Researchers' Degrees-of-Flexibility and the Credibility of Difference-in-Differences Estimates: Evidence From the Pandemic Policy Evaluations

Joakim A. Weill[1],[*] Matthieu Stigler[2], Olivier Deschenes[3] and Michael R. Springborn[4]

[1] Agricultural and Resource Economics, University of California, Davis, USA. Email: jweill@ucdavis.edu

[2] Center on Food Security and the Environment, Stanford University, USA

[3] Department of Economics, University of California, 2127 North Hall, Santa Barbara, CA 93106, IZA, and NBER.

[4] Environmental Science and Policy, University of California, Davis, USA

October 2021

[*] We thank Safegraph, Cuebiq, PlaceIQ and Google Mobility for generous data access, as well as Jimena Rico-Straffon for outstanding research assistance. We also thank Marianne Bitler, Clément de Chaisemartin, Karen Ortiz Becerra, Dalia Ghanem, Jess Rudder and Ashish Shenoy for thoughtful comments, and participants at the Berkeley Initiative for Transparency in the Social Sciences, UC Environmental Economics webinar, UC Davis Econometrics reading group and NatuRE Policy Lab. This research was supported by the University of California, Grant R00RG2419.

Abstract

The COVID-19 pandemic brought unprecedented policy responses and a large literature evaluating their impacts. This paper re-examines this literature and investigates the role of researchers' degrees-of-flexibility on the estimated effects of mobility-reducing policies on social-distancing behavior. We find that two-way fixed effects estimates are not robust to minor changes in usually-unexplored dimensions of the degree-of-flexibility space. While standard robustness tests based on the sequential addition of covariates are very stable, small changes in the outcome variable and its transformation lead to large and sometimes contradictory changes in the estimates, where the same policy can be found to significantly increase or decrease mobility. Yet, due to the large number of degrees-of-flexibility, one can focus on a set of results that appears stable, while ignoring problematic ones. We show that recently developed heterogeneity-robust difference-in-differences estimators only partially mitigate these issues, and discuss how a strategy of identifying the point at which a sequence of ever more-stringent robustness tests eventually fail could increase the credibility of policy evaluations.

1 Introduction

The COVID-19 crisis has brought unprecedented responses by national and sub-national authorities. At the center of this public action has been mobility restrictions and lockdown orders aimed at reducing the risk of virus transmission through constrained social contacts. More than a year after their initial implementations, and as subsequent waves of COVID-19 cases surged in many countries around the world, such measures continue to be reintroduced and their efficacy debated. Therefore a key question is whether, and to what degree, these mobility-reducing policies cause reductions in measured mobility.

Due to the urgency of this question, and the near-real time availability of data, the scientific literature focusing on the effect of what we refer to as 'mobility-restricting policies' (MRPs)[1] on mobility grew at near exponential rate: as of March 2021, a Google Scholar search of "mobility" and terms related to MRPs implemented in response to the COVID-19 pandemic ("shelter-in-place", "stay-at-home", or "lockdown") returned over 23,000 results. A large number of these studies focus on estimating the impacts of mobility-restricting policies on measured mobility with the level of analysis varying from city to worldwide. In particular, due to the sharp temporal adoption of MRPs, many of these studies leverage a difference-in-differences (DD) based design. This previous literature does not reach a consensus; results range from finding that "there is little evidence... that stay-at-home mandates induced distancing" (Gupta et al., 2020) to "clear effects of stay-at-home orders on social distancing" (Allcott et al., 2020). At the same time, several studies highlight existence of pre-trends, often interpreted as indicating that voluntary mobility reductions were an important factor (e.g. Gupta et al. 2020; Andersen 2020; Abouk and Heydari 2021).

In this paper, we re-examine the evidence from this early literature and assess the impact of mobility restrictions on county-level mobility in the United States, leveraging MRPs uniquely gathered across state, county, and city levels. First, we document and correct important gaps in the source data on mobility restriction policies used in the previous literature. For example, one

[1] These policies are also described as "non-pharmaceutical interventions (NPIs)" in the literature

commonly used data set indicated that around one third of U.S. counties enacted an emergency declaration by the end of March 2020 (NACo, 2020); our data gathering showed that the true number was twice as large. Using the corrected data, we consider a broad set of five MRPs: State emergency declarations, State shelter-in-place orders, State earliest restriction or closure, County emergency declarations, and County shelter-in-place orders. We supplement these MRPs data with an extensive set of 20 daily mobility indicators derived from anonymized mobile device ping information provided by Safegraph, Google Mobility, PlaceIQ, and Cuebiq.

Second, we estimate the impact of MRPs on observed mobility using standard two-way fixed effects and event-study methods, following the approach used in the previous literature. We examine a wide range of specifications that were used in the previous literature (although in distinct individual papers), focusing on linear models, log-linear models, seven-day moving averages, and first-differences. We further supplement these analyses by considering all mobility outcomes observed at the county-day level, and also estimate the impact of MRPs individually and in combination. The later is especially important since MRPs are almost always overlapping (e.g., counties with a county shelter-in-place order may also have a county emergency declaration or a state shelter-in-place order) and temporally dependent (e.g., emergency declarations typically precede shelter-in-place orders).

Most importantly, the wide range of specifications we consider allows us to assess whether the estimates of the impact of MRPs on mobility in the prior literature paint a representative picture of the evidence. In particular, we emphasize the wide range of researchers' degrees of flexibility (or freedom) that exist over many modelling and specification choices: the outcome variable (mobility measure) and its functional form in the regression model, covariates included, estimator, and the spatial and temporal unit of analysis (e.g., state or county, daily or weekly). Rather than focusing on a "clean" and robust set of results in support of a consistent narrative, we report estimates along many dimensions of the degree-of-flexibility space, while highlighting contradictory findings, failing robustness tests, and ambiguous results.

At the county scale, we find that some conclusions drawn about the efficacy of mobility restricting policies in the previous literature using two-way fixed effect regressions are not robust to seemingly minor variations in the estimating model, casting doubts on the internal validity of the prior estimates. Depending on the mobility outcome and policy considered, several estimates have the "wrong" sign, where the MRP is found to increase mobility. For a given mobility measure, several estimated effects switch sign, but remain statistically significant, when using the log of the outcome variable instead of the level. Most alarmingly, the standard robustness tests based on the sequential addition of covariates fail to provide warning signs of specification sensitivity. The fact that robustness on covariates does not imply robustness on transformations of the outcomes illustrates that focusing on "one-dimensional" robustness checks can be misleading.

Third, we find that event-study estimates in this context are plagued with pre-trends, which have been interpreted by previous studies as evidence of voluntary social distancing in anticipation of the policy announcement. These issues are most pronounced using the standard event-study estimators. We apply recent heterogeneity-robust DD estimators and find that they partially mitigate these issues, but only for certain outcome variables and functional forms. Overall, we document that at the county-day level, the existing tools for difference-in-differences analysis —both standard and recently developed methods—are insufficient for making rigorous and specific conclusions regarding the causal impact of MRPs on mobility. To be clear, we do not assert that MRPs were ineffective; strong patterns in the data are highly suggestive of a substantial impact on mobility. Rather, our conclusion is that DD-based evaluation methods are not yet a sufficient match for the complexity presented by multiple overlapping policies and feedback processes to the behavioral outcome of interest at a fine-scaled spatio-temporal level (i.e., county-by-day). The ultimate concern of broader policy interest is not just whether MRPs reduced mobility but also whether actual infections and deaths were curtailed. However, this latter step requires rigorous and robust treatment of the first step, which we show is much less straightforward than initially expected.

Besides our direct contribution to the literature on the efficacy of mobility-restricting policies, our paper brings to light issues related to the meta-scientific literature on researchers' degrees of flexibility and the credibility of economics research (see Christensen and Miguel (2018) and Kasy

(2021) for recent overviews). In any empirical study, researchers are confronted with a multitude of decisions on how to handle the data and conduct the analyses, a *garden of forking paths* in Gelman and Loken (2013)'s terms. The resulting findings are conditional on the specific path of decisions made along the way, leaving open the question of whether different paths would have lead to very different conclusions.

The sensitivity of empirical results to researcher degrees of flexibility have been highlighted in several replication studies (Aiken et al., 2015; Foote and Goetz, 2008; Clemens and Hunt, 2019), as well as in more exhaustive "many-analysts" experiments, where multiple research teams are asked to analyse the same data set (see e.g., Silberzahn and many authors (2018); Botvinik-Nezer and many authors (2020); Huntington-Klein et al. (2021)). These studies are based on artificial research exercises, where each researcher's goal is not to write a full-length article but instead to produce an isolated analysis and set of findings. One might hope that in the process of actual research, major instability of the results over alternative researcher's decisions would be discovered and highlighted, either in different papers or within the "robustness" section of a given paper. The recent COVID-19 literature provides a natural experiment to assess whether this is actually the case: Due to the importance and urgency of the COVID crisis, a remarkable number of studies focusing on the same research question were released and/or published in a very short period of time. This represents a unique opportunity to investigate in a real setting both the influence of researchers' degrees of flexibility on the study's conclusions; and whether the sensitivity of the results is reflected in each paper or across the combined literature.

By estimating MRP impacts on mobility along all main dimensions of the researchers degrees of flexibility in our setting, we can point to several concerns that have not been fully addressed in the previous literature. We find that apparently innocuous researcher's decisions such as modelling the dependent variable with the log transformation or leaving it in levels can have a large impact on the results, even impacting the signs and statistical significance of the estimated effects. Further, we find that for almost every *well behaved* result obtained, we can find an alternative and equally reasonable specification which gives opposite sign or null results. These issues are mitigated but not fully resolved by the use of recent heterogeneity-robust DD estimators (e.g., Callaway and Sant'Anna 2020). Finally we find pervasive pre-trends in the data for virtually all MRP and mobility outcomes, which can only be satisfactorily addressed by adopting a particular specification.

We do not take a stance as to why the previous literature largely focuses on highlighting the 'clear' impacts of MRPs on mobility. This could be due to random factors, the supply side of research production (e.g., selective reporting of results in agreement with the researchers' priors, selective submission of results that provide a consistent story, biased sampling of the specifications space when exploring possible models), or the demand side (anticipated need by researchers to provide 'clean' results and story as a necessary step for publication). Our own previously published research in this area highlighted significant impacts of emergency declarations (one prominent example of MRP) on mobility outcomes, though also documented large pre-trends in event-studies (Weill et al., 2020). Furthermore, the econometric research focused on issues with two-way fixed effects specifications developed extremely rapidly, and in parallel with the applied COVID-19 policy research. Many of the shortcomings linked with the use of two-way fixed effects estimators under staggered adoption designs have only recently became apparent.

Concerns about selective reporting of results, p-hacking, and pre-testing are not new. Leamer (1978) recommended more than 40 years ago that researchers should report the results of all estimations that they tried, as opposed to focusing on a few chosen ones. Similarly, in the context of event-studies, where the causal interpretation of a policy effect relies on the absence of pre-trends, recent research shows that focusing on pre-trend tests that "work" can induce substantial bias in the reported estimates (Roth, 2019). Unfortunately, the recommendation made by Leamer is far from being a standard practice today. When featuring alternative specifications to the main model, published studies tend to report "robustness tests" that are based on the inclusion of alternative covariates in the regression model. These reported robustness tests also generally agree with the main findings of the research. As we show in this paper, failure to consider robustness analysis along all relevant dimensions (outcomes under study, functional form, covariates, and estimator) can lead one to wrongly infer that the estimated effect is stable.

The remainder of this paper proceeds as follow. First, we summarize a selected set of existing empirical studies of the impact of MRPs on mobility (directly below). We then present our analysis, beginning with the data (Section 2) and results of the standard two-way fixed effects models (Section 3). Next, we explore robustness tests (Section 4) and contradictory results from event studies (Section 5). Finally, we round out the analysis with results from heterogeneity-robust estimators (Section 6).

1.1 Review of existing empirical evidence

Table 1 summarizes notable studies that address similar questions using difference-in-differences designs include a set of working papers (Gupta et al., 2020; Andersen, 2020; Elenev et al., 2021) and peer-reviewed articles (Painter and Qiu, 2021; Villas-Boas et al., 2020; Allcott et al., 2020; Abouk and Heydari, 2021; Dave et al., 2021; Goolsbee and Syverson, 2021). While the shared methodology might suggest a relatively uniform approach, examination of the literature indicates the absence of a consensus on many empirical design choices. For example, each analysis uses a different transformation of the key outcome/dependent variable (mobility metric), including levels (Gupta et al.); natural log (Allcott et al.; Goolsbee and Syverson; Elenev et al.); the change relative to a fixed pre-COVID baseline in level or log (Villas-Boas et al.; Dave et al.); the daily change relative to a daily 2019 counterfactual (Andersen); and daily first differences (Painter and Qiu; Abouk and Heydari). Four of these articles use weighted regression population weights, one used weights based on the number of visits to stores in January (Goolsbee and Syverson), another via a synthetic control method (Villas-Boas et al.), while the rest do not. Mobility data are typically drawn from a single source, with Safegraph being the most common. Most studies focus only on mobility response measured at the state level. Exceptions include Allcott et al. who focus on the county level but combines county and state policies into one metric. Gupta et al. also focuses on the county level and considers state and county policies, albeit not with full event study regressions where the different policy effects are estimated simultaneously. These analyses typically rely on existing databases of reported policies (e.g., such as those collected by the Kaiser

Family Foundation or the New York Times) and none (to our knowledge) gather substantial new data on which counties or cities enacted local policies.

