apply the same re-weighting strategy as the data become sparse and it is common to observe MA beneficiaries with a vector of characteristics for which there is no match in the TM sample. We therefore instead follow a standard approach of constructing propensity scores for enrollment in MA as a function of a rich set of observables, and then apply the re-weighting strategy to the propensity score rather than to the entire vector of variables.

Specifically, given a vector of observables x_i, we estimate a logit model of MA_i on x_i. That is, we assume that $p_i = \Pr(MA_i = 1) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$ and estimate β by maximum likelihood. We estimate the logit model separately for each county, to allow the relationship between enrollment in MA and observables to differ across counties. We then use our estimate of β to generate the propensity score for individual i, denoted by $\hat{p}_i$. Online Appendix Figure A3 presents the distribution of the propensity score for the TM and MA populations. We then repeat the same re-weighting procedure used in the first empirical strategy, but now with respect to $g_i' = (z_i, \hat{p}_i')$, where the propensity score $\hat{p}_i$ is binned into bins $\hat{p}_i'$ of width 0.01. That is, instead of assuming that (conditional on county) the risk score captures all relevant information that may affect selection, we now replace it with the propensity score of enrolling into MA.

A critical decision, obviously, regards the set of variables x_i that enter the propensity score calculation. The risk score r_i is based on a rich set of observables, including very detailed health measures as well as age, gender, and dual eligibility in Medicaid. These observables are used with a particular functional form to produce the risk score. Using the same underlying variables to generate the propensity score is a natural and less restrictive way to correct for selection in our setting. In practice, however, the results in Table 2, panel B, column 2, show that this approach yields quite similar results to our first approach that adjusts only for selection on priced observables, which we reproduce in Table 2, panel B, column 1.

C. *Using Mortality to Address Selection on Unobservable Health*

It is less obvious how to correct for selection on unobservables that affect the propensity to enroll in MA and may also be correlated with health care spending. Our main approach to address it is to leverage the fact that we can observe mortality outcomes for individuals in both TM and MA. As we saw in Table 1, mortality is lower for MA enrollees than for TM enrollees; it is also lower conditional on county and risk score (not shown). While clearly imperfect, it may provide a rough sense as to how much additional selection may affect the interpretation of the results, and this could vary for different types of health care utilization outcomes.

Table 3—Spending Differences for Different Groups of Enrollees

	Percent MA enrollees (1)	TM, unweighted (2)	TM, weighted[a] (3)	TM, mort. weighted[a] (4)	MA (5)	Difference ((5) − (3))/(3) (6)	Difference ((5) − (4))/(4) (7)
Number of enrollee-months (000s)	25,394	180,608	180,608	180,608	25,394		
Total spending	100%	911	855	706	642	−24.9%	−9.0%
Panel A. Spending ($/month) by enrollee characteristics							
Male	43%	916	857	696	673	−21.4%	−3.3%
Female	57%	907	853	713	619	−27.4%	−13.2%
65–74	56%	723	661	534	540	−18.2%	1.2%
75–84	33%	1,022	967	874	731	−24.4%	−16.4%
85+	11%	1,264	1,276	1,137	898	−29.6%	−21.0%
Urban[b]	77%	942	887	733	645	−27.3%	−12.0%
Rural[b]	23%	851	752	622	634	−15.7%	1.9%
Panel B. Realized distribution of spending ($/month)							
Proportion with no spending		0.37	0.38	0.43	0.46	19.6%	7.7%
Median spending		93	84	64	38	−54.3%	−40.2%
75th percentile		332	317	262	222	−30.0%	−15.0%
90th percentile		1,314	1,233	977	849	−31.1%	−13.1%
95th percentile		3,433	3,124	2,396	2,161	−30.8%	−9.8%
97.5th percentile		8,349	7,571	5,835	5,690	−24.8%	−2.5%
99th percentile		18,510	17,332	14,672	13,614	−21.5%	−7.2%

Notes: Results are based on baseline sample (see Table 1, columns 8 and 4). All statistics are at the enrollee-month level. All spending numbers are in dollars/month.

[a] Weighting based on our preferred weighting, as in column 4 of both panels in Table 2.

[b] Rural/urban assignment is based on whether the enrollee zip code is in an MSA.

in 2010 (column 2 of Table 1, which includes those outside of our baseline sample), this translates to $101.5 billion in annual (2010) health care spending in TM relative to $76.3 billion in health care spending in MA, a difference of $25.2 billion in annual health care spending.

The differences are still positive, but not as large, if in addition we adjust for unobserved health. Doing this (Table 2, panel B, column 4) indicates that health care spending in MA is only $64 (9 percent) lower than in a comparable (on county and predicted mortality rate) sample of TM enrollees. Recall that MA insurers are paid based on risk scores, so the higher difference in spending that arises from adjusting for selection on priced observables (Table 2, panel A) is more directly associated with the profits of MA insurers from the current set of MA enrollees, while using mortality to adjust for unobserved health may be more relevant in the context of a counterfactual of moving MA enrollees to TM (or vice versa), assuming that it indeed captures much of the selection on unobserved health.

In the remaining tables, we compare differences across types of consumers or care. The relative patterns are similar with either adjustment approach, although naturally the quantitative differences are smaller across the board when we additionally adjust for unobserved health.

Differences by Consumer Type.—Panel A of Table 3 reports the spending differences for different types of enrollees. Each row represents a different subsample

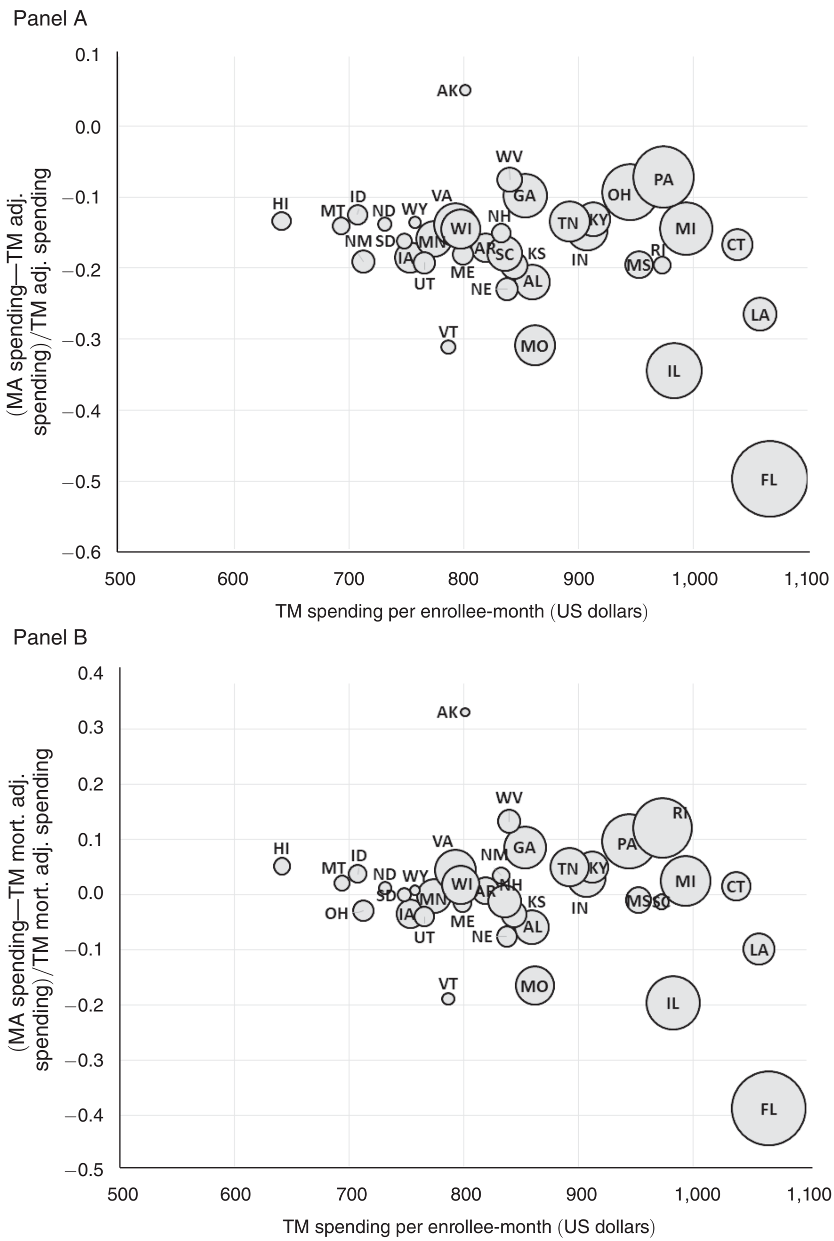

FIGURE 3. TM-MA SPENDING DIFFERENCES ACROSS STATES

Notes: The figure plots the (percentage) difference between average MA spending and (re-weighted) TM spending per enrollee-month against average TM spending for each of the 36 states in our baseline sample. The y-axis in panel A compares MA spending to TM spending that is re-weighted to match the MA population on county and risk score, using our preferred weighting (see Table 2, panel A, column 4). Panel B does the same but using predicted mortality to adjust for selection on unobservables (see Table 2, panel B, column 4), as described in Section III. The size of each bubble is proportional to the number of total Medicare enrollees in the state. The x-axis reports average (unadjusted) TM spending in the state (see Table 2, panel A, column 1).

Table 5—Differences in Health Care Utilization

	TM, unweighted	TM, weighted[a]	TM, mort. weighted[a]	MA	Difference	
					((4) − (2))/(2)	((4) − (3))/(3)
	(1)	(2)	(3)	(4)	(5)	(6)
Total spending ($/month)	911	855	706	642	−24.9%	−9.0%
Inpatient days	0.200	0.181	0.143	0.144	−20.6%	0.4%
Any inpatient admission	0.027	0.025	0.020	0.021	−16.0%	3.7%
Days conditional on any	7.4	7.4	7.2	6.9	−5.5%	−3.1%
Skilled nursing facility (SNF) days	0.336	0.296	0.219	0.131	−55.9%	−40.3%
Days conditional on any	47.3	46.7	45.4	20.6	−55.8%	−54.6%
Emergency department (ED) visits	0.049	0.045	0.037	0.038	−15.8%	1.9%
Outpatient ED visits	0.031	0.028	0.024	0.024	−14.8%	−0.2%
Inpatient ED visits	0.018	0.017	0.013	0.014	−17.5%	5.6%
Physician visits	1.22	1.21	1.10	1.01	−16.8%	−8.0%
Any physician visits	0.545	0.540	0.503	0.486	−10.0%	−3.5%
Number of visits conditional on any	2.24	2.25	2.18	2.08	−7.5%	−4.6%

Notes: Results are based on baseline sample (see Table 1, columns 8 and 4). All statistics are at the enrollee-month level, but all days associated with a given encounter are attributed to the original admission date, even if it extends beyond the month.

[a] Weighting based on our preferred weighting, as in column 4 of both panels in Table 2.

visits in an outpatient setting are also lower in MA than in TM, with the difference approximately equally driven by the extensive margin (a lower rate of MA enrollees who see a physician at least once a month) and the intensive margin (a lower average number of physician visits by MA enrollees who visit the physician at least once).

Interestingly, additional adjustment for unobserved health essentially eliminates utilization differences for inpatient-related measures, just as it did for inpatient-related spending (Table 4). This pattern is consistent with our adjustment for unobserved health fully adjusting for health differences between TM and MA enrollees, and MA insurers having no discretion over inpatient utilization, which is fully driven by health events.

Overused and Underused Care.—In Table 6, we explore differences in potential low-value and high-value care. Panel A examines utilization of diagnostic testing and imaging services, where overuse may be a concern (e.g., Brot-Goldberg et al. 2017, US Government Accountability Office 2008). Table 6, panel B, examines utilization of various measures of preventive care, an area where underuse may be a concern (Brot-Goldberg et al. 2017).[11] We see lower utilization in MA for both low-value and high-value care. Diagnostic tests and imaging procedures are lower in MA by similar percentages as total spending. Preventive care exhibits no obvious pattern relative to overall care; rates of most preventive care are lower in MA, although there is variation across the measures.

[11] We show rates of preventive care by enrollee-month to be consistent with the analysis in the rest of the paper. Naturally, recommended care is not at a monthly level but typically at an annual (or biannual) level. The analysis looks similar if instead we examine these measures on an annual basis (not shown).

Table 7—Differences in Spending per Episode of Care

	TM, unweighted	TM, weighted[a]	TM, mort. weighted[a]	MA	Difference	
					((4) − (2))/(2)	((4) − (3))/(3)
	(1)	(2)	(3)	(4)	(5)	(6)
Total spending ($/month)	911	855	706	642	−24.9%	−9.0%
Spending per SNF day	381	379	383	378	−0.2%	−1.4%
Spending per outpatient ED visit	782	768	760	837	9.0%	10.1%
Inpatient[b]						
Spending per admission	10,134	10,151	10,206	10,093	−0.6%	−1.1%
Spending per day	1,901	1,903	1,950	1,908	0.3%	−2.1%

Notes: Results are based on baseline sample (see Table 1, columns 8 and 4). All statistics are at the enrollee-month level, but all expenditures or days associated with a given encounter are attributed to the original admission date, even if it extends beyond the month.

[a] Weighting based on our preferred weighting, as in column 4 of both panels in Table 2.

[b] Inpatient spending here includes only payments to the hospital; it does not include associated physician payments as in prior tables.

in discouraging both high-value and low-value care utilization (Brot-Goldberg et al. 2017) and Medicaid coverage for the previously insured encourages increases in ED visits of all types, including (and perhaps particularly) nonemergency visits (Taubman et al. 2014).

B. (*Lack of*) *Mean Price Differences for Hospital Admissions for Specific Diagnoses*

Table 7 shows spending per encounter in MA and TM. Given the close similarity between the percentage difference in utilization measures in Table 5 and the percentage difference in the corresponding spending measures in Table 4, it is not surprising that spending per encounter is quite similar between MA and TM. Inpatient spending per admission, inpatient spending per day, and SNF spending per SNF day are essentially the same in MA and TM. Interestingly, spending per outpatient ED visit is 9–10 percent higher in MA; this may reflect utilization management for MA patients that discourages relatively less severe cases from coming to the ED or from being admitted from the ED to the hospital. We also note that neither re-weighting approach makes much difference for inpatient spending; the spending per encounter statistics are quite similar already in the raw comparison of means.

This similar spending per encounter for MA and TM enrollees suggests that prices may be similar in MA and TM. However, spending per encounter can also be affected by differences in providers seen or in reason for the visit. To hone in on differences in "prices"—or unit payment rates—we compare payments in MA and TM for admission to the *same* hospital with the *same* DRG.[12] Under TM, hospitals are paid by CMS based on a preset formula that is a product of a hospital-specific

[12] For this pricing analysis, we focus on the approximately 4,000 hospitals in our baseline sample that are paid (by TM) under Medicare's prospective payment system (PPS). These represent about 95 percent of all inpatient admissions in MA and cover essentially all standard (non-specialty) hospitals.

Figure 4 shows our estimate of the average price in TM and MA overall, and for the top 20 DRGs (by their share of MA admissions); online Appendix Table A2 provides the underlying numbers. In reporting DRG-specific average prices, we weight the admissions in each DRG by the state's share of MA admissions in all DRGs, so that any differences in average prices across DRGs within MA (or within TM) reflect price differences for a common "state basket," and are not contaminated by differences in the geographic distribution of admissions by DRG across states. The national average price is computed by weighting each DRG by its (national) share of MA admissions.

Inpatient prices are extremely similar in MA and TM. The national average admission price is \$9,945 in TM and \$10,054 in MA. The price for an average MA admission is only 1.1 percent higher in MA relative to TM. The largest difference among the top 20 DRGs is for chest pain (DRG #313), for which the average MA price is about 6 percent lower than in TM. For 10 of the top 20 DRGs, the average price in MA is within 2 percent of that in TM.