Table 1: Selected Studies Analyzing the Impact of MRPs on Mobility Outcomes

Article	Dependent variable transformation	Population weighted	Mobility data source	MRPs	Units
Gupta et al., 2020	no	no	PlaceIQ, Safegaph, Google, Apple	county and state	county and state
Painter and Qiu, 2020	first-difference	no	Safegraph	state	county
Andersen, 2020	deviation from 2019 counterfactual; log for visits	yes	Safegraph	state	state (event studies)
Villas-Boas et al., 2020	%-deviation from pre-COVID-19 period (2/10-3/8)	no (weight with SCM)	Unacast, Google	state	state
Allcott et al., 2020	log	yes	Safegraph	county	county
Abouk and Heydari, 2021	first-difference	yes (TWFE)	Google	state	state
Dave et al., 2021	deviation from pre-COVID-19 period (2/6-2/12)	yes	Safegraph	state	state
Goolsbee and Syverson, 2021	log	yes (various)	Safegraph	county and state	county
Elenev et al., 2021	raw shares; log for visits	yes and no	Sagegraph	county and state merged	county

Notes: SCM = synthetic control method, TWFE = two-way fixed effects.

Issues with pre-trends are apparent and discussed in several of these articles. Many observe that variation in mobility is not fully or even largely explained by MRPs. For example, Allcott et al. find that “the magnitude of policy effects is modest, and most social distancing is driven by voluntary responses”. Similarly Abouk and Heydari conclude that “(a)lthough evidence for reduced social contact in the United States is strong... people in most states had already started to reduce the time they spent outside their homes before any NPI (non-pharmaceutical intervention) was implemented”, and Goolsbee and Syverson concludes that “[...] the vast majority of this drop is due to individuals’ voluntary decisions to disengage from commerce rather than government-imposed restrictions on activity.” . Only one article focuses on the prospect of time varying treatment effects in its main results: Andersen observes that “it is possible that difference-in-differences and event study estimates are biased by comparisons between early and late treated units”. Within this literature we could find no discussion focused on estimates that have an unexpected sign or that switch sign across plausible alternative specifications, as we do below. Taking the previous literature as a starting point, we have designed our paper to fully explore the implications of empirical design choices (data sources, estimator, functional form choice, policy under consideration, etc) on the estimated impact of mobility-restricting policies on mobility. Finally, one article explicitly investigates the role of spillover effects between nearby counties within the framework of two-way fixed effects models (Elenev et al.).

2 Data Sources and Preliminary Analysis

Our empirical analysis of the impact of mobility restricting policies on mobility outcomes is conducted with a comprehensive set of data files on mobility outcomes collected from mobile device signals, combined with extensive MRP data assembled at the city, county and state level. This section describes the data sources, defines the primary outcome and control variables, and then presents some summary statistics.

2.1 Mobility Measures

We assembled a daily data set of 20 mobility measures at the county level collected during the first-wave of the COVID-19 pandemic in the United States. We restrict our analysis to start in January 2020 and to end on April 21st, 2020. This corresponds to the time span between the earliest date the mobility data are available, and the date where some jurisdictions began to lift distancing orders on April 21st, 2020. Our mobility measures are all based on anonymized and aggregated mobile device signal ("pings") data and come from four different sources. The variables are listed in Table 2 along with their source and summary statistics. Additional information on each of the variables appears in Table A.1 in the Appendix. Safegraph data are provided at the census-block group level, which we aggregate to the county level. PlaceIQ data were used by Couture et al. (2020) to derive the *Device Exposure* variable at the county-level. Google Mobility data are provided at the county-level over the period February 15 - April 21 (expressed in changes relative to the 5-week period from January 3 to February 6, 2020). Cuebiq data are provided at the county level.

The summary statistics in Table 2 include the number of county-day observations (N) for which each of the mobility measures are available between January and April 2020. Differences in the temporal and spatial coverage of these measures are evident, with Safegraph and Cuebiq providing the best spatial coverage with almost all counties represented (over 285,000 observations, amounting to 3069 counties and 93 days). Google mobility data are not consistently available for all days and counties due to Google's anonymity constraints. Many of these variables are designed to capture the amount of time spent at or away from home by individuals,[2] other variables correspond to specific activities outside the home,[3] while some other variables can be interpreted as mobility and social mixing indices.[4]

In the fourth column we report the sample mean of the various mobility measures. For example, the Safegraph variable *Completely Home* indicates that 27% of sampled devices spend the day completely at home, while the *Median Distance* travelled outside the home is 10,683 meters on average between January 20 to April 21. Similar to the information from Safegraph, Cuebiq data indicates that 28% of device users are *Staying Around Home*. The Google Mobility data shows both positive and negative averages, as each of its mobility indicators are normalized relative to same day of the week between January 3 to February 6, 2020, before the effects of the COVID-

[2] *Staying Around Home (%) Average Time Not Home*, *Median Not Home Dwell Time*, *Completely Home (%)*, *Median Home Dwell Time*

[3] *Grocery and Pharmacy Retail and Recreation Workplaces*, *Full Time Work (%)*

[4] *Device Exposure*, *Median Distance*, *Mobility Index*

Table 2: Summary Statistics and Sources for Mobility Outcomes

Source	Variable	Re-signed	N	Counties	Mean
Safegraph (SG)	Average Time Not Home	yes	285,405	3069	284.55
	Away at least 3 hours (share)	yes	285,405	3069	0.15
	Completely Home (share)	no	285,405	3069	0.27
	Full Time Work (share)	yes	285,405	3069	0.06
	Median Distance	yes	285,405	3069	14104.85
	Median Home Dwell Time	no	285,405	3069	616.82
	Median Home Share (share)	no	285,405	3069	0.74
	Median Not Home Dwell Time	yes	285,405	3069	139.43
	Part Time Work (share)	yes	285,405	3069	0.09
PlaceIQ (PQ)	Device Exposure	yes	183,117	1969	89.82
Google (GM)	Grocery & pharmacy	yes	134,843	2425	1.19
	Parks	yes	42,445	944	8.84
	Residential	no	77,185	1526	7.95
	Retail & Recreation	yes	139,935	2517	-12.21
	Transit stations	yes	64,358	1105	-13.11
	Workplaces	yes	164,680	2722	-18.87
Cuebiq (CQ)	Mobility Index	yes	285,413	3069	3.65
	Staying Around Home (share)	no	285,413	3069	0.28
	Staying Around Neighborhood	no	285,413	3069	0.37
	Traveling more than 10 miles	yes	285,413	3069	0.34

Notes: Table 2 reports summary statistics for the 20 mobility indicator variables analyzed in this paper. The number of observations corresponds to the number of counties in which a mobility indicator is observed multiplied the number of days it is observed. The “re-signed” column indicates if a mobility indicator will be re-signed in the analysis below so that increase in the variable can be interpreted as increase in social distancing. In the tables and figures below re-signed variables will be identified with an asterisk next to their name.

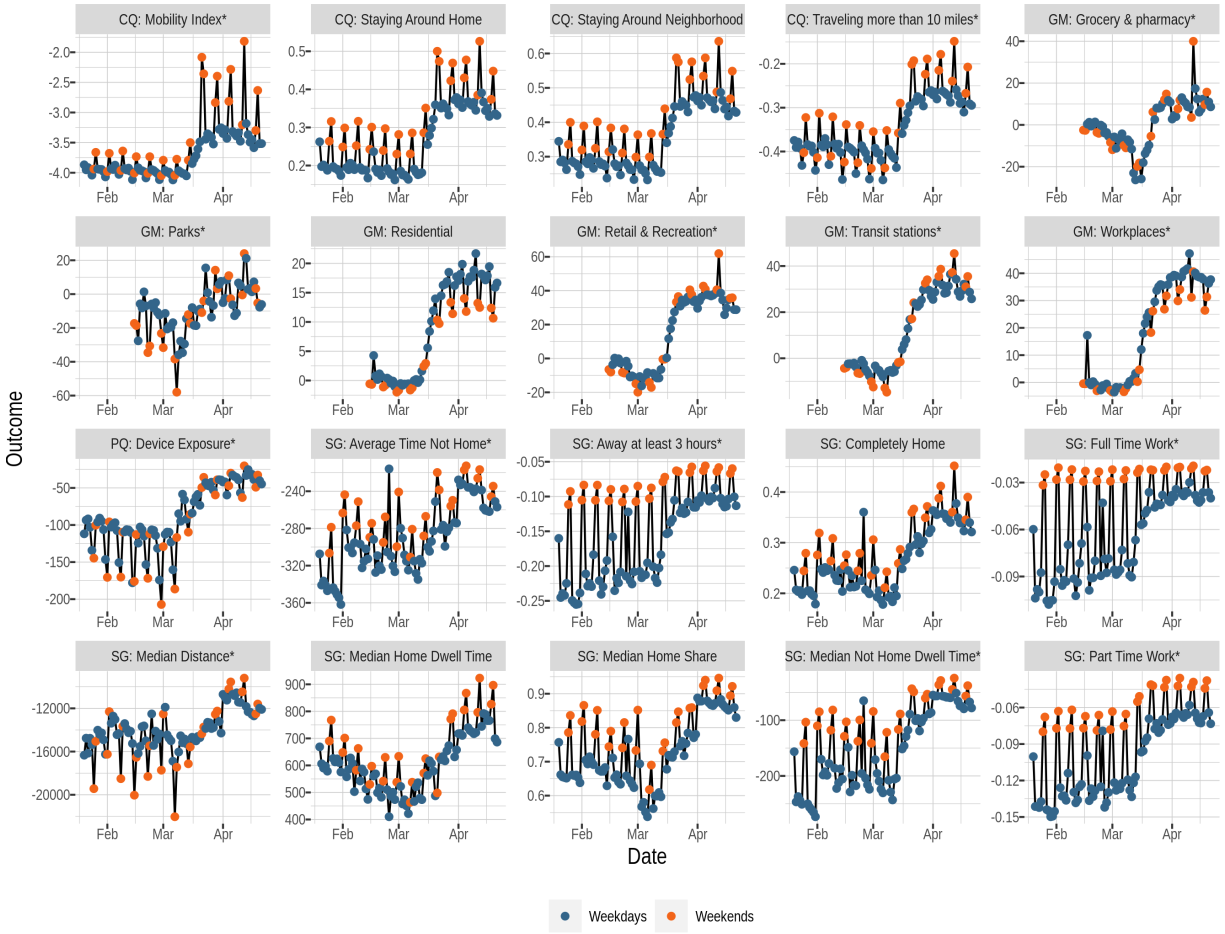
19 pandemic were fully realized in the U.S. Interestingly, within these relative Google measures we see that *Parks* and *Residential* time have positive averages, while *Retail & Recreation*, *Transit stations*, and *Workplaces* all have negative averages, consistent with changes in mobility behavior to increase social distancing.

Figure 1 plots the mobility and activity patterns in the anonymized “pings” data, with weekdays shown by red dots and weekend days shown with blue dots. **To ease comparisons between outcome variables, for the remainder of the paper we analyze “re-signed” mobility outcomes so that increase in the variable can be interpreted as increase in social distancing**. For example, *Traveling more than 10 miles* was multiplied by minus one: on the first row and fourth column of Figure 1, the variable indicates an initial average level of about -0.4 in January (meaning that on average at the county level, 40% of people moved further than 10 miles per day), and then increases quickly to -0.3 in April. When working with the log of the outcomes, we first take the log and then multiply by minus one. To keep track of this convention, variables that were re-signed are highlighted with an asterisk next to their name. Several notable features emerge from the simple analysis in Figure 1: First and foremost, there is an important reduction in mobility and social activity that occurs around the middle of March 2020, when the majority of the U.S. population was under some form of MRP. This reduction in mobility is evident for all of the outcome variables. For example, the Cuebiq *Staying Around Home* nearly doubles, and the Place IQ *Device Exposure* is halved. Second, the implied mobility by these indicators appears to be reduced more strongly during weekdays, consistent with the large increase in unemployment claims and in work from home observed at that time. Finally, for most of the mobility indicators, there is no reversion to pre-mid-March mobility and activity levels as of the of April 2020.

In addition to Figure 1, Figure A.16 in the appendix presents Pearson correlation coefficients between these different outcome variables used to measure mobility. Except for Google Mobility *Parks* and Safegraph *Median Distance*, all outcome variables are strongly correlated, with correlation coefficients well above 0.5. However, the analysis in Sections 4 and 5 will show that the choice of mobility outcome variable has an important impact on the estimated effect of MRPs on mobility.

Table 3 report the sample averages of the control variables used in the empirical analysis below. In our sample of 3069 counties, there are on average 2.14 confirmed cases of COVID-19 per day, and 27 % of counties-days observations have at least one confirmed case of COVID-19 (8% experienced at least one death attributed to COVID-19). Naturally, all these indicators of pandemic severity have means close to zero before the Mar 11, and much larger means afterwards. The remaining rows of Table 3 report the average daily temperature (in °F), precipitation and snow (both in hundreds of inches).