The close similarity of inpatient admission prices between MA and TM echoes similar findings by Baker et al. (2016) and is interesting given that it is frequently conjectured that because the public sector has greater bargaining power, public fee-for-service may achieve lower prices than private insurance (e.g., Philipson et al. 2010). Consistent with this conjecture, prior empirical work has shown that for the same service, TM tends to reimburse at substantially lower prices than commercial (under-65) private insurance both in the outpatient setting (Clemens and Gottlieb 2017) and the inpatient setting (Cooper et al. 2015). In contrast, we do not find that TM prices are substantially lower than MA prices.[14] One potential explanation for this discrepancy is that regulation requires hospitals to accept fee-for-service Medicare rates for Medicare beneficiaries when they are not included in the MA plan's network; as a result, MA plans may have greater bargaining power—and thus obtain lower rates—than commercial plans that serve the under-65 population. Berenson et al. (2015) provides more details on this institutional environment, and reports on results from a survey of hospital and MA plan executives, which are very consistent with our findings.

Geographic Variation in Hospital Prices.—We also compare geographic variation in inpatient prices for MA and TM. We construct average state prices in MA and TM following a parallel process to what we did for measuring DRG prices; here, we weight the admissions in each state using the DRG's national share of MA admissions, so that comparisons of state-level average prices within MA (or within TM) are not contaminated by differences in the mix of DRGs across states.

Figure 5 shows the results; online Appendix Table A3 shows the underlying numbers. Pricing variation across states (weighted by Medicare enrollment) is about 20 percent lower in MA than in TM. Specifically, the coefficient of variation across states is 0.067 in MA, compared to 0.082 in TM. By contrast, recent work has

[14] Of course, our MA sample is limited to three large insurers, and their bargaining power may not be representative of smaller MA insurers; however, Cooper et al.'s (2015) analysis of commercial pricing was also limited to the same three large insurers, and in that study, average inpatient prices were almost twice as high as in TM.

Table 8—Potential Channels for Cost Saving

	TM, unweighted (1)	TM, weighted[a] (2)	TM, mort. weighted[a] (3)	MA (4)	Difference ((4) − (2))/(2) (5)	Difference ((4) − (3))/(3) (6)
Panel A. Hospital discharge destinations						
Home	0.0136	0.0122	0.0104	0.0109	−10.4%	5.4%
Home health service organization	0.0053	0.0049	0.0039	0.0038	−23.3%	−4.2%
SNF	0.0067	0.0061	0.0047	0.0038	−37.6%	−17.5%
Other post-acute care	0.0014	0.0013	0.0010	0.0004	−70.5%	−63.4%
Other (including hospice, death)	0.0027	0.0024	0.0018	0.0018	−27.3%	−2.9%
Panel B. Surgeries and specialists						
Total surgeries	0.037	0.033	0.029	0.039	18.1%	33.0%
Outpatient surgeries	0.029	0.026	0.023	0.032	25.5%	41.2%
Inpatient surgeries	0.008	0.007	0.007	0.007	−7.2%	4.8%
Primary care visits	0.379	0.370	0.334	0.355	−3.8%	6.5%
Specialist visits	0.840	0.844	0.764	0.655	−22.4%	−14.3%

Notes: Results are based on baseline sample (see Table 1, columns 8 and 4). All statistics are at the enrollee-month level. All spending numbers are in dollars/month. Panel A reports (unconditional) hospital discharge destinations.
[a] Weighting based on our preferred weighting, as in column 4 of both panels in Table 2.

By contrast, in TM there are virtually no restrictions on physician clinical decisions or patient choices of care.

We have already seen evidence of one "signature" of MA mechanisms to reduce care utilization: all these mechanisms should constrain patient entry into care, particularly expensive care, so that the average person using that care in MA is in worse health, and has higher cost than the average person using that care in TM. In other words, MA enrollees should have fewer encounters, but have greater spending (or utilization) per encounter. Consistent with this, we found that spending per outpatient ED visit was in fact slightly higher in MA than in TM (see Table 7).

In Table 8, we provide additional evidence consistent with restrictions on utilization. In panel A, we explore differences between TM and MA in the distribution of discharge destinations of hospitalized patients. Destinations are roughly ordered in how expensive they are (from cheaper to more expensive). Inpatients covered by MA are disproportionately discharged to less expensive destinations. In particular, discharges to SNFs (or other post-acute care) are substantially less common, while discharges home (or to home health services) are relatively more likely.

In addition to limiting use of care, MA may also constrain the type of service, encouraging use of less expensive substitutes. Panel B points to some patterns that are suggestive of such channels. First, we analyze the frequency of surgeries. We find the surgery rate to be in fact higher, not lower, in MA by about 20 to 30 percent. However, inpatient surgeries are similar and outpatient surgeries are much higher, which is suggestive of MA insurers using outpatient surgeries to substitute away from inpatient surgeries and perhaps (given the fact that overall number of surgeries is higher) from other types of expensive, nonsurgical admissions as well. Second, we examine two types of physician visits: primary care and specialist visits. We already saw in Table 5 that MA enrollees are associated with fewer physician visits. The results in Table 8 show that this is driven primarily by fewer specialist visits; rates of primary care visits are similar.

preference argument would suggest that consumers who choose MA are better off in MA than in TM. Other inferences are harder to make. Quality of the health care experience is difficult to assess; our measures of preventive care point to reductions there that are similar in magnitude to those for other forms of care. We calculated that the mean actuarial benefit to consumers (i.e., rebates that are passed on to consumers in the form of other benefits) was $51 per enrollee-month, but, of course, the rebate may be valued differently from its actuarial value, and MA plans have other attributes that will affect consumer surplus, such as limited networks. The implications of privately provided Medicare for both consumers and producers is an important area for further work.

REFERENCES

Ayanian, John Z., Bruce E. Landon, Robert C. Saunders, L. Greg Pawlson, and Joseph P. Newhouse. 2013. "Quality of Ambulatory Care in Medicare Advantage HMOs and Traditional Medicare." *Health Affairs* 32 (7): 1228–35.

Baker, Laurence C., M. Kate Bundorf, Aileen M. Devlin, and Daniel P. Kessler. 2016. "Medicare Advantage Plans Pay Hospitals Less Than Traditional Medicare Pays." *Health Affairs* 35 (8): 1444–51.

Boards of Trustees. 2011. *2011 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds.* Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. Washington, DC, May.

Berenson, Robert A., Jonathan H. Sunshine, David Helms, and Emily Lawton. 2015. "Why Medicare Advantage Plans Pay Hospitals Traditional Medicare Prices." *Health Affairs* 34 (8): 1289–95.

Billings, John, Nina Parikh, and Tod Mijanovich. 2000. *Emergency Department Use: The New York Story.* Commonwealth Fund. New York, November.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin Handel, and Jonathan T. Kolstad. 2017. "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics." *Quarterly Journal of Economics* 132 (3): 1261–1318.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? Evidence from the Medicare Advantage Program." *American Economic Review* 104 (10): 3335–64.

Cabral, Marika, and Mark R. Cullen. 2017. "The Effect of Insurance Coverage on Preventive Care." *Economic Inquiry* 55 (3): 1452–67.

Centers for Medicare and Medicaid Services (CMS). 2009. "Announcement of Calendar Year (CY) 2010 Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies." https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/announcement2010.pdf.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2017. "In the Shadow of a Giant: Medicare's Influence on Private Physician Payments." *Journal of Political Economy* 125 (1): 1–39.

Clemens, Jeffrey, Joshua D. Gottlieb, and Timea Laura Molnár. 2017. "Do Health Insurers Innovate? Evidence from the Anatomy of Physician Payments." *Journal of Health Economics* 55: 153–67.

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen. 2015. "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured." National Bureau of Economic Research (NBER) Working Paper 21815.

Curto, Vilsa, Liran Einav, Amy Finkelstein, Jonathan Levin, and Jay Bhattacharya. 2019. "Health Care Spending and Utilization in Public and Private Medicare: Dataset." *American Economic Journal: Applied Economics.* https://doi.org/10.1257/app.20170295.

Curto, Vilsa, Liran Einav, Jonathan Levin, and Jay Bhattacharya. 2014. "Can Health Insurance Competition Work? Evidence from Medicare Advantage." National Bureau of Economic Research (NBER) Working Paper 20818.

Duggan, Mark, Jonathan Gruber, and Boris Vabson. 2018. "The Consequences of Health Care Privatization: Evidence from Medicare Advantage Exits." *American Economic Journal: Economic Policy* 10 (1): 153–86.

Einav, Liran, Amy Finkelstein, and Maria Polyakova. 2018. "Private Provision of Social Insurance: Drug-Specific Price Elasticities and Cost Sharing in Medicare Part D." *American Economic Journal: Economic Policy* 10 (3): 122–53.