Figure 1: Trends in Daily Mobility Outcomes, January-April 2020



Notes: Some mobility outcomes are resigned (indicated by a star next to the variable name) so that increases in a given outcome are interpreted as increases in social distancing. All mobility outcomes are in levels, except for Google Mobility which is only available in % change relative to a baseline.

2.2 Mobility Reducing Policies

We obtained data on the date of declaration of COVID-19 mobility reducing policies (MRPs) in the U.S. from February-April at the state, county and city level. For all three levels, our policies of interest include emergency declarations (EDs) and shelter-in-place orders (SIPOs). For state-level policies we used data provided by Fullman et al. (2020). For counties we used data collated by the National Association of Counties (NACo, 2020) as a starting point. The NACo data constitute an incomplete set of county-level policies implemented. No pre-existing data set was available for cities. To fill in these gaps in county and city MRPs, we conducted a manual search for the dates

Table 3: Summary Statistics for Main Control Variables

Variable	N	Counties	Mean	Min	Max
Cases	285417	3069	2.15	0.00	2174.00
Cases > 0	285417	3069	0.27	0.00	1.00
Cases sqrt	285417	3069	0.34	0.00	46.63
Deaths > 0	285417	3069	0.08	0.00	1.00
Precipitation	282997	3043	0.11	0.00	5.16
Snow	282997	3043	0.09	0.00	16.59
Temperature (mean)	282997	3043	43.80	-18.39	87.75

of either type of declaration for (1) all U.S. counties lacking either a ED or SIPO in the NACo data, as well as (2) cities with a population of 50,000 or greater.[5] When existent, these dates were typically found in media reports, government websites or government meeting minutes. Overall, across 3,069 U.S. counties, the original NACo data set included 989 EDs and 147 SIPOs; these counts were slightly more than doubled (through manual search) in our final data set, to 1954 EDs and 334 SIPOs. Across the 783 U.S. cities with a population of 50,000 or greater, we identified 633 with EDs and 122 with SIPOs.

Governmental responses were not limited to EDs and SIPOs. For example, some states and counties issued separate orders for restrictions on restaurants, gatherings, schools and non-essential businesses. For consistency and tractability, given our coverage of three levels of government, we focus mainly on EDs and SIPOs. However, at the state level we also consider the "earliest restriction or closure" (ERC) as provided by Fullman et al. (2020), which takes the earliest state policy that either closed schools, restaurants, bars or restricted gatherings.

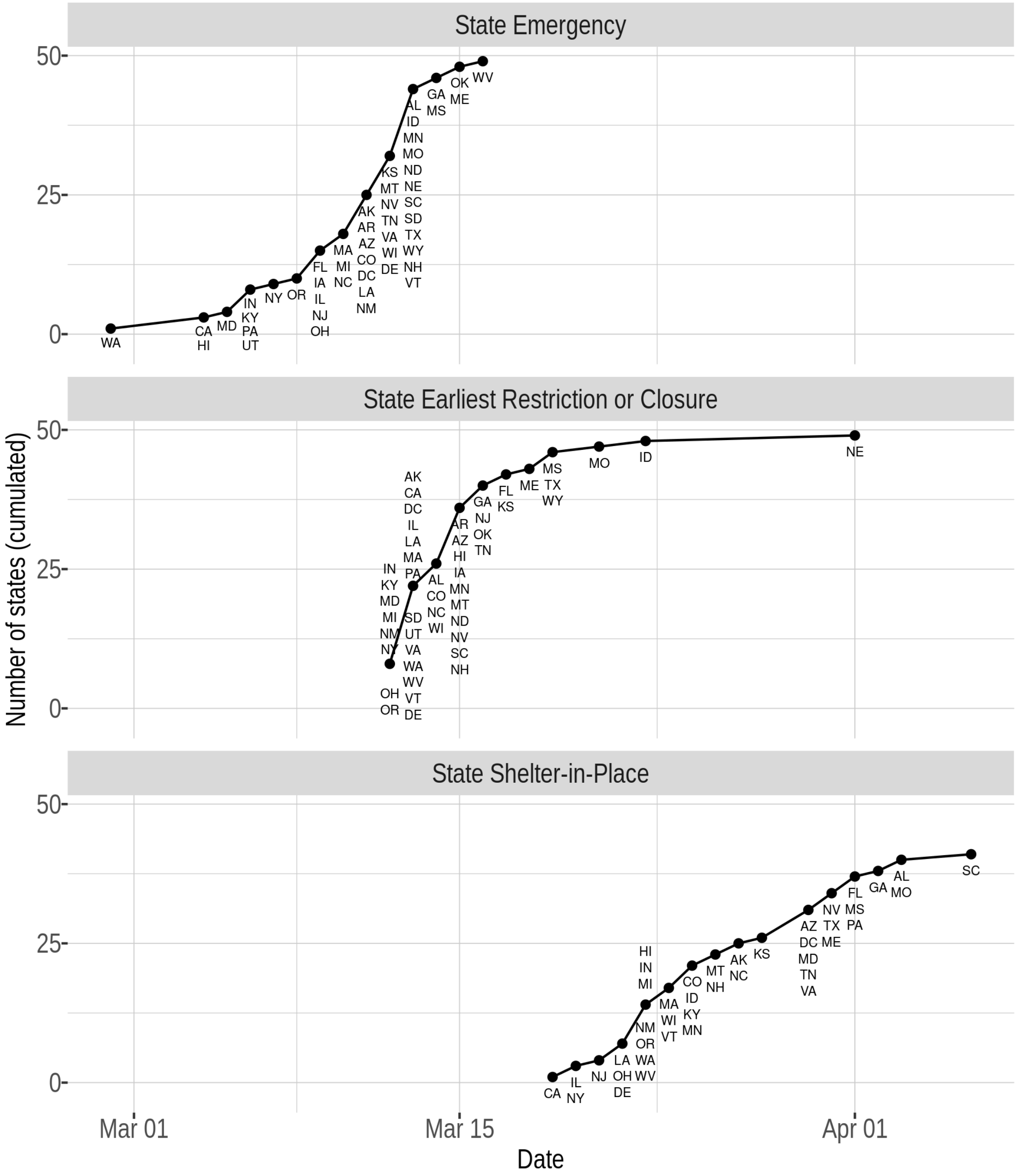
Figure 2 shows the cumulative adoption of state-level policies as a function of date, by policy and state (as indicated by state abbreviation). For example, Washington state was the first to declare an ED, followed by California and Hawaii. It is also evident that there is an ordering in MRP policy adoption, with the ordering starting with EDs, followed by ERCs, and then SIPOs. Notably, all states declared an ED and an ERC at some point before April 1st, but not all states imposed a

[5] We used population in 2019 as estimated by the U.S. Census Bureau (2020). Where county and city governments have been consolidated into one jurisdiction, we treat the entity as a county in our data set.

SIPO.

Figures 3 and 4 show the timing of county-level EDs and SIPOs adoptions, respectively. The main finding here is that county-level EDs are much more common than SIPOs. We also observe that EDs are less common in the less-populated areas of the central U.S., as well as the far northeastern U.S. This is not surprising as these regions were not heavily impacted by COVID-19 before April 2020.

Figure 2: Cumulative Number of States Adopting a Given Mobility Restricting Policy by Date



Notes: Cumulative number of state-level MRP adoption by date. New adoptions on a given date are represented with state abbreviations.

Figure 3: Timing of County Emergency Declarations in the U.S. from February 14 to June 11, 2020.

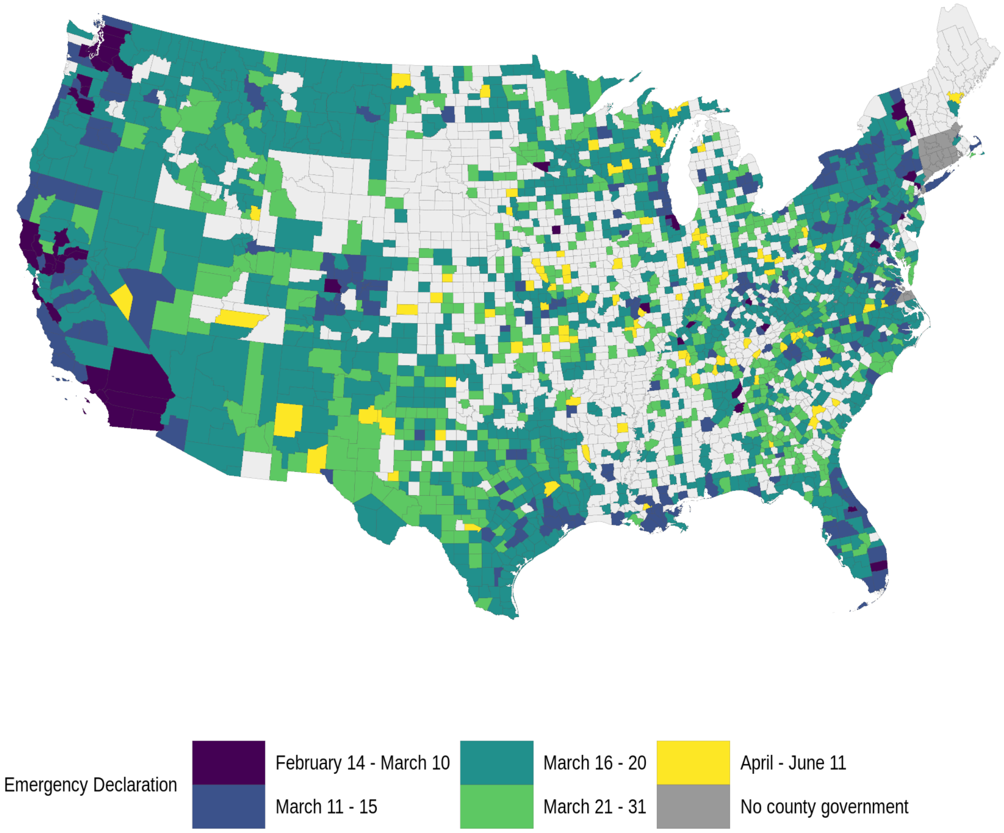


Figure 4: Timing of County Shelter-In-Place Orders in the U.S. from February 14 to June 1, 2020

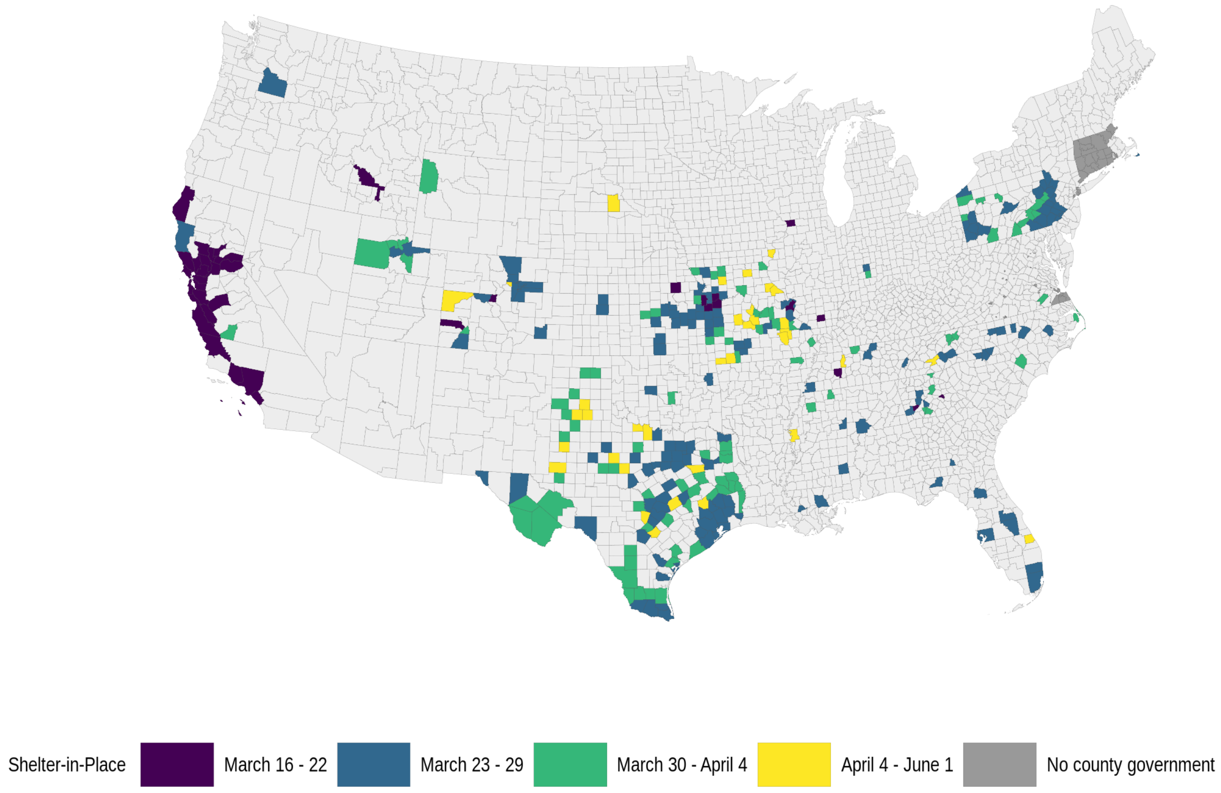
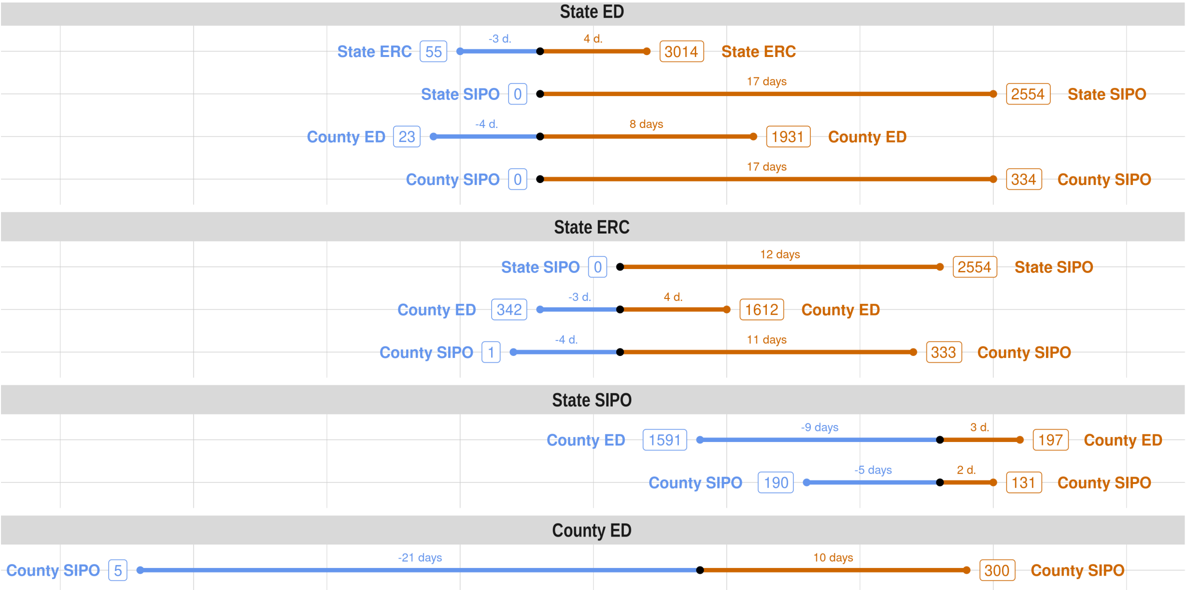


Figure 5 displays the relative timing of the adoption of state and county policies. For each of the MRPs we consider, we compare pairs of policies (say “A” and “B”) and reports the number of counties where policy A was adopted before policy B (blue line) and the number of counties where policy A was adopted after policy B (brown line). In addition, we report the average relative time (in days) between the adoption of policies A and B. For example, the top panel for state ED shows that 3014 counties (out of 3069 in the full sample) were in a state that adopted an ERC later than ED (with an average gap of 4 days), while 55 counties saw the opposite ordering. Looking at the comparison between state ED and the four other MRPs, it is evident that state EDs were generally adopted first, typically with long gaps before the adoption of a follow-up policy (e.g., 17 days for state SIPO and county SIPO). A similar pattern is observed comparing state ERC (typically the second MRP policy adopted after state ED) to the other MRPs. For example, all state SIPOs were adopted after the state ERC (12 days later on average), and all but one of the county SIPOs were put in place after the state ERC. Finally, county ED measures tend to be introduced before state SIPOs, while the relative timing of state and county SIPO adoptions tends to be balanced. Like the pattern for state-level policies, county SIPOs are introduced later that county ED measures, on average.

Figure 5: Pairwise Comparisons of the Relative Timing of Mobility Restricting Policy Adoption



Notes: For each of the four MRPs listed in bold, the figure shows the average relative adoption time (reported in days) and the number of counties (reported in the boxes). Adoptions that come before are shown in blue and those that come after in brown. For example, in the top panel for state ED, we report that of all counties that ever adopted county ED, 1931 counties adopted state ED before county ED (with an average gap of 8 days), while 23 counties adopted county ED before state ED (with an average gap of 4 days).

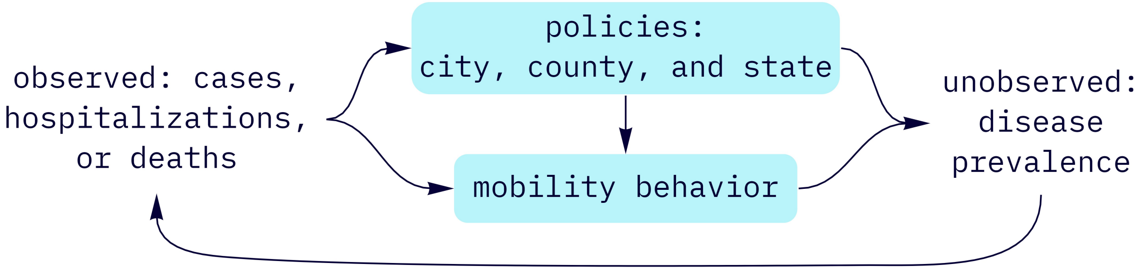
There are three main conclusions from this section that have important implications for the empirical analysis below. First, considering state-level policies, all states implemented some form of MRP (although not all implemented the more restrictive SIPO), while at the county-level, 62 % of counties (representing 79 % of the U.S. population) implemented one or more form of MRP. Second, a substantial share of the U.S. population was subjected to multiple MRPs simultaneously (e.g., state ED, county ED, and county SIPO). Third, states and counties adopted MRPs in a staggered manner, with early adopters implementing the policies several weeks before later adopters. Further, there is typically a natural ordering of policy adoption, for example, state/county SIPOs always follows state/county EDs. These features of the policy response to the first wave of the

COVID-19 pandemic complicate the estimation of the causal effect of various MRPs on mobility. Different jurisdictions often adopted a wide range of different policies, but not all of them: as a result, one might wish to estimate which of these MRPs work best, which motivates including them simultaneously in the analysis.

3 Policy impact estimation framework

Figure 6 represents our conceptual causal chain. We focus here on the link between MRPs and mobility behavior, and highlight two challenges that impede straightforward estimation of this relationship. First, rather than a single event involving one policy at one spatial scale, the early response to the COVID-19 pandemic included multiple policies at multiple scales. Furthermore, the vast majority of these policies appeared in a compressed window of two months, limiting the variation in "treatment timing" to inform the impact. Second, the conceptual causal chain we posit in Figure 6 raises the prospect of endogeneity. Our focus is estimating the impact of MRPs on mobility behavior. These policies by cities, states and counties are clearly a response to perceived risk as informed by COVID-19 cases, hospitalizations and deaths. However, we might also expect individuals to respond directly to these perceived risks, in addition to policies. Finally, these policies and changes in mobility behavior are expected to influence disease prevalence and hence future perceived risk as further disease outcomes are observed. In future work, it may be that a structural approach to modeling this system would provide further traction in disentangling these effects.

Figure 6: Conceptual causal chain. This article focuses on the relationship between policies and mobility behavior (shaded).


observed: cases, hospitalizations, or deaths
policies: city, county, and state
mobility behavior
unobserved: disease prevalence

3.1 Empirical model specification

As a starting point for our model specification, we follow the previous literature and aim to estimate the impact of mobility reduction policies (MRPs), enacted by different jurisdictions (states and counties) at different points in time, using a linear two-way fixed effect estimator (TWFE):

$$Y_{it} = \alpha_i + \delta_t + \sum_{p=1}^{P} \beta_p D_{ipt} + X_{it}' \gamma + \varepsilon_{it} \tag{1}$$

where i denotes county, t represents the day (from January 20 - April 21, 2020), and p is the index for the five policies we analyze. Y_{it} is one of the 20 different mobility and activity outcome variables we described in Tables 2 and Appendix Table A.1. D_{ipt} is a dummy variable equal to 1 when county i is "treated" by a given MRP policy p within the set P: State ED (Emergency Declaration), State ERC (Earliest Restriction and Closure), State SIPO (Shelter in Place Order), County ED, and County SIPO on date t. To begin we follow the previous literature's practice of analyzing the impact of MRPs individually (i.e., by including them one at the time in the model), but we will also report estimates of β_p when all the impact of all policies are estimated simultaneously.

The vector of control variables X_{it} are observable predictors of mobility and activity, which include daily precipitation, snowfall and mean temperature. In some specifications we also include binary indicators for whether the first COVID-19 case and, separately, the first COVID-19 mortality has been recorded in a county.[6] Including these variables in the models have important implications in terms of the required identifying assumptions. On the one hand, such controls may help alleviate the omitted variable concerns described in the conceptual framework. On the other hand, these controls may violate the strict exogeneity assumption, since mobility in previous periods may have causal effect on future COVID-19 cases and deaths.

The fixed effect α_i controls for time-invariant characteristics of each county, including typical mobility patterns, rural or urban status, population density, the availability of transportation infras-

[6] Some counties never experience COVID-19 mortality events in the sample period, in which case the binary indicator remains a zero throughout.

tructure, and all other county-specific drivers of mobility that do not change over time.[7] The date fixed effect δ_t captures time-varying drivers of mobility and activity, including any national-level event or announcement that influences mobility equality across locations (e.g., announcements from the CDC), pronounced weekday/weekend differences documented in Figure 1, and any nationwide weather-driven trend.

Like the previous literature reviewed earlier, the goal of the analysis is to identify the causal effect of MRPs on mobility and activity measures, as represented by the vector of parameters β. Identification of β requires a strict exogeneity assumption for the error term in Equation (1), requiring that it is uncorrelated with the various policy indicators, conditional on the controls and fixed effects included in the regression. This can be interpreted as the standard no pre-trends assumption whereby the mobility outcomes in "control" counties provide a valid counterfactual for the mobility outcomes in "treatment" counties that adopted MRPs. This assumption can be tested by estimating the event-study analog of Equation (1) and testing for pre-trend differences directly. Furthermore, when policies are included one at a time, the *interpretation* of β as the average treatment effect on the treated relies on assuming that treatment effect is constant across units and/or over time. We return to these important points below.

4 Results using the TWFE estimator

For the remainder of this paper, we will systematically investigate how researcher's decisions influence the estimated effects of MRPs on mobility. The space of researcher's degrees of flexibility in this context is extremely large. We only focus on a subset of this space, which still includes multiple dimensions: outcome variables, transformations of the outcome variables, combinations of MRPs to consider, included covariates, regression weighting, and estimator used. Due to these many dimensions we cannot present all the results at once. In the following sections, we will present successive "slices" of the multidimensional space of researchers degrees of flexibility, thus

[7] In practice, many of those characteristics may change slowly over time (e.g., population density). For the purpose of this analysis with a sample period that covers a three month period, we consider them time invariant.

highlighting the numerous “forking paths” of the garden as in (Gelman and Loken, 2013).

Table 4 begins the empirical analysis by reporting the TWFE estimates of β_p for a log-linear specification with the dependent variable *Completely Home (%)*. The log-linear specification is one of many used in the previous literature, and we also consider alternative specifications further below.

There are 8 columns in Table 4, each corresponding to a different combination of the MRPs and control variables. In column (1) we consider only state-level MRPs and we find that all three state-level policies (ED, SIPO, and ERC), considered in isolation of the city or county policies, lead to statistically significant increases in the fraction of time spent completely at home. For example, the state SIPO policies are associated with a 4.3% (0.043 log points) increase in the fraction of time spent at home. In column (2) we consider only county-level MRPs and find stronger effects, e.g., a 9.0% increase in the fraction of time spent at home due to a county SIPO. Column (3) considers city-level MRPs. Since the unit of observation for the mobility data is the county, we model the effect of county population share residing in a city with an MRP. The estimated coefficient is positive and statistically significant, indicating that a 10 percentage point increase in a county’s population living in a city with mobility reduction policy increases the fraction of time spent at home by 2.3%.