American Economic Journal: Applied Economics 2019, 11(2): 238–265
https://doi.org/10.1257/app.20170223

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

By Anthony Heyes and Soodeh Saberian*

We analyze the impact of outdoor temperature on high-stakes decisions (immigration adjudications) made by professional decision-makers (US immigration judges). In our preferred specification, which includes spatial, temporal, and judge fixed effects, and controls for various potential confounders, 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. Results are consistent with established links from temperature to mood and risk appetite and have important implications for evaluating the influence of climate on "cognitive output." (*JEL* K37, K41, Q54)

We investigate the link from outdoor temperature to decisions made by experienced professional decision-makers, working in good-quality, climate-controlled, indoor spaces. If decisions with durable consequences are systematically influenced by irrelevant factors, the potential for welfare loss is obvious. The question we investigate is the following: do decision outcomes, the substance of which have nothing to do with contemporaneous temperature, depend causally on how hot it is outside on the day the decision is made? Examining the universe of files (just under 207,000) evaluated over a four-year period by the 266 immigration court judges at the 43 US Federal Immigration Courthouse locations spread across most major US cities our answer is a resounding yes—with high significance and robustness, and a substantial effect size. As such, we evidence a subtle and pernicious channel through which variations in climate (across space and through time) can damage well-being, by influencing decisions.

The analysis contributes to our developing understanding of how decisions can be sensitive to apparently irrelevant considerations. For example, Mani et al. (2013)

* Heyes: Department of Economics, University of Ottawa, 120 University Private, Ottawa, Ontario, Canada, K1N 6N5 (email: aheyes@uottawa.ca); Saberian: PhD candidate, Department of Economics, University of Ottawa, 120 University Private, Ottawa, Ontario, Canada, K1N 6N5 (email: ssabe101@uottawa.ca). Alexandre Mas was coeditor for this article. Heyes is Canada Research Chair (CRC) in Environmental Economics at University of Ottawa and part-time Professor of Economics at the University of Sussex. He acknowledges financial support from the CRC Program and from SSHRC under Insight Grant # 435-2012-472. We are grateful to Alberto Salvo, Ben Olken, Maya Papineau, Matthew Neidell, John List, Patrick Baylis, Pierre Brochu, Sandeep Kapur, Jason Garred, Abel Brodeur, two referees from this journal, and seminar participants at CREE 2017 at Ivey Business School at Western University, University of Exeter, University of Sussex, National University of Singapore, and McGill University for helpful conversations. Errors are ours.

† Go to https://doi.org/10.1257/app.20170223 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

- The quality of data and the procedural details of the immigration system allow us to avoid a plethora of identification challenges, allowing for clean, persuasive causal inference.

Our main approach uses high-frequency data to estimate a linear probability model with a variety of fixed effects, though we also provide some nonparametric results. In addition, we develop variants in which the independent variables of interest are the Heat Index (a measure used by the US National Weather Service that combines temperature and humidity nonlinearly into a metric designed to capture how hot it "feels") and the difference between realized temperature on a particular date and local norms for that date. Our central identifying assumption is minimal—that temperature realizations are as good as random after accounting for spatial and temporal fixed effects.

The analysis uncovers a substantial effect of short-term (daily) variations in temperature on decision outcomes. In our preferred specification, which include city-by-month and judge fixed effects, as well as controls for case characteristics and other potential environmental confounders, same-day, outdoor temperature has an impact on decision outcomes. Our results suggest that a 10°F increase in temperature reduces the likelihood of a decision favorable to the applicant by 1.075 percent, which is equivalent to a 6.55 percent decrease in the grant rate (the grant rate in the data as a whole is 16.4 percent). To put this into perspective, in our sample, the difference in grant rate between a judge at the twenty-fifth percentile in terms of leniency, and one at the seventy-fifth percentile, is 7.9 percent. Consistent with some existing studies of temperature susceptibility varying by gender (Yu et al. 2010; Xiong et al. 2015) the effect is particularly pronounced for female judges. To allay concerns that there might be something unique to the immigration setting that is driving the results, we repeat the exercise for decisions made in 18,461 Parole Suitability Hearings at the 39 locations of the California Department of Corrections and Rehabilitation (CDCR), arriving at parallel conclusions.

Why are these results important? As a straight piece of law and economics the research contributes to an assessment of the consistency of US immigration (and California parole) practices. The Sixth Amendment to the US Constitution lays out "fair trial" as a fundamental right. The Administrative Procedurs Act (APA) (1946) determines that any adjudication or decision by an agent of the US government should not be "arbitrary or capricious." Agency decisions should be "... rationally connected to the facts before it" (Committee on Capital Markets Regulation (CCMR), 2016, 2). The immigration court system is "about" decisions, and natural justice, as well as the law, dictates that decisions on a particular file should be based solely on the merits of the case ("the facts and nothing but the facts"). There is no plausible reason why a particular file should have any different prospect of success

settings. For example, in some professions, an employee might choose to defer work from a hot day to a cooler day (or work in the evening), or decide to work at home in response to weather conditions.

learning in classrooms. Seppänen et al. (2006) conducts a meta-analysis of the 24 papers that a particular search protocol elicits on this topic (including those just listed). Of these, 9 take place in the lab, the rest are in offices or schools, and between them they generate just over 100 effect size estimates. Their systematic review of the literature generates an estimate of the indoor temperature associated with highest productivity being at 21.75°C (71.5°F), with a decrement of performance of around 9 percent when temperature is 30°C (86.1°F).[5] In general, heat stress has a much greater influence than does cold stress on the performance of cognitive tasks (see Hancock and Vasmatzidis 2003 for a review).

Turning to decision-making in particular, Cheema and Patrick (2012) presents five studies of consumer behavior in which they manipulate laboratory temperatures. In higher temperatures subjects are: (i) less likely to engage in gambles (particularly complex gambles), (ii) less likely to choose innovative products over established ones, and (iii) more likely to rely on "system 1" (heuristic or habit-based) processing (Pocheptsova et al. 2009). In our setting, in which the rejection rate of immigration applications is around 83 percent, such that the granting an applicant leave to stay can plausibly be regarded as the less-habitual, more innovative, and more risky choice, this would point to a negative relation between high temperatures and grant rates.

While evidence of the effect of contemporaneous indoor temperature on brain-intensive tasks is suggestive for us, none of it is directly relatable. Studies that cast light on how daily *outdoor* temperature affects indoor mental performance are rare. Graff Zivin et al. (2018) finds that an (outdoor) temperature above 79°F on a particular day damages performance of children on math (but not reading) tasks. Park (2016) investigates the relationship between daily outdoor temperature and high school exit exams in New York City, and finds that compared to a 72°F day, taking an exam on a 90°F day reduces a typical student's performance by 0.19 standard deviations.

Turning away from cognition, separate strands of research evidence: (a) a causal link from ambient temperature (and other dimensions of weather) to "mood," broadly defined, and then; (b) a causal link from mood to decision-making. Baylis (2015) links temperature to measures of hedonic state (mood) using geo-located Twitter activity. His four sentiment metrics based on phraseology, emoticon use, and profanity each become more negative once outdoor temperatures exceed 70°F (with little to no effect for colder temperatures). Denissen et al. (2008) finds a similar effect when they analyze online diary entries of 1,233 students. Relatedly, a number of behavioral finance papers (for example, Hirshleifer and Shumway 2003, Cao and Wei 2005, Floros 2011) link daily variations in weather—typically cloud cover and sunshine, but also temperature and humidity—to stock price movements via changes in emotional state.

[5] The first of these numbers accords with anecdotal introspection. In a more recent review, Cheema and Patrick (2012, 985) notes that: "Prior studies find that an ambient temperature of 72°F, one at which most people appear comfortable, may be most conducive for automatic tasks." For instance, Allan et al. (1979) finds that performance on a paired-association memory task peaks at 72°F. Other evidence suggests a difference between temperatures that are optimal for comfort and those that are optimal for performance. Specifically, Pepler and Warner (1968) shows that people perform office work best at 68°F, although they report feeling cold."

Figure 1. Location of Immigration Courts (Excluding Honolulu)

female. The mean grant rate (the rate at which a decision is made that favors the applicant) in the database as a whole is 16 percent.