Since the MRPs are often overlapping (e.g., counties with a county SIPO may also have a state ED or SIPO), we then estimate a model where all MRPs are included in the same regression. Column (4) reports the results of this analysis, and shows that estimating the impact of MRPs on mobility separately for state, county, or city-level policies or estimating the effects jointly leads to very similar estimates. Among the state- and county-level binary policies, SIPO policies are the strongest determinants of reduced mobility, as shown by the positive coefficients on the log fraction of time spent completely at home. The estimates in column (5) confirm this result and show that controlling for daily weather variables (precipitation, snow and mean temperature) which are also drivers of mobility, does not substantially change the estimated MRPs effects. Columns (6) to (8) complete the analysis by entering indicators for the date at which a county experiences its

Table 4: Two-Way Fixed Effects Estimates of the Impact of MRPs on Log *Completely Home (%)*

	Dependent variable:							
	Completely Home (log)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
State Emergency	0.008***			0.010***	0.011***	0.009***	0.013***	0.011***
	(0.003)			(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
State Shelter-in-Place	0.043***			0.032***	0.033***	0.027***	0.026***	0.022***
	(0.004)			(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
State Earliest Restriction or Closure	0.022***			0.017***	0.020***	0.016***	0.022***	0.018***
	(0.005)			(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
County Emergency		0.030***		0.018***	0.018***	0.010***	0.017***	0.010***
		(0.004)		(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
County Shelter-in-Place		0.090***		0.066***	0.064***	0.063***	0.055***	0.056***
		(0.008)		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
County Pop. Share Under City Policy			0.227***	0.193***	0.188***	0.165***	0.136***	0.122***
			(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Temperature (mean)					$-$0.001***	$-$0.001***	$-$0.001***	$-$0.001***
					(0.0001)	(0.0001)	(0.0001)	(0.0001)
Precipitation					0.063***	0.064***	0.063***	0.063***
					(0.001)	(0.001)	(0.001)	(0.001)
Snow					0.035***	0.035***	0.035***	0.035***
					(0.001)	(0.001)	(0.001)	(0.001)
Cases $>$ 0						0.084***		0.077***
						(0.003)		(0.003)
Deaths $>$ 0							0.080***	0.069***
							(0.004)	(0.004)
Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	285,405	285,405	285,405	285,405	282,997	282,997	282,997	282,997

Notes: Standard errors clustered at the county level. *p<0.1; **p<0.05; ***p<0.01

first COVID-19 case or confirmed death. Again, the addition of these factors does not lead to a meaningful change in the estimated MRP effects.

The evidence in Table 4, albeit for only one mobility outcome, might lead one to conclude that MRPs had the intended effect by significantly changing mobility behavior through an increase in the fraction of time spent at home. In this case, the estimated MRPs effects are stable and robust across a wide range of specifications, including the joint estimation of MRP effects, and the State SIPO policies are the strongest determinants of reduced mobility. Next, we document that the remarkable stability of the TWFE estimates of the impact of MRPs on log *Completely Home (%)* generally applies to all the other mobility outcomes.

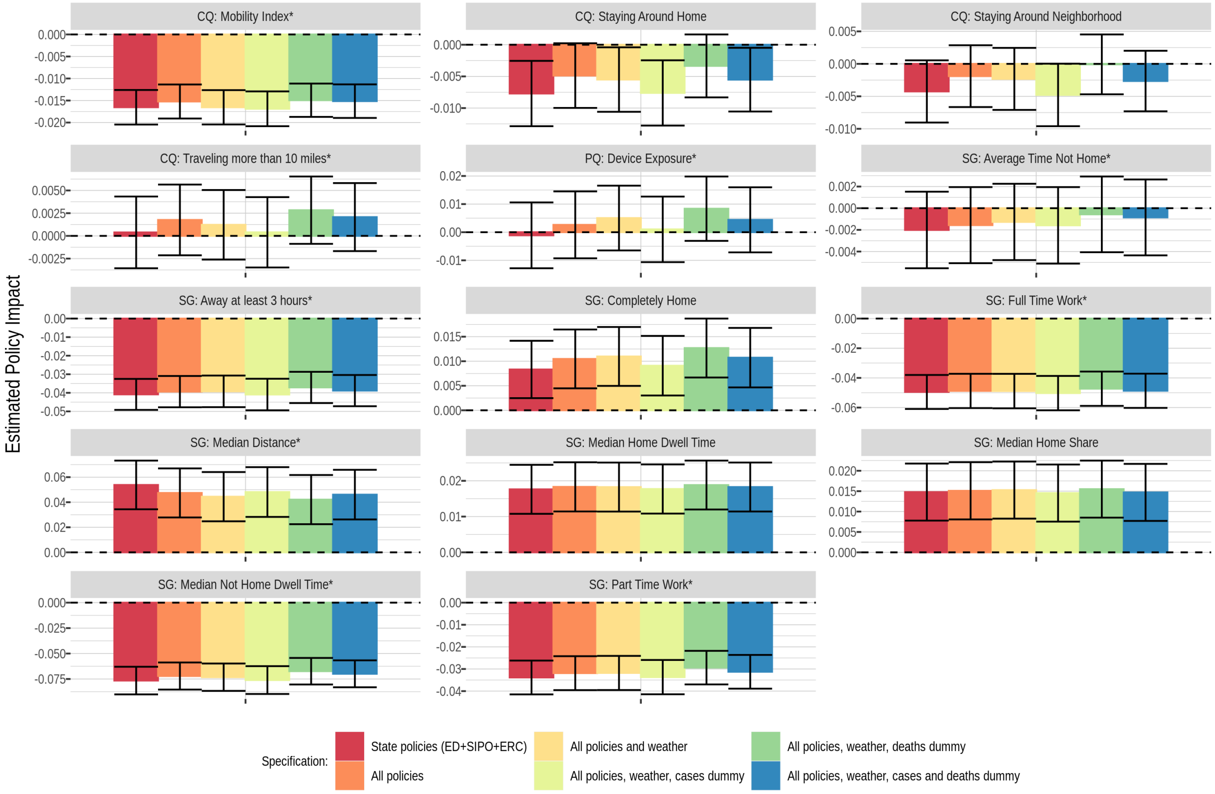
4.1 Robustness of the TWFE estimates

In this section we continue the analysis of the log-linear specification and graphically report in Figure 7 the estimates of the impact of State ED along with their 95% confidence intervals for the 14 mobility and activity indicators that take strictly positive values.[8] Each subplot reports estimates specific to a given mobility outcome. The height of each bar correspond to the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

The subplot in the upper left corner is for the Cuebiq (CQ) *Mobility Index*. We use different colors to represent the specification of the predictors of mobility (detailed in the legend). Starting from the left, the red bar shows the estimate of the effect of State ED on the *Mobility Index* when only the state-level MRPs are included in the model (along with the county and date fixed effects), corresponding to column (1) in Table 4. The specification shown by the orange bar adds all other policies (county and city-level), and the one in yellow adds the weather controls. The specifications shown with the pale green and darker green add in turn the indicator for the date of the first case or first death related COVID-19 in each county. Finally, the bar to the right (blue) corresponds to column (8) in Table 4 and includes all policies and predictors in the model. For example, in the case of the Safegraph *Median Distance*, we find that the TWFE estimates point to a robust and statistically significant positive impact, indicating that the state ED policies were effective in reducing mobility, by about 0.02 log points, or 2% (recall that variables were re-signed so that positive impacts can be interpreted as increase in social distancing).

[8] Since the Google Mobility indicators are reported as a change relative to a reference week and thus contain negative values, we cannot analyze them with a log-linear model. We examine models where the dependent variable is in levels below.

Figure 7: Estimated Impact of State ED on Log Mobility Outcomes



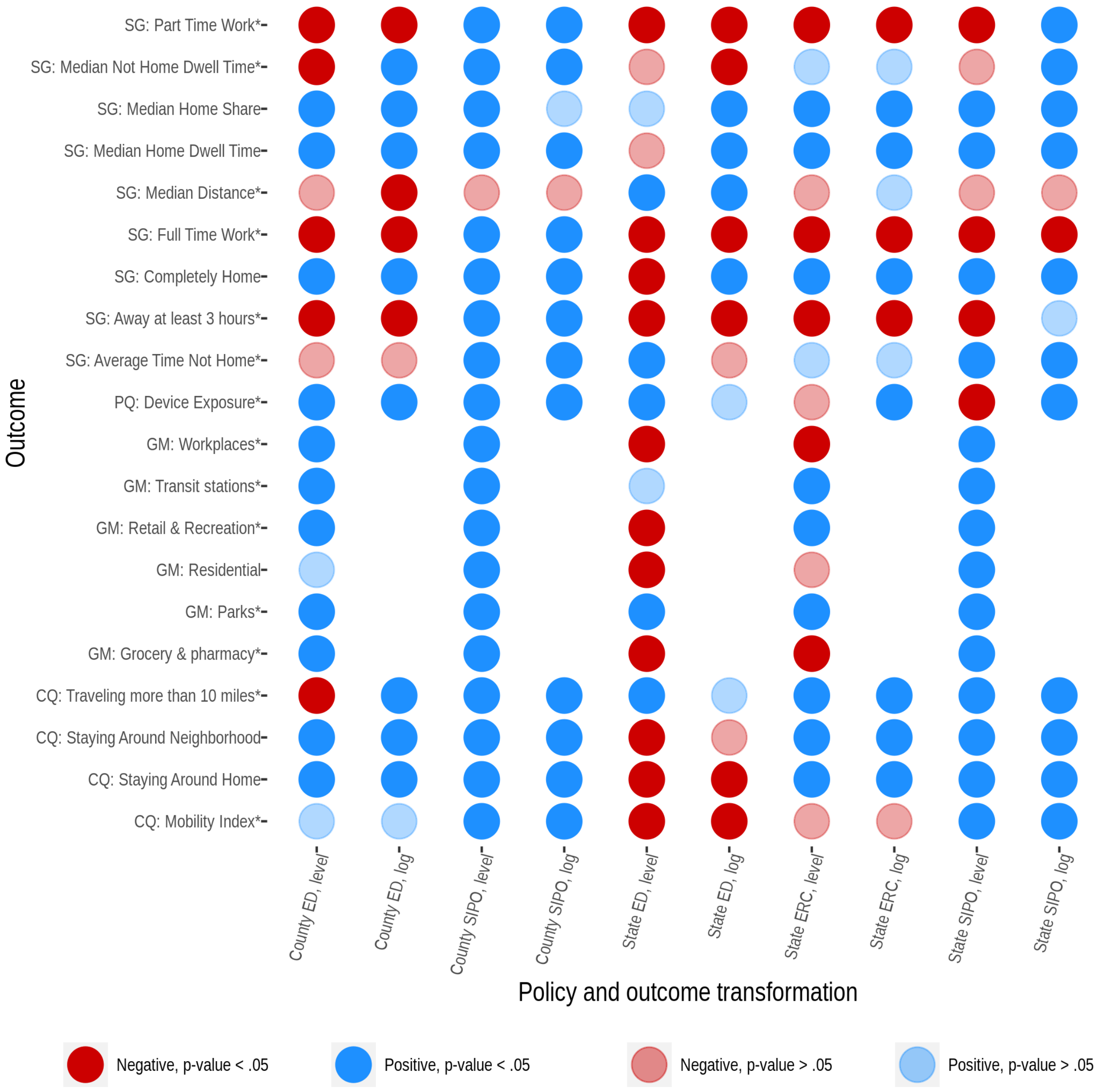
Notes: Figure 7 reports the TWFE estimates of the impact of State ED on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Remarkably, the robustness of the TWFE estimate of the impact of State ED policies on mobility across the log-linear specifications that include in turn all other MRP (state, county, and city) is evident for all 14 mobility outcomes in Figure 7. Specifically, for most of the mobility outcomes (i.e., within each individual subplot bar graph), the point estimates are similar and the 95% confidence intervals are generally overlapping, and do not meaningfully change across the 6 specifications.

In the Appendix, we report a series of additional figures for the four remaining MRPs, similarly structured as Figure 7 that further document the stability of the TWFE estimates of the impact of MRPs on our 14 mobility outcome modelled with a log-linear specification: State SIPO (Figure A.1), State ERC (A.2), County ED (A.3), and County SIPO (A.4). Similar to the evidence in

Figure 7, we find that for all MRPs, the TWFE estimates generally have roughly the same rough magnitudes across specifications.

Figure 8: Estimated Sign and Statistical Significance of Estimated Impact of MRPs on Mobility Outcomes



Notes: Figure 8 reports the TWFE estimates of the impact of all MRPs on 20 mobility outcomes using either a linear or log-linear specification. All models include all MRPs included simultaneously and the weather and Covid-19 severity controls

Based on the analysis of a single outcome (for example, log *Completely Home (%)*), it would seem natural to conclude that MRPs played an important role in reducing mobility and increasing social distancing during the first-wave of the COVID-19 pandemic, as other researchers have noted (e.g., Holtz et al. 2020). However, further analysis raises two key concerns. First, the stability of point estimates across specifications in Figure 7 is a misleading, or at least incomplete, signal of robustness. While such stability is often used as a basis for establishing a causal relationship, we show later in our event study analysis that assumptions needed for credible difference-in-difference estimation may not hold. Second, we find persistent issues with the *direction* of our main effect of interest: for several outcomes and policies, the MRPs are estimated to *decrease* distancing rather than increasing it as one would expect. To summarize this evidence visually, Figure 8 shows the sign and statistical significance of the estimated coefficient on the MRPs from regressions where the dependent variable enters either in levels, or through a log-transformation (as in Figure 7 and all other related figures in the Appendix).

Throughout Figure 8, we rely on the "full" specification used in the final column of Table 4 (all policies included simultaneously and including weather and Covid-19 severity controls). We therefore examine multidimensional space of researcher's degrees of flexibility to focus on the impact of the choice of outcome variable and MRP considered, investigating two possible transformations of the outcome. Remember that all dependent variables have been re-signed to indicate mobility reduction, so that we expect MRPs to lead to an increase in social distancing and increase the dependent variables. The rows in the figure correspond to each of our 20 outcome variables while the columns indicate a given MRP and whether the outcome variable is modeled in log or level. Blue dots indicate the expected sign (positive, i.e., more social distancing) while red dots indicate the unexpected sign. For either color, the darker version of the color indicates the estimated MRP coefficient was statistically significant at the 5% level.

Statistically significant coefficients of unexpected sign (shown in dark red) are prevalent in Figure 8. This is true for the majority of outcome variables for State ED in both level and log-linear specification. Incorrectly signed impacts are also the majority result across policy and outcome transformation for four outcome variables (*Part Time Work, Median Distance, Full Time Work,* and *Away at Least 3 Hours*). In addition to the prevalence of unexpected signs, we also observe signs switching from statistically significant in one direction to significant in the other direction as we vary either the outcome variable, or the policy and dependent variable transformation. For example, the impact of State SIPO on *Device Exposure* is negative for the specification in levels and positive (as expected) for the log-linear specification, and statistically significant in both cases. Similar sign switches of the estimated impact of the same MRP on the same outcome are observed for *Part Time Work*, *Median Not Home Dwel Time*, *Completely Home*, and *Travelling More than 10 Miles*.

What can explain such pervasive reversal of estimated MRPs impact across specifications? A first possibility comes from the recent econometric literature documenting that when treatment adoption is staggered and when there is heterogeneity between cohorts of units that become treated at different times, the TWFE estimator fails to recover a meaningful average treatment effect (de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; Sun and Abraham, 2020; Borusyak et al., 2021). As shown in Figure 5 all MRPs analyzed in this paper and in the previous literature were implemented in a staggered manner. Furthermore, counties that implemented MRPs at different times were likely impacted differently by these policies, which could introduce between adoption cohorts treatment effect heterogeneity. Thus a first consideration is that the estimates reported in the previous section might fail to recover the ATT due to issues inherent to how the TWFE estimator handles treatment effect heterogeneity.[9]

A second possibility is that the parallel trend assumption—one of the central assumptions underlying the identification of β_p in the TWFE estimator—is invalid for some outcomes or MRPs. In particular, in the case we highlighted earlier where the impact of state ED policies on the *Completely Home (%)* variable in the log-linear specification is positive and statistically significant but negative and statistically significant in the specification in levels could indicate that the parallel trend assumption holds in the log-linear specification, but not in the level specification (or vice

[9] In subsection A.4 in the Appendix, we perform the Goodman-Bacon decomposition that highlights how the implicit use of counties treated earlier as controls for units treated later can be problematic. In particular, we find that removing all comparisons that involve the use of counties treated earlier as control units from the TWFE estimator frequently changes the *sign* of the estimated effect for the State ED and State ERC policies (the two MRPs that do not have a well-defined control group throughout our sample period). However, the sign and magnitude of the estimated effect remain stable for most other outcome variables, suggesting that a different assumption might also be violated.

versa).[10] Recent work by Roth and Sant'Anna (2021) highlight that in general, the parallel trend assumption is sensitive to the specific functional form chosen in the estimating equation. The next section presents a detailed analysis of pre-trends for all the mobility outcomes and MRPs.

5 Event study analysis

5.1 Specification

The DD analysis of the impact of MRPs on the mobility outcomes presented in the previous section relies on the parallel trend assumption, which states that trends in the outcome (mobility indicator) before the adoption of an MRP (known as "pre-trends") are the same in the treatment (adopters) and control (non-adopters) counties. An event study analysis provides a simple graphical approach at documenting pre-trends, which we now implement and supplement with F-tests on the pre-trend coefficients in order to test the parallel trend assumption.

To proceed, we use the same framework as in Equation (1), but index days relative to the event (date of adoption of an MRP) by k. Therefore D_{ipk} is a dummy variable equal to 1 when county i is k days away from being "treated" by a given MRP policy p within the set P. As before, the policy set is $P = \{\text{state ED, state SIPO, state ERC, county ED, county SIPO}\}$. Formally, $D_{ipk} = \mathbb{1}\{t - MRP_{ip} = k\}$, where MRP_{ip} is the policy p for county i. We take the common event study approach of including a single dummy for all relative days before our event window (over 20 days pre-adoption of a policy) denoted by $k = -21^{-}$, and another for all relative days after, $k = 21^{+}$ (over 20 days post-adoption):

$$Y_{it} = \alpha_i + \delta_t + \left(\sum_{p \in P} \sum_{k=-21^{-}}^{k=21^{+}} \theta_{pk} \cdot D_{ipk} \right) + X_{it}^{\prime} \gamma + \varepsilon_{it} \tag{2}$$

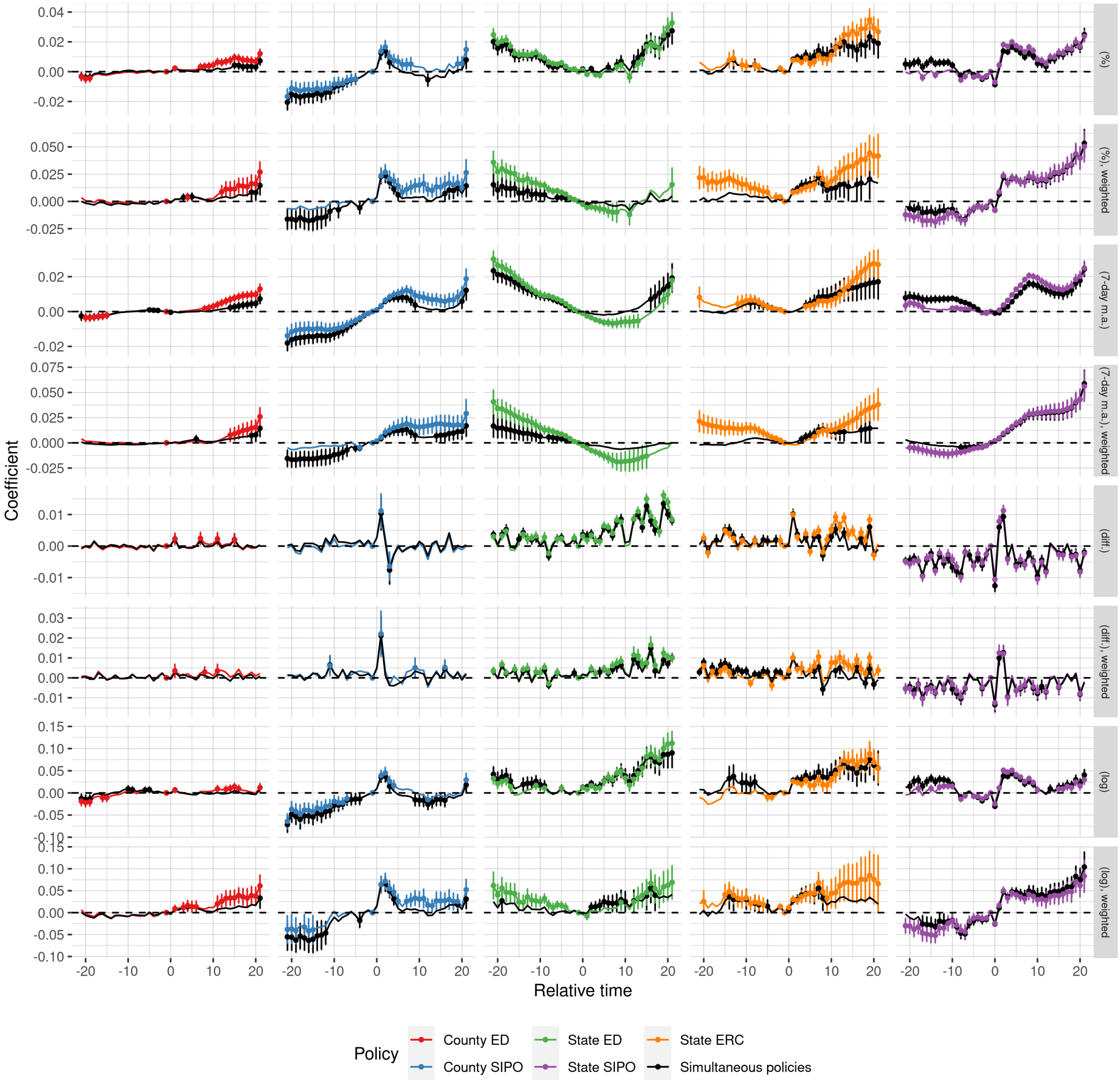
[10] This concern over possible failure of the parallel trend assumption is heightened by the observation that many individuals may have curtailed their mobility behaviors even if their state or county of residence was not under an MRP, or in anticipation to the pre-announcement of a SIPO or other MRPs.

The day before the adoption of a policy ($k = -1$) is omitted to serve as baseline. Like in the standard DD analysis, we include county fixed effects (α_i) to control for unobserved time-invariant differences between counties, and day fixed effects (δ_t) to control for unobserved time-varying drivers of mobility and activity that are common to all counties. We also include the full set of time-varying control variables (X_{it}) we considered earlier (see last column in Table 4.) Finally, ε_{it} is a county-day specific error term. Standard errors are clustered at the county level.

5.2 Results from the event study analysis

To begin, we focus on a single outcome, *Completely Home (%)* and graphically present estimates of the event-study θ_{pk} coefficients (along with the corresponding 95% confidence intervals) in Figure 9. We consider 8 different possible specifications of the dependent variable, including the level and log-transformed specifications analyzed earlier. We also consider specifications where the dependent variable is first-differenced, and specifications where the dependent variable is a 7-day moving average of the mobility outcome (as opposed to the daily-level value). We either weight the underlying regressions by county-level population, or estimate them without weights. We implement these 8 specifications because they have been used in the previous literature. The labels on the right-end margins of each row in Figure 9 indicate the specification of the dependent variable and use of weights. The 5 columns in Figure 9 correspond to the five MRPs we analyze. Note that in each column, we report the estimated event-study coefficients for the relevant MRP estimated in isolation (color) and in combination with the four other MRPs (black). To the best of our knowledge, only a couple of papers have studied the identification of treatment effects under multiple treatment. In an application focused on the impacts of school bonds on housing prices, Cellini et al. (2010) derive estimators of the ”treatment-on-the-treated” and ”intent-to-treat” when multiple bonds measures have been passed, assuming that dynamic effects only depend on the length of time elapsed since treatment. Using simulation studies, Sandler and Sandler (2014) highlight that in the case of multiple treatment, ignoring one or more treatment produces biased estimates in general.

Figure 9: Estimated Event-Study Coefficients for *Completely Home (%)*



Notes: Figure 9 reports the event-study estimates of θ_{pk} from equation (2) above for the *Completely Home (%)* outcome under several specifications. The transformation and use of regression weights varies by row and the MRP varies by column. The estimates are presented for both the MRP estimated in isolation (shown by the colored lines) and jointly with all MRPs (shown by the black lines). Whiskers represent the 95% confidence intervals.

There are three main results that readily emerge from Figure 9. First, focusing on the post MRP adoption period (to the right of 0 on the x-axis), most specifications and MRPs show results that are consistent with MRPs increasing the *Completely Home (%)* variable (although there are a few exceptions such as the 7-day moving average model for state ED and the first-differenced models for state SIPOs). Second, with the exception of county ED, statistically significant and sizable pre-trends are apparent for most specifications and MRPs. Third, including all policy variables simultaneously (black estimates) as opposed to individually, using population weights, or first-differencing the outcome does not lead in general to a meaningful improvement in supporting the parallel trends assumption.