Our data comes from asylumlaw.org. Asylumlaw no longer operates but was: "A website run by an international consortium of agencies that helps asylum seekers in Australia, Canada, the United States, and several countries in Europe. It provides links to legal and human rights resources, experts, and other information valuable for asylum seekers." [8] The data contains date of hearing, identity of judge, nationality of applicant, and category of application.[9]

Asylum decisions made by immigration judges are decisive and those that are denied asylum are subject to removal. Judges sit alone, and there are no formal quotas with respect to their grant rate. While the activities of judges are subject to the overall supervision of the US Attorney General, this is an area of law in which individual judges are widely regarded as having a high degree of personal discretion and independence in the way in which they evaluate files (see Ramji-Nogales et al. 2007, and Chen et al. 2016). Though the characteristics of cases that judges in different locations are likely to see will of course vary, the degree of discretion is supported anecdotally by the wide variation in grant rates of judges both between and within particular courthouses. For instance, over the study period in the Los Angeles courthouse there are five judges that granted asylum to fewer than 4 percent, while three others granted in over 67 percent.

Judges typically determine multiple cases on a given day. The judge is presented with a file, may (or may not) ask questions of the applicant, then enters an adjudication. Within a court all cases are in principle randomly assigned to the judges

[8] The dataset was kindly provided by Kelly Shue (University of Chicago Booth School of Business) in personal correspondence.

[9] There are two types of cases in immigration courts: affirmative cases in which the applicant presents in the courts on her/his own and defensive cases in which the applicant is instructed to attend on the initiative of the immigration authorities.

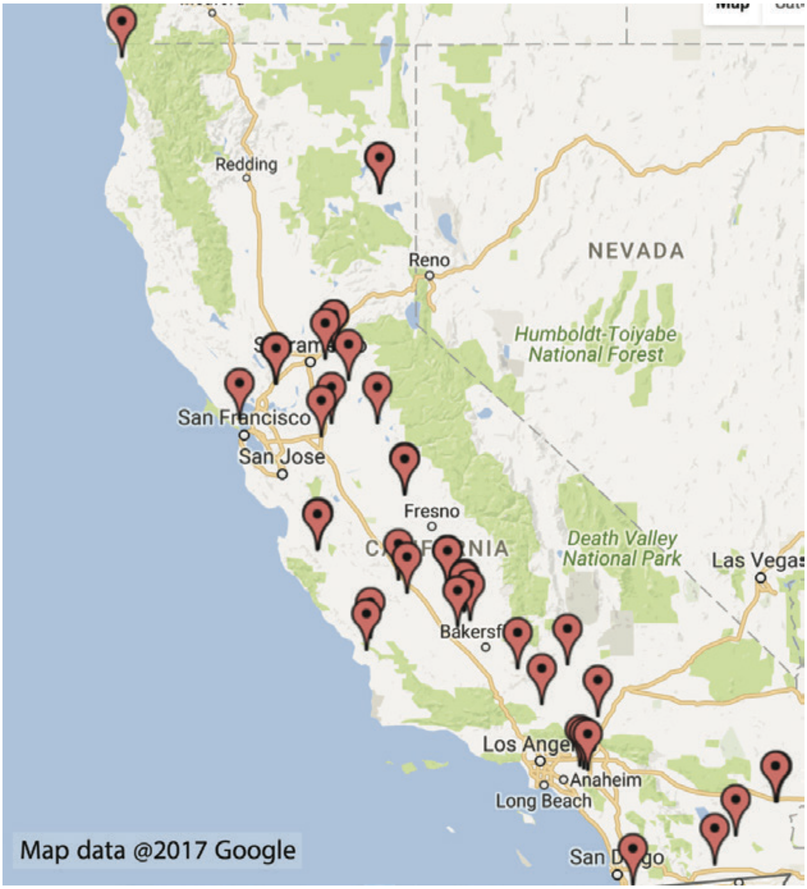

Figure 2. Location of Parole Hearing Venues

Our data contains the date of hearing, identity of panel members, inmate unique identifier, location of hearing, hearing type, and outcome.[13]

C. *Environment*

Our main research question is whether the adjudication on a file responds to the outdoor temperature on the day on which it is evaluated. To accomplish this, we combine our decision dataset with temperature and a variety of other environmental controls.

The location of asylum decisions from which we construct our dependent variable is drawn from the 43 US cities in which the US Department of Justice operates immigration courthouses. These are widely dispersed (see Figure 1) and subject to diverse weather conditions.

[13] There are two types of hearing that we consider: (i) Initial parole (which is scheduled one year before eligibility), and (ii) Subsequent parole that is scheduled if there is any consideration in the initial session.

Table 1—Summary Statistics

	Mean	SD
Grant indicator	0.164	0.371
Temperature (°F)	57.37	15.721
Heat index (°F)	57.77	16.423
Air pressure (pa)	29.688	0.759
Dew point (°F)	49.372	17.202
Precipitation (mm)	0.003	0.014
Wind speed (km/h)	4.557	3.441
Sky cover (percent)	55.44	0.276
Ozone (ppm)	0.0220	0.0120
CO (ppm)	0.917	0.496
$\text{PM}_{2.5}$ (μ/m^3)	14.957	11.569

We will also be controlling for air-quality conditions. Daily pollution data is published online by the United States Environmental Protection Agency (USEPA).[16] The dataset includes daily measures of particulate matter less than 2.5 microns in width $(PM_{2.5})$, carbon monoxide (CO), and ozone (O_3) throughout the United States for the period of 2000 to 2004.

Table 1 presents summary statistics.

III. Methods

A. *Empirical Strategy*

We estimate the following linear probability model:

$$(1) \qquad g_{it} = \beta_0 + \beta_1 temp_{it} + W_{it}\beta_2 + P_{it}\beta_3 + X_{it}\beta_4 + \gamma_i + \Psi_{ct} + \theta_t + \epsilon_{it},$$

where g_{it} is a binary variable that takes the value one if the judge's decision in an asylum application i on date t is granted, zero otherwise.

The key independent variable is the mean 6 AM to 4 PM temperature on the date the case is considered, $temp_{it}$. For most of our discussion, β_1 is the coefficient of interest.

To allow for the possibility that other dimensions of weather rather than temperature might impact decisions, we include a vector of weather controls, W_{it}. It contains dew point temperature (a standard measure of humidity), precipitation, wind speed, air pressure, and sky cover on date t, in the vicinity of the courthouse in which application i is adjudicated, all calculated on a 6 AM to 4 PM average basis. Pollution exposure can also influence cognitive function, mood, and/or decision-making (Archsmith et al. 2018, Chang et al. 2019, Ebenstein et al. 2016). To allow for this possibility we include P_{it}, which is a vector of pollution controls. It comprises mean daily measures of ozone (O_3), carbon monoxide (CO), and particulate matter $(PM_{2.5})$.

[16] The data is available at https://aqs.epa.gov/api.

Table 2—Fixed Effect Estimates: 6am–4pm Average

	Preferred (1)	1-Day lag (2)	1-Day lead (3)	All (4)
Temperature$_t$/1,000	−1.075	−1.454	−1.208	−1.617
	[0.274]	[0.406]	[0.382]	[0.486]
Temperature$_{t-1}$/1,000	–	0.361	–	0.372
	–	[0.278]	–	[0.277]
Temperature$_{t+1}$/1,000	–	–	0.139	0.159
	–	–	[0.260]	[0.260]
F-statistic of joint significance of weather variables	3.41	3.07	2.99	2.73
p-value	0.0026	0.0036	0.0044	0.0059
Observations	206,924	206,924	206,924	206,924

Notes: The unit of analysis is an immigration case. Dependent variable is a dummy taking value one if decision is favorable to applicant, zero otherwise. Temperature is the 6 am to 4 pm average in the city in which the case is adjudicated, on the day of adjudication, in Fahrenheit. The temperature measure is divided by 1,000 to reduce decimal places. All regressions control for weather, pollution, and time fixed effects. Weather covariates include dew point, air pressure, wind speed, precipitation, and cloud cover measured as 6 am to 4 pm averages in the city in which the case is adjudicated, on the day of adjudication. Pollutant covariates include controls for ozone, carbon monoxide, and $PM_{2.5}$, measured as calendar daily averages at the air quality monitoring station closest to the courthouse of adjudication, on the day of adjudication. Time fixed effects include day of week and year dummies relating to the day of adjudication. Regressions also include city-month fixed effects, name of judge adjudicating case, type of application, and nationality of applicant. Sample is all cases adjudicated at 42 mainland US federal immigration courthouse locations from January 1, 2000 to September 30, 2004. Standard errors are clustered on city-month in brackets.