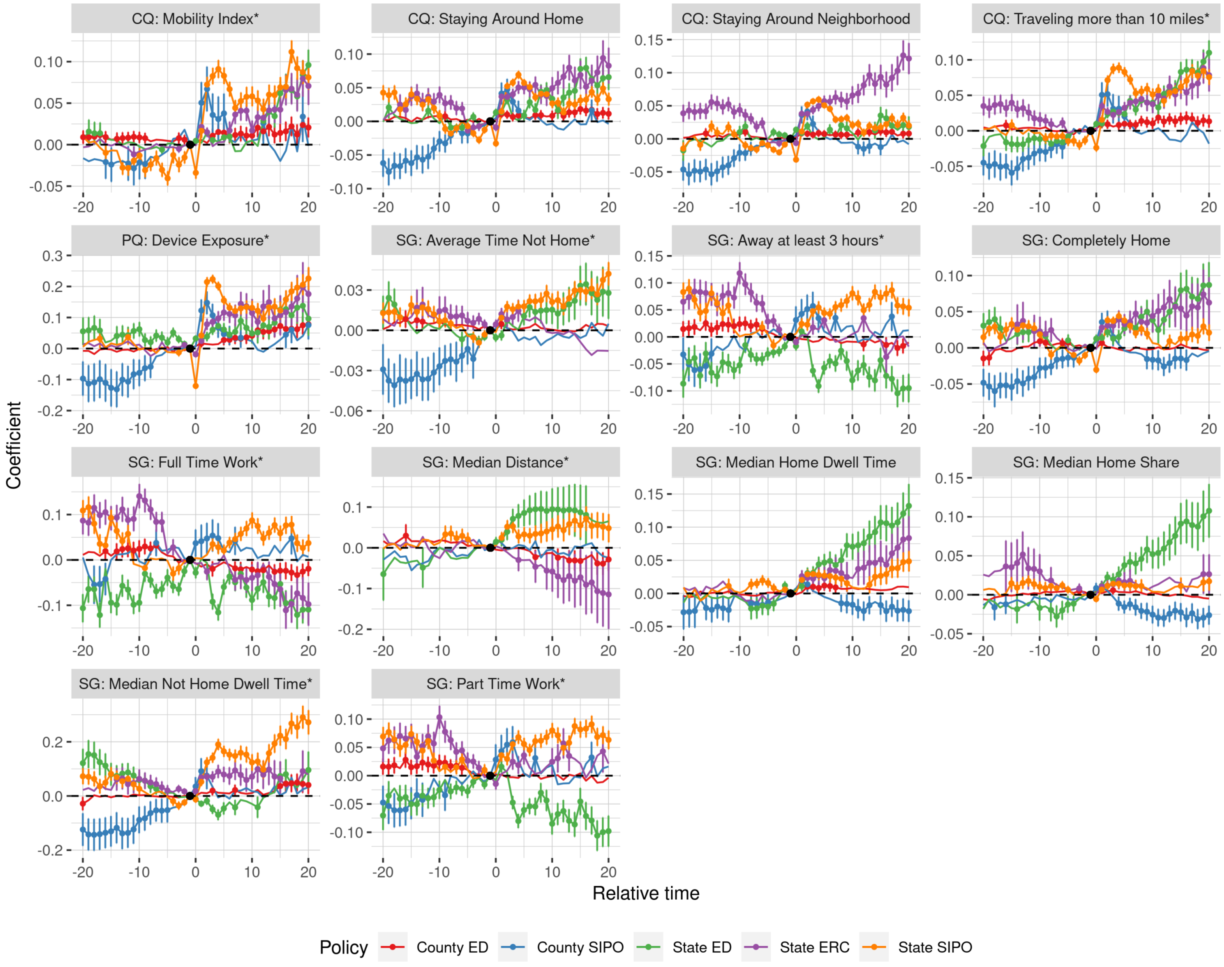
To assess whether the issues with pre-trends and instability of the estimates identified previously could simply be due to our choice of dependent variable, we present additional slices of the degrees of flexibility space below. Figures 10 and Appendix Figure A.10 continue the event-study analysis by reporting estimates for all mobility outcomes, for the two initial specifications of the dependent variable (log-linear and in level). Each box in the figures correspond to a given mobility outcome and shows the estimated event-study coefficients (θ_{pk}) and 95% confidence intervals for the 5 MRPs, each from a separate regression.

Examining first the results from the specification in levels for the outcomes, one can observe the complex dynamics in the impact of MRPs on mobility and activity outcomes in the post adoption period, as was also shown in Figure 9. Some policies have rapid impacts in decreasing mobility, with impacts lasting from a few days, to the full 20 days in the post-adoption period. Notably, some of the post MRP adoption estimates have the "wrong" sign, indicating that the introduction of an MRP increased mobility and reduced social distancing (e.g., the state ED effect on Safegraph *Away at least 3 hours*, or the county SIPO effect on Safegraph *Median Home Dwell Time* and *Median Home Share (%)*). At the same time, it is also surprising to observe that the different MRPs can have differently signed impacts on the same mobility outcomes (e.g., Safegraph *Full Time Work*), which is increased by State SIPO and reduced by State ED.

Most importantly, however, it is again evident that sizable and statistically significant pre-trend differences exist for at least one policy for each outcome. This adds to the evidence of failure of the parallel trend assumption already shown in Figure 9 for the Safegraph *Completely Home (%)* variable for all MRPs and a wider range of specifications. That is, looking at all available mobility outcomes and for most policies analyzed, a large number of pre-trend coefficients are statistically different from zero at the 5% level. For example, for the Cuebiq *Travelling More Than*

10 Miles outcome, we can easily detect significant pre-trends for all five MRPs. Moreover, the pre-trend patterns tend to vary across policies, again in the case of Cuebiq *Travelling More Than 10 Miles* outcome, the pre-trend coefficients are mostly positive for state ERC and county ED, while negative for state ED, State SIPO, and County SIPO.

Figure 10: Estimated Event-Study Coefficients for All Mobility Outcomes, Based on the Log-Linear Specification



Notes: Figure 10 reports the event-study estimates of θ_{pk} from equation (2) above for 14 mobility outcomes, using the log-linear specification. Each estimated impact of MRPs are estimated from separate regressions. Whiskers represent the 95% confidence intervals.

Figure A.10 in the Appendix is structured similarly as Figure 10 but focuses on models with a specification in levels, which includes mobility outcomes from Google [11] For all outcomes,

[11] Recall that due to the log-linear specification of the model, outcomes from Google Mobility are not included since they can on zero and negative values.

we detect statistically significant pre-trends for one or more of the policies. Moreover, for any given outcome, the pre-trends can be consistently positive or negative, depending on the MRP considered. This heterogeneity in the impact of different MRPs on the same outcome is also observed in the post-adoption period.

Figure A.11 summarizes the information about the pre-trends across the 8 specifications considered in Figure 9 and all outcomes. The figure reports the p-values from F-tests testing the null hypothesis that the pre-MRP-adoption coefficients are jointly equal to zero. It is configured as a heat-map, with the columns indicating the five MRPs estimated individually and then estimated jointly (far right column). The rows correspond to the eight specifications for the dependent variable that were considered in Figure 9. Each cell summarizes the results of 14 F-tests on pre-trends coefficients, corresponding to the 14 mobility outcomes we analyze from Safegraph, Cuebiq, and PlaceIQ.[12] The number in each cell corresponds to the fraction of tests (out of 14) where the p-value on the null hypothesis exceeds 0.05, indicating that the null hypothesis would be rejected at the usual 5% significance level.

The results further confirm some of the evidence already documented: significant pre-trends are pervasive. For example, *all* 14 out of 14 "no pre-trends" tests are rejected for all specifications of the dependent variable for the state ED, state SIPO, state ERC, and simultaneous policies models (as shown with a fraction of zero in the cells). For county SIPO, the parallel trend assumption is rejected for all 14 outcomes in half of the specifications (first-difference, first-difference weighted, log-linear weighted, and 7-day moving average weighted). The MRP for which the parallel trend assumption appears most supported is county ED, although we fail to rejected at most for 6 of the 14 outcomes in 1 specification. All together, 672 F-tests on pre-trends coefficients were conducted to construct Figure A.11. Out of those, we fail to reject the null hypothesis of "no pre-trends" only 34 times (out 672 tests) at the 5% significance level. Note that for simplicity we do not account for multiple hypothesis testing, since controlling for it would only reduce the number of rejected cases, reinforcing the result that pre-trends are pervasive.

[12] The Google Mobility outcomes are ignored here since we cannot apply the log transformation on them.

Differences-in-differences and event-study approaches, implemented with standard TWFE estimators to the mobility outcomes to estimate the impact of MRPs provides definitive evidence of estimates that are "robust" to the inclusion of covariates. But we also observe consistent violation of the parallel trends assumption—rejection of the null hypothesis of no pre-trends—and major specification issues where the sign of the effect depends on choices made on the transformation of the dependent variable, e.g., linear versus log-linear versus first-difference. This combined pattern of stability of estimates within-specification, but instability across specifications (for the same outcome), and statistically significant pre-trends is observed for most outcomes considered. We interpret this evidence that one or more of the assumptions required to interpret the DD and event-study estimates as Average Treatment Effects on the Treated (ATT) are likely violated.

6 Heterogeneity-robust estimators for staggered adoption designs

The recent econometric literature highlights that under the presence of staggered treatment and treatment effect heterogeneity, the TWFE regressions underlying DD and event-study analyses typically fail to recover the ATT (Sun and Abraham, 2020; de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021). The bias arises from the implicit construction of control groups which include units that are themselves under the effect of a treatment.[13] For example, in our setting, TWFE regressions will *implicitly* use counties treated on March 15 to serve as controls for counties treated on March 20; this might be a bad comparison if the treatment effect is dynamic, or if the treatment effect is heterogeneous across units treated on different days. This same literature also provides various solutions to address the deficiencies of the TWFE estimators. New event-study methods, for example from Callaway and Sant'Anna (2020), reduce the problem to

[13] Goodman-Bacon (2021) shows this issue with a decomposition result that reframes the TWFE estimator as a weighted average of all possible two-by-two difference-in-differences estimators. Sun and Abraham (2020) propose a general decomposition of the event study estimator for alternative choices of specifications that nests TWFE as a special case.

estimating “clean” difference-in-differences for each group of units that receive treatment at the same time (also called “cohorts”) by comparing these units to proper “controls” (which are units either not yet treated, or never treated). The cohort-specific difference-in-differences estimates can then be re-aggregated to produce event study estimates that are free of the bias due to using treated units as implicit controls.[14]

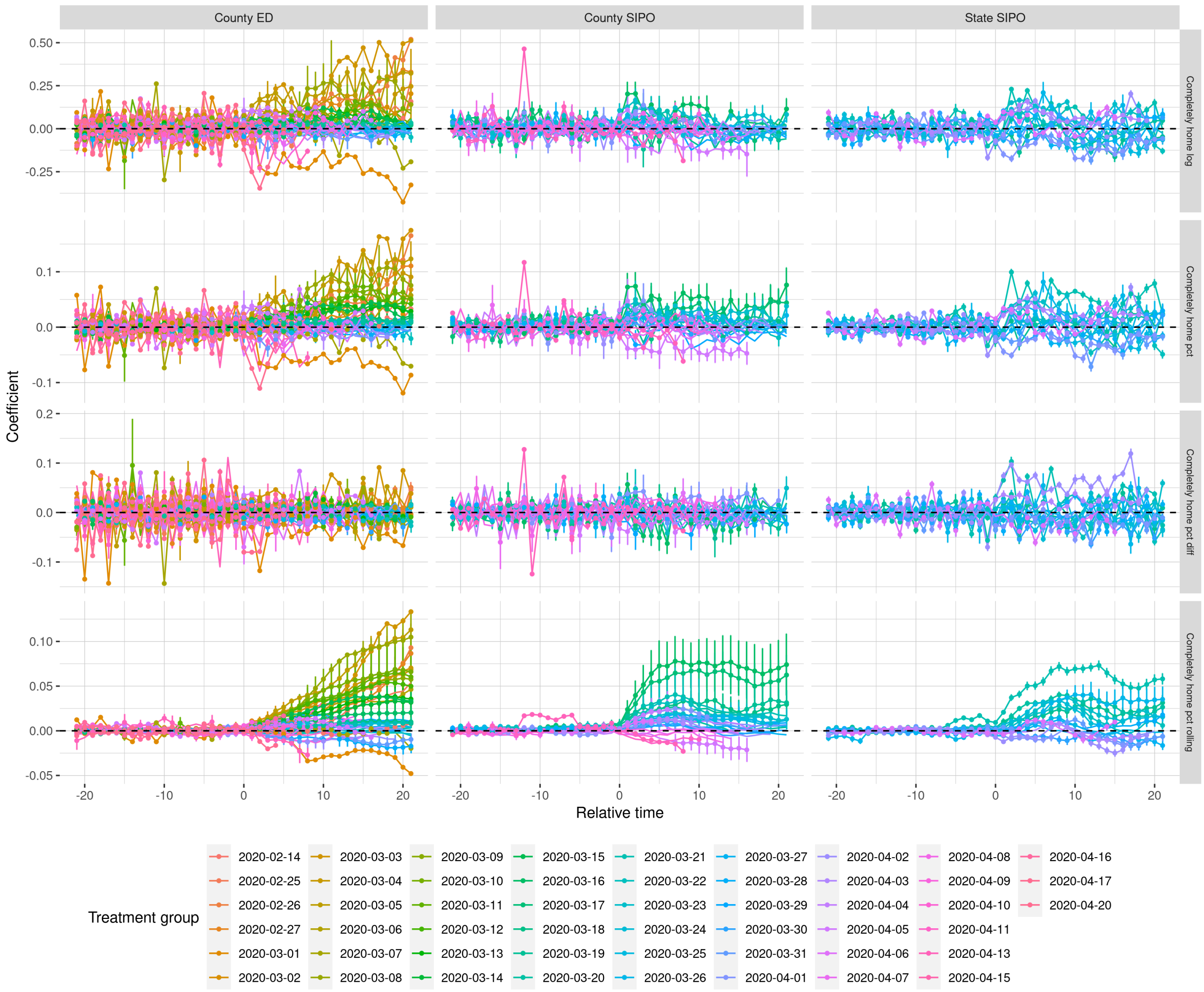
Applying the Callaway and Sant’Anna (CS) estimator in our setting with the control variables leads to several observations being dropped, sometimes to the extent that a given model will not be estimated.[15] With this as a limitation, we only consider CS estimates that consider one MRP at a time.In addition, because the CS estimator and related estimators only compare units that are treated to units that are never or not-yet-treated, they explicitly restrict the number of post-treatment periods for which we can estimate a treatment effect. In our sample all units are eventually treated by a State ED and a State ERC, and the last unit to receive a state ED does so on March 16. This means that no treatment effect of State ED can be estimated after March 16 with the CS estimator. To ease comparison with the previous results in the paper, we report CS estimates focusing on the three MRPs for which there are always proper “control” units during the sample period, namely County ED, County SIPO, and State SIPO.