IV. Results

A. *Linear*

The base results are summarized in Table 2. Column 1 is the preferred specification, incorporating the full suite of controls: time fixed effects, weather, and pollution controls.[21]

The coefficient in column 1, −1.075, implies that a 10°F increase in 6 am–4 pm temperature on the day a decision is made reduces the likelihood of a grant decision by 1.075 percent. Recall that the average grant rate in the sample is 16.4 percent, so this implies a 6.55 percent decrease in grant rate. The effect of a 10°F rise in temperature is comparable in size to those found by Eren and Mocan (2018) for an unexpected loss by the local NCAA football team (which induced a temporary 6.4 percent increase in severity of juvenile sentencing). Several studies point to between-judge variation in asylum grant rates (Ramji-Nogales et al. 2007 and Chen 2017). In our sample, the difference in grant rate between a judge at the twenty-fifth percentile in terms of leniency, and one at the seventy-fifth percentile, is 7.9 percent.

Columns 2 and 3 of Table 2 report the results of including lag or lead. In each case the point estimates on the lagged terms are much smaller in absolute value than those on the main measure, mixed in sign, and never approach significance

[21] All of our main specifications are estimated on the whole 58 months of data. The terrorist attacks of September 11, 2001 fall during our study period and can be expected to have impacted the operation of the immigration system in the United States. While we do not report them here, we have run the main specifications on the pre- and post-9/11 portions of the dataset, observing consistent patterns across them.

Table 3—Alternative Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Temperature_t/1{,}000$	−0.717	−0.727	−0.780	−0.806	−1.037	−0.893	−1.082	−0.939	−1.075
	[0.270]	[0.273]	[0.269]	[0.249]	[0.278]	[0.215]	[0.271]	[0.285]	[0.274]
Hausman-test	0.76	0.69	0.44	0.63	0.40	0.36	0.90	0.09	–
p-value	0.384	0.406	0.506	0.426	0.528	0.549	0.343	0.760	–
Observations	206,924	206,924	206,924	206,924	206,924	206,924	206,924	206,924	206,924
Nationality FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Day of week FEs	N	Y	Y	Y	Y	Y	Y	N	Y
Type of application FEs	N	N	Y	Y	Y	Y	Y	Y	Y
Judge FEs	N	N	N	Y	Y	Y	N	Y	Y
City-month FEs	N	N	N	N	Y	N	N	Y	Y
Judge-month FEs	N	N	N	N	N	N	Y	N	N
City FEs	N	N	N	N	N	Y	Y	N	N
Year FEs	N	N	N	N	N	N	Y	Y	Y
Year-month FEs	N	N	N	N	N	Y	N	N	N
Date FEs	N	N	N	N	N	N	N	Y	N

Notes: The unit of analysis is an immigration case. Dependent variable is a dummy taking value one if decision is favorable to applicant, zero otherwise. Temperature is the 6 AM to 4 PM average in the city in which the case is adjudicated, on the day of adjudication, in Fahrenheit. The temperature measure is divided by 1,000 to reduce decimal places. All regressions control for weather and pollution. Weather covariates include dew point, air pressure, wind speed, precipitation, and cloud cover measured as 6 AM to 4 PM averages in the city in which the case is adjudicated, on the day of adjudication. Pollutant covariates include controls for ozone, carbon monoxide, and $PM_{2.5}$, measured as calendar daily averages at the air quality monitoring station closest to the courthouse of adjudication, on the day of adjudication. Each specification contains other controls as indicated. Column 9 coincides with column 1 from Table 2, our preferred specification. Sample is all cases adjudicated at 42 mainland US federal immigration courthouse locations from January 1, 2000 to September 30, 2004. Standard errors are clustered on city-month in brackets.

Table 4 explores the sensitivity of results to some alternative but plausible specifications.

Much of the related literature on short-term effects of weather and air quality on human outcomes has used the calendar day as its unit of analysis (for example, Hirshleifer and Shumway 2003, Ebenstein et al. 2016, and Park 2016). While this is not our preferred approach, a substantial portion of each calendar day occurs after the court is closed; for example, for comparability, we report in Table 4, column 2 the results of repeating the exercise on a calendar day basis. As would be expected given the introduction of additional imprecision into the way in which the regressor of interest is measured, the estimated coefficients are attenuated somewhat, but retain sign and significance, and are similar in magnitude to Table 2 (−0.750 instead of −1.075 for the preferred specifications).[24]

Decision locations are dispersed widely across the country and in places that exhibit very different weather patterns. This implies that a 90°F day in Phoenix may not have the same effect as such a day in Boston. The inclusion of city-month and year fixed effects should control for unobservable characteristics of that location at that time of year (such as "normal" weather conditions). However, to probe this further, we estimate a variant in which the independent variable of interest is the deviation of 6 AM–4 PM temperature on decision day from the average 6 AM–4 PM temperature for that location in that week of the year. The results of this exercise are

[24] In a further variant, we conducted the exercise using eight-hour blocks (midnight–8 AM, 8 AM–4 PM, 4 PM–midnight). The results (not reported here) parallel those presented.

In an additional exercise, we explore the role of the gender of the judge. For this exercise we reestimate the preferred regression specifications on the subsample of decisions made by female judges (72,229 decisions made by 95 individuals) and male judges (134,695 decisions made by 171 individuals) separately. In Online Appendix Table A.2. the results of these exercises are summarized in columns 2 and 3, respectively. In each case the point estimate is negative and significant at the 5 percent level. However, the female coefficient is around 6 percent bigger in absolute value. The Hausman test (reported in the lower panel of onine Appendix Table A.2) confirms that the coefficient values are significantly different at the 5 percent level (p-value 0.0325). This is consistent with prior research that temperature-sensitivity is particularly pronounced amongst females (Yu et al. 2010, Xiong et al. 2015). The result also goes some way to address a concern that the patterns that we observe are driven not by the effect of temperature on judgement, but that temperature is instead influencing outcomes by impacting (for example) the comportment of the applicant or his lawyer. If that (or other external-to-judge mechanisms) were the channel we would not expect to see differences based on the gender of the judge.

B. *Nonlinear*

In addition to the conventional linear estimate, we also examine possible nonlinearity in the relationship between temperature and decision outcomes by reestimating using temperature bins 5°F in width, with the 50–55°F bin as the reference category.

The results of this exercise are presented in column 1 of Online Appendix Table A.2 and illustrated in Figure 5. Point estimates are statistically significant and positive when temperature is in the range of 25–30°F and 40–45°F and negative when it exceeds 55°F. They are also meaningful in size. Other things equal, taking a case heard on a day where outdoor temperature is between 50–55°F and dropping it instead into a day where the temperature exceeds 85°F reduces the likelihood of a favorable decision by 6.31 percent.

The negative effects of temperature appear close-to-linear and most of the robustness checks and other exercises that we conduct below will be centered on the linear results.

C. *Robustness*

Table 5 reports the results of a battery of robustness tests.

Pollution.—Recent research points to a possible link from short-term pollution exposure to mood and cognitive function, either of which might influence decision outcomes (Heyes et al. 2016; and Szyszkowicz et al. 2010). While our main specifications include controls for ambient levels of the main pollutants (O_3, $PM_{2.5}$, and CO), concern may remain that we have failed to control adequately for air-quality effects, and that these are confounding our results. If that were the case, then we would expect dropping the whole set of pollution controls to substantially affect our estimate of β_1. In column 1, we report the result of reestimating the preferred specification, but omitting the vector of pollution controls. The estimated coefficient on temperature retains sign and significance, and value changes only a little (-0.910 instead of -1.075).

California.—Of our 43 venues, 6 are located in California (accounting for around 32 percent of all decisions). To rule out that we are picking up something idiosyncratic to California, particularly since our external validity exercise is going to rely on Californian parole data, we reestimate our preferred specification excluding decisions made at courts in that state. This excludes around 71,000 of the 207,000 decisions in the sample. The result of this exercise are reported in column 2 of Table 5. Again, when estimated on the restricted sample, the estimate of β_1 retains sign and significance, and is little-changed in value (-1.159 instead of -1.075). So the pattern that we observed in the data is not being 'driven' by anything particular to California.

Weather.—Columns 3, 4, and 5 probe further the potential confounding role of rain and cloud.

Existing research points to cloud cover as influencing mood (Lambert et al. 2002, Kent et al. 2009, and Hirshleifer and Shumway 2003). We include a continuous variable that captures extent of cloud cover in our main specification to control for this. However, as a further test we reestimate the central specification on those decisions made on "clear sky" days, the subset of days when daily cloud cover is less than 5 percent (results in column 3). The point estimate of β_1 for the subsample estimation remains negative and significant. Though larger in absolute value (-2.738 instead of -1.075), suggesting that elevated temperature has a more pronounced impact on the decision on blue sky days versus non-such days, the difference between the two values is not significant at the 5 percent level.

Similarly rain can influence mood (Denissen et al. 2008). While a continuous measure of precipitation is included in the vector of weather controls, column 4 reports the result of reestimating the preferred specification on the subset of decisions (133,890 of them) made on days in which local recorded precipitation is zero. On such days rain cannot plausibly be argued to have influenced outcomes. The estimated coefficient retains sign and significance and is changed slightly in absolute value (-1.304 compare to -1.075). Column 5 reports the results of pushing this further by repeating the same exercise this time excluding days on which recorded precipitation on either the day of decision or the day before were nonzero (111,361 decisions). Again, the point estimate on the coefficient of interest is somewhat larger in absolute value (-1.281 instead of -1.075), but retains sign and significance.

Heat Index (*HI*).—The way in which temperature is experienced by the human body can itself depend on the water content of the air. Humidity is known to affect both mood and labor productivity (Howarth and Hoffman 1984, Tsutsumi et al. 2007, and Wan et al. 2009). We therefore investigate the joint effect of temperature and humidity in our setting by dropping temperature and dew point from our preferred specification and replacing it with the Heat Index (HI). The HI is used by the US National Weather Service and combines air temperature and relative humidity, via a nonlinear algorithm, into a single metric designed to capture how close to linear pattern with the negative effect becoming significant for values of HI exceeding 80°F.

Outlier Judges.—We note in the data section that judges do not have specific quotas with respect to what their grant rates should be; indeed this is an area of the legal system in which judges, sitting alone, are regarded as exercising a very high degree of personal discretion (Ramji-Nogales et al. 2007). To convince ourselves that the result that we are claiming are not being driven by "extreme" judges, we conduct two outlier analyses.[26] In the first, we exclude those decisions made by judges who have a grant rate across the whole study period in either the top or the bottom quartile (just retaining the 'middle half' of judges when ranked in terms of moderation).[27] Column 8 of Table 5 reports the results of this exercise, again, sign and significance is retained and the value of the coefficient is little disturbed (-0.707 instead of -1.075). In the second, we conduct the same exercise but exclude the top and bottom deciles of judges.[28] The results of this is reported in column 9 of Table 5. Again, the sign and significance is retained and the value of the coefficient is little disturbed (-1.064 instead of -1.075).

D. *Placebos*

As further falsification tests we perform three placebo exercises.[29] First, we replace the decision-day temperature series with the temperature at the same location 100 days after decision day, and 100 days before. Second, we replace the decision-day temperature in the vicinity of the courthouse in which the decision was made with the temperature on the same day, but taken from the weather monitoring station *most distant from it* "as the crow flies." For example, for Hartford (Connecticut) the placebo temperature is taken from the NOAA measuring station at Davenport (California) 4,238.72 miles away and for Dallas (Texas) the placebo temperature values are taken from Port Angeles (Washington) 2,792.42 miles away.

The results of these exercises are reported in Table 6. In each case the absolute value of the estimate of the coefficient of interest is several times smaller, signs are mixed, and in no case is statistical significance achieved.

E. *Parole*

Until now we have focused on judges evaluating immigration files. We are not going to claim broad generality of results, though we believe they are highly

[26] For example, suppose there existed a judge who is so extreme that he never found in favor of the applicant (his grant rate was 0 percent). The grant rate of that judge could not go lower upon exposure to high temperature because he is already at the lower bound. Recall that we already have judge fixed effects in all of our main specifications.

[27] This excludes decisions made by judges who have overall grant rates below 8.1 percent or above 22 percent.

[28] This excludes decisions made by judges who have overall grant rates below 4.7 percent or above 31 percent. Note that while we exclude the top and bottom decile of judges, we do not lose exactly 20 percent of our sample of decisions. This is because different judges are associated with different numbers of decisions across the study period.

[29] For this exercise, we limit analysis to mainland US locations (exclude weather stations in Puerto Rico and Hawaii). We ran a wide variety of other placebos with similar (insignificant) results.

Table 7—Parole: Calendar Day

	Preferred (1)	1-Day lag (2)	1-Day lead (3)	All (4)
$Temperature_t/1{,}000$	−1.560	−2.188	−1.586	−2.378
	[0.468]	[0.779]	[0.746]	[1.116]
$Temperature_{t-1}/1{,}000$	–	0.763	–	0.802
	–	[0.720]	–	[0.752]
$Temperature_{t+1}/1{,}000$	–	–	0.0319	0.194
	–	–	[0.762]	[0.793]
Observations	18,461	18,461	18,461	18,461

Notes: The unit of analysis is a parole case. Dependent variable is a dummy taking value one if decision is favorable to applicant, zero otherwise. Temperature is daily average at the monitoring station closest to the decision venue, in Fahrenheit. The temperature measure is divided by 1,000 to reduce decimal places. All regressions control for weather, pollution, and time fixed effects. Weather covariates include dew point, air pressure, wind speed, precipitation, and cloud cover daily averages. Pollutant covariates include controls for ozone, carbon monoxide and nitrogen dioxide, measured as daily averages at the air quality monitoring station closest to the venue of decision on the date of decision. Time fixed effects include day of week and year dummies relating to the day of decision. Regressions also include venue-month fixed effects, commissioners' name, type of application, and name of inmate. Sample consists of data on all parole hearings conducted by the Board of Parole Hearing (BPH) between January 3, 2012 and December 18, 2015 is from the California Department of Corrections and Rehabilitation (CDCR). Standard errors are clustered on venue-month in brackets. significant at 10 percent significant at 5 percent significant at 1 percent.

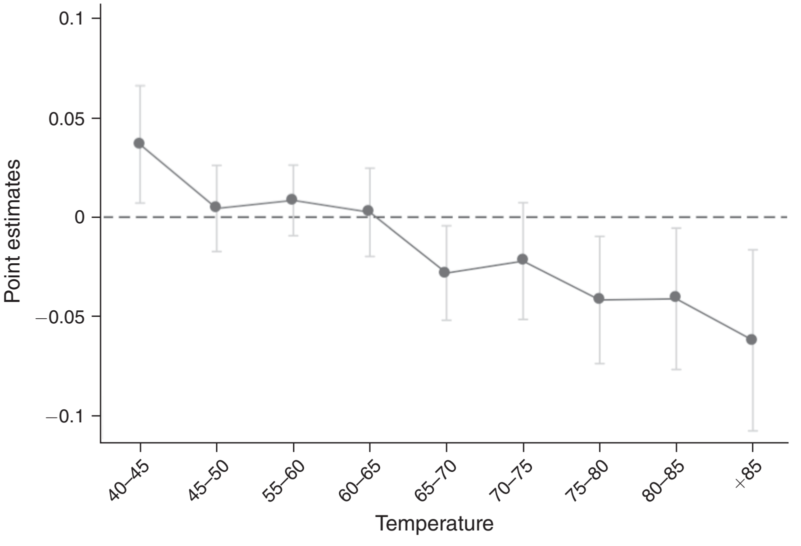

Figure 7. Nonlinear Estimates: Parole, Temperature, Calendar Day

Notes: This figure plots the coefficients on the temperature indicator variables from estimation of a nonlinear variant of the specification reported on column 1 from Table 7. The nonlinear variant replaces the continuous temperature measure with a series of temperature indicator variables of width 5° Fahrenheit. Gray lines show the 95 percent confidence intervals based on standard errors clustered on venue-month.

V. Conclusions

Temperatures vary across space and through time. We present what we believe to be the first evidence, in either a naturally occurring or artificial setting, that words that this is from a societal perspective of a high-stakes setting, the large effect sizes imply that the welfare losses are, in turn, large.

Away from the world of work, decisions are central to human well-being. We all routinely make decisions about what to buy, how to invest, how to vote, when to quit our jobs, etc. If decisions with durable impacts are systematically affected by irrelevant, transient factors, then the potential for individual and welfare loss across many settings is obvious.

One area in which we have been agnostic throughout the paper is channels. Pinning down the mechanism(s) from outdoor temperature to indoor decision processes would be a useful ambition of future work, and probably initially best-suited to laboratory or laboratory-in-the-field methods. The two broad channels that we noted in the introductory review that are consistent with the results relate to mood and cognitive acuity. High temperatures may stimulate temper, irritability (for example in Baylis 2015, Twitter users are more likely to use profanity), and other emotions that might induce a judge to be less well-disposed toward a typical applicant. In addition, depressive mood has been linked to reduced risk appetite. In both the immigration and parole settings denying a request can be plausibly regarded as the risk-averse course of action. Mental fatigue and other effects of heat can reduce mental acuity, which can increase mistakes, and also themselves induce transient increases in risk aversion.

Just as we have sought not to oversell the results, neither should we overstate the limitations. It is widely believed that world average temperatures are rising, as are the frequency of very hot and very cold days. Understanding the full set of social and economic outcomes that extreme temperature can influence is crucial to forming a measured view of the implications of such climate change. That outdoor temperature can have a large, significant, and apparently robust effect on indoor decisions, even when subjects operate in a climate-controlled setting, has potential for how we think about the links from climate to human well-being. The bounds on those effects, and the mechanisms underpinning them, are important foci of ongoing research.

REFERENCES

Abd-Elfattah, Hoda M., Faten H. Abdelazeim, and Shorouk Elshennawy. 2015. "Physical and Cognitive Consequences of Fatigue: A Review." *Journal of Advanced Research* 6 (3): 351–58.

Administrative Office of the United States Courts. 1996. *Standard Level Features and Finishes for U.S. Courts Facilities.* Washington, DC: US General Services Administration.

Allan, J.R., T.M. Gibson, and R.G. Green. 1979. "Effect of Induced Cyclic Changes of Deep Body Temperature on Task Performances." *Aviation, Space, and Environmental Medicine* 50 (6): 585–89.

Allen, Margaret A., and Gloria J. Fischer. 1978. "Ambient Temperature Effects on Paired Associate Learning." *Ergonomics* 21 (2): 95–101.

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 Resource Economists* 5 (4): 827–63.

Ariely, Dan, and George Loewenstein. 2006. "The Heat of the Moment: The Effect of Sexual Arousal on Sexual Decision Making." *Journal of Behavioral Decision Making* 19 (2): 87–98.

Baylis, Patrick. 2015. "Temperature and Temperament: Evidence from a Billion Tweets." Energy Institute at HAAS Working Paper 265.

Cao, Melanie, and Jason Wei. 2005. "Stock Market Returns: A Note on Temperature Anomaly." *Journal of Banking and Finance* 29 (6): 1559–73.

Jahedi, Salar, Cary Deck, and Dan Ariely. 2017. "Arousal and Economic Decision Making." *Journal of Economic Behavior and Organization* 134: 165–89.

Kahol, Kanav, Mario J. Leyba, Mary Deka, Vikram Deka, Stephanie Mayes, Marshall Smith, John J. Ferrara, and Sethuraman Panchanathan. 2008. "Effect of Fatigue on Psychomotor and Cognitive Skills." *American Journal of Surgery* 195 (2): 195–204.

Kent, Shia T., Leslie A. McClure, William L. Crosson, Donna K. Arnett, Virginia G. Wadley, and Nalini Sathiakumar. 2009. "Effect of Sunlight Exposure on Cognitive Function among Depressed and Non-depressed Participants: A REGARDS Cross-Sectional Study." *Environmental Health* 8 (1): 34–36.

Lambert, G.W., C. Reid, D.M. Kaye, G.L. Jennings, and M.D. Esler. 2002. "Effect of Sunlight and Season on Serotonin Turnover in the Brain." *Lancet* 360 (9348): 1840–42.

Loewenstein, George. 1996. "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes* 65 (3): 272–92.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341 (6149): 976–80.

Mathot, Kimberly J., Marion Nicolaus, Yimen G. Araya-Ajoy, Niels J. Dingemanse, and Bart Kempenaers. 2015. "Does Metabolic Rate Predict Risk-Taking Behaviour? A Field Experiment in a Wild Passerine Bird." *Functional Ecology* 29 (2): 239–49.

Oakley, J. B., and A.F. Coon. 1986. "The Federal Rules in State Courts: A Survey of State Court Systems of Civil Procedure." *Washington Law Review* 61 (5): 1367–68.

O'Brien, Erin M., and Jodi A. Mindell. 2005. "Sleep and Risk-Taking Behavior in Adolescents." *Behavioral Sleep Medicine* 3 (3): 113–33.

Park, Jisung. 2016. "Temperature, Test Scores, and Educational Attainment." Unpublished.

Pepler, R.D., and R.E. Warner. 1968. "Temperature and Learning: An Experimental Study." *ASHRAE Transactions* 74 (2): 211–19.

Pocheptsova, Anastasiya, On Amir, Ravi Dhar, and Roy F. Baumeister. 2009. "Deciding without Resources: Resource Depletion and Choice in Context." *Journal of Marketing Research* 46 (3): 344–55.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2010. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60 (2): 295–411.

Seppänen, O., W.J. Fisk, and Q.H. Lei. 2006. "Room Temperature and Productivity in Office Work." *Proceedings of Healthy Buildings* 1: 243–47.

Simon, Dan. 2012. *In Doubt: The Psychology of the Criminal Justice Process*. Cambridge, MA: Harvard University Press.

Szyszkowicz, Mieczysław, Jeff B. Willey, Eric Grafstein, Brian H. Rowe, and Ian Colman. 2010. "Air Pollution and Emergency Department Visits for Suicide Attempts in Vancouver, Canada." *Environmental Health Insights* 4: 79–86.

Tchen, Nadine, Helen G. Juffs, Fiona P. Downie, Qi-Long Yi, Hanxian Hu, Irene Chemerynsky, Mark Clemons, et al. 2003. "Cognitive Function, Fatigue, and Menopausal Symptoms in Women Receiving Adjuvant Chemotherapy for Breast Cancer." *Journal of Clinical Oncology* 21 (22): 4175–83.

Tsutsumi, Hitomi, Shin-ichi Tanabe, Junkichi Harigaya, Yasuo Iguchi, and Gen Nakamura. 2007. "Effect of Humidity on Human Comfort and Productivity after Step Changes from Warm and Humid Environment." *Building and Environment* 42 (12): 4034–42.

US Senate and House of Representatives. 1946. *Administrative Procedure Act, Pub. L. 79–404, 60 Stat. 237.* Washington, DC: US Senate and House of Representatives.

Viner, Russell M., Charlotte Clark, Stephanie J.C. Taylor, Kam Bhui, Emily Klineberg, Jenny Head, Robert Booy, and Stephen A. Stansfeld. 2008. "Longitudinal Risk Factors for Persistent Fatigue in Adolescents." *Archives of Pediatrics and Adolescent Medicine* 162 (5): 469–75.

Wan, J.W., Kunli Yang, W.J. Zhang, and J.L. Zhang. 2009. "A New Method of Determination of Indoor Temperature and Relative Humidity with Consideration of Human Thermal Comfort." *Building and Environment* 44 (2): 411–17.

Weaver, Leslie Jo, and Craig Hadley. 2009. "Moving Beyond Hunger and Nutrition: A Systematic Review of the Evidence Linking Food Insecurity and Mental Health in Developing Countries." *Ecology of Food and Nutrition* 48 (4): 263–84.

Weinreb, Linda, Cheryl Wehler, Jennifer Perloff, Richard Scott, David Hosmer, Linda Sagor, and Craig Gundersen. 2002. "Hunger: Its Impact on Children's Health and Mental Health." *Pediatrics* 110 (4): 41–50.

Wyer, Robert S., and Donal E. Carlston. 1979. *Social Cognition, Inference, and Attribution*. Hillsdale, NJ: Lawrence Erlbaum Associates.