Figure 11 reports the CS estimates for the cohort-specific impact of these three MRPs on the *Completely Home (%)* mobility outcome. These CS estimates are analog to the standard event-study regression specification results presented above in Figure 9. Here, each MRP cohort (defined as each individual MRP adoption date in the sample) is represented by a separate line and color. We continue to present estimates for various transformation of the dependent variable as the previous sections have shown how seemingly minor specification changes (e.g., linear versus log-linear) can lead to widely different estimates.

[14] de Chaisemartin and D’Haultfœuille (2020a) propose a similar estimator motivated from the point of view of a social planner who could discount future treatment effects, à la Manski (2005). Sun and Abraham (2020) propose an alternative estimator obtained by interacting event-study coefficients with dummies for each treatment cohort. Both alternatives weight cohort-specific treatment effects by relative cohort size.

[15] This occurs because the CS estimator requires a certain overlap of the categorical covariates between treatment and control groups. While the overlap holds over the whole sample, it can fail to hold in some subsamples of smaller size. In that case, the problematic subsample is simply dropped.

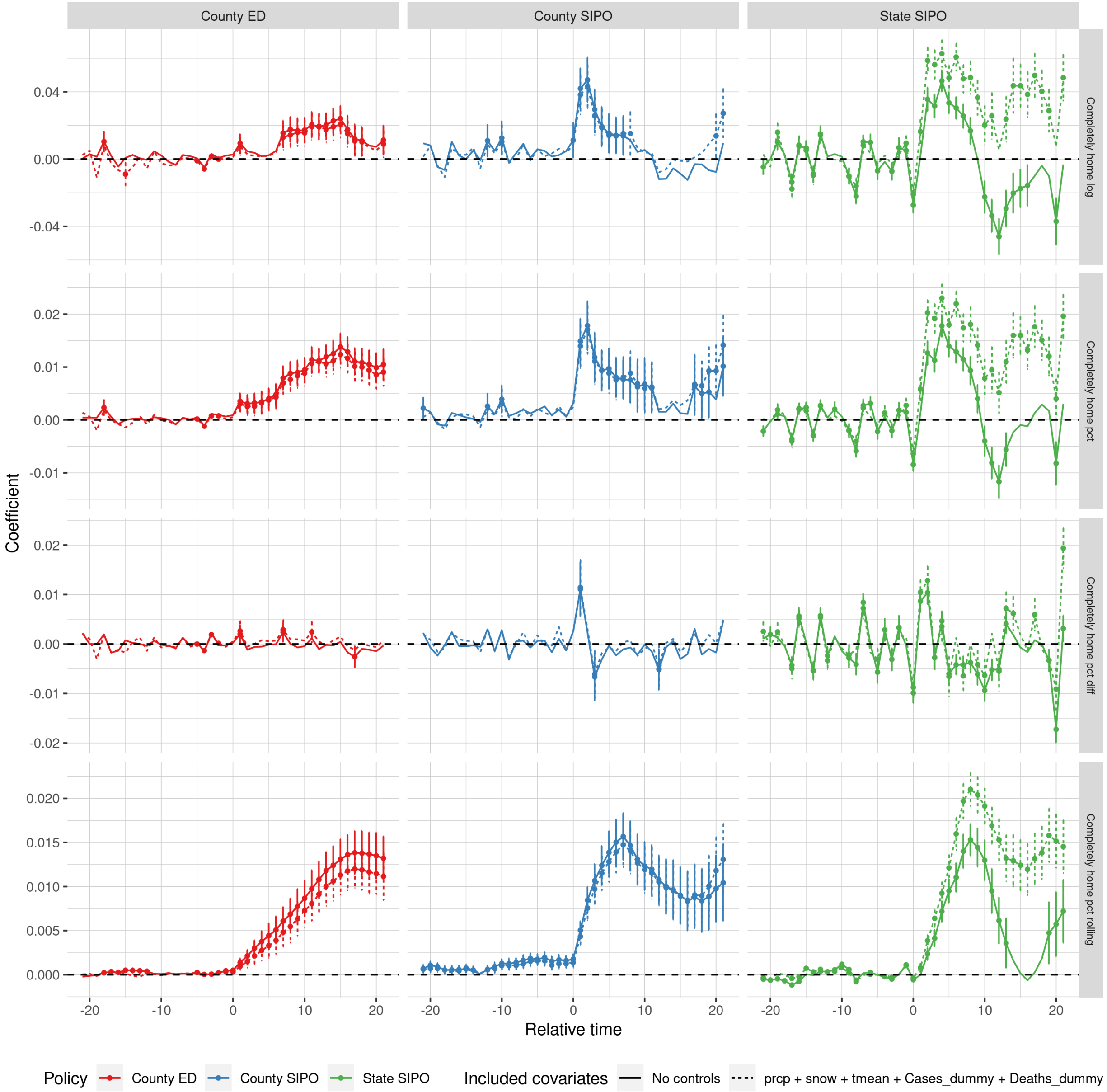
Figure 11: Callaway and Sant'Anna Estimates of MRP Impacts by Treatment Cohort



The CS estimates in Figure 11 are noisy, as might be expected since they represent around 50 smaller, cohort sub-samples from the overall data set. At this level, we also find that significant pre-trend effects are still present. The estimates for the 7-day moving average specification (row four) appear better-behaved, with smaller pre-trend coefficients and some evidence of ATTs that indicate that MRPs reduced mobility. Yet, even for this specification, we still observe estimates suggesting that MRPs *increased* mobility for some adoption cohorts. We also find that the estimated ATTs tend to be higher in places that implemented the policies earlier. This could be due to a greater response to treatment among early adopters of MRP, but it could also be due to increased mobility among the control units. Finally, because we are using units that are "not yet treated" as controls, this could also reflect changes in the composition of the control group.

Figure 12 is structured as Figure 11 but combines the cohort-specific estimates into a single estimate using the aggregation method proposed in Callaway and Sant'Anna (2020) to obtain heterogeneity-robust event-study estimates of the ATTs. Again, we focus on the *Completely Home (%)* mobility outcome for illustrative purposes. For each MRP and transformation of the dependent variable, two set of estimates are reported: with and without covariates. While the overall picture is better than in the case of standard event study estimates, the aggregated CS estimates still point to important pre-trend differences, especially for County and State SIPO. Notably, however, the pre-event coefficients are small relative to post-treatment coefficients.

Figure 12: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts



Taken together, the evidence in Figure 12 for *Completely Home (%)* is mixed regarding the impact of MRPs on time spent completely at home. In the case of County ED, the estimates point to a positive impact on time spent completely at home (except for the specification using first differences of the outcome in row 3). Similar patterns are observed for County SIPO, although the estimates are noisier and pre-trend differences are more noticeable. For State SIPO the estimates are more dependent on the inclusion or exclusion of the covariates (weather variables and indicators for county-level COVID-19 cases and mortality).

Among all specifications considered, the CS event study estimates for the 7-day moving average transformation appear most consistent with the "no pre-trend" hypothesis. As shown in Figure A.12 in the Appendix, this holds across most outcome variables, although unexpected signs still appear for a subset of outcome variables, especially looking at the impacts of County ED. This suggests that—in combination with the heterogeneity-robust estimation—temporal smoothing over a longer periods longer than 1 day helps the credibility of the estimator. In contrast, regressions using the CS estimator still appear very noisy when the outcome is in level, in log, or first-differenced (see Figures A.14, A.13, and A.15 in the Appendix respectively). One implication is that time-aggregating the data may help improve the credibility of the parallel-trend assumption. Notably, a similar finding emerges from the previous literature where event studies conducted with state-level data (as opposed to county-level) generally produce well-behaved estimates (e.g., Andersen 2020; Abouk and Heydari 2021; Dave et al. 2021). Thus, it may be that fine scale event study estimates in our setting (daily at the county level) are beyond the capacity of the data and difference-in-difference based methods, unless one is willing to relax the parallel-trend assumption.

7 Conclusion

In this article we analyze the impact of mobility reducing policies (MRPs) on mobility outcomes during the onset of the COVID-19 pandemic, leveraging a uniquely broad coverage of policies spanning the state, county, and city levels in the United States. Generalizing from previous research which focused on selected mobility outcomes and used specific functional forms and TWFE estimators, we analyze twenty mobility measures derived from aggregated mobile device signal data using a wide range of specifications and estimators. We also analyze five different types of MRPs both in isolation and jointly, consistent with the manner they were implemented at the state, county, or city level.

In this context, we find that the standard TWFE and event-study estimates are highly sensitive to choices of mobility outcome and transformation of the outcome (e.g., linear versus log-linear) being implemented. In particular, we uncover systematic violations of the standard parallel-trend assumption for virtually all outcomes and MRPs. We also document that due to researcher's degrees of flexibility, it is possible to focus on specific outcome variables and functional forms that produce "well-behaved" estimates, and that usual robustness-tests based on the sequential addition of covariates points to the stability of those estimates. Yet, for several outcomes we obtain widely different results for the impact of MRPs on the same outcome (often with differing sign), but analyzed through TWFE estimator on different functional forms. Based on this, one could conclude that both the magnitude and the sign of the ATT associated with the effect of mobility reducing policies on mobility remains largely unknown.

An emerging econometric literature highlights that in the presence of treatment effect heterogeneity and staggered treatment adoption, standard difference-in-differences and event-study estimators can be severely biased (Goodman-Bacon, 2021; Sun and Abraham, 2020; de Chaisemartin and D'Haultfœuille, 2020b; Borusyak et al., 2021). Indeed, we hypothesize that the unreliability of the basic TWFE estimates we uncover is likely due to heterogeneity in the treatment effect of MRPs between groups of counties that receive treatment at different times. When we implement the recent estimator by Callaway and Sant'Anna (2020), we generally obtain more stable estimates of the overall ATT of MRPs, where the direction of the estimated impact is more robust across specifications. At the same time, these important new methods have noteworthy limitations. In particular, we find that due to the interaction and overlap between different MRPs, heterogeneity-robust difference-in-differences estimators fail to credibly estimate the impacts of multiple treatments. Further, several of the estimated pre-MRP coefficients (i.e., "pre-trend" coefficients) are statistically different from zero using these more recent estimators, which weakens the case for a causal interpretation of the estimates. Finally, the transformation of the outcome variable still has large impact on the estimated ATT, sometimes changing the sign of the estimates.

Taken together, there are three main implications to these results. First, since the onset of the pandemic, dozens of research articles used standard difference-in-differences and event-study methods and broadly concluded that MRPs caused reductions in mobility, suggesting they were an important tool to curb the spread of COVID-19. In addition, a related literature used similar methods to estimate the impact of MRPs on COVID-19 cases and deaths. Our reading of the evidence suggests the previous literature appears to have focused on a set of specifications that are highly sensitive to varying minor modelling decisions. Indeed, most prior studies of the impact of MRPs found that such interventions reduce mobility, but this conclusion requires the researcher to favor one specification over other plausible ones. As we find here, for most mobility outcomes, there exist alternative specifications that provide contrary results. Overall, this suggests that robustness of recent evidence of a causal impact of MRPs on reducing mobility is not unequivocal.

Second, a lot of recent attention in the econometric literature concerned with staggered treatment adoption designs has focused on deriving treatment effect heterogeneity-robust estimators. Our empirical results indicate that such estimators can produce estimates that are substantially different than those relying on standard DD methods (this point is also highlighted in Baker et al., 2021), and indeed improve the credibility of the ATT estimates. However, our research highlights important limitations of these new estimators. To begin, they currently do not account for multiple, potentially interacting treatments. This is a serious limitation in the context of COVID-19, where states, counties, and cities all implemented different kinds of MRPs. More generally, this limitation applies in other settings, such as the study of federal and state overlapping policies (e.g., air quality regulations, education policy, health insurance programs for low-income families). Furthermore, causal identification with these estimators continues to rely on a parallel-trend assumption, and the validity of such assumption usually depends on a specific functional form specification for the dependent variable (Roth and Sant'Anna, 2021). When the chosen functional form substantially impacts the treatment effect estimates or the existence of a "pre-trend", there is little guidance from theory to select a specific transformation. The Change-in-Changes model developed by Athey and Imbens (2006) provides an alternative approach that does not depend on the scale of the outcome variable, but this estimator currently does not easily accommodate the inclusion of covariates and of many different treatment groups. Continuing to improve and generalize methods for staggered treatment adoption designs and multiple treatments is therefore an important avenue for future research.

Third, although difference-in-differences research designs are straightforward to implement with readily available data and can increase the credibility of observational studies (Angrist and Pischke, 2010), our findings highlights that researchers' degrees of flexibility can still have a marked influence on the sign and magnitude of the estimated treatment effects. We explore the role of researchers' degrees of flexibility along four different margins: choice of outcome variable, functional forms for the dependent variable, covariates, and estimation procedures. This is only a subset of the margins along which degrees of flexibility exist.[16] Estimating models for the full-set of possible specifications is often impossible, and researchers' degrees of flexibility in themselves are not to be abhorred, as they are an irreducible part of research. However, the scale of available discretion in choices made by researchers and their impacts on the estimates, even after conditioning on a "credible research design", raises the question of when to believe the reported estimates. The problem of pre-testing specifications and reporting only a set of results that "fits" within the researcher's overall narrative is well-known (Christensen and Miguel, 2018; Kasy, 2021). To mitigate it, Leamer (1983) argued that applied economists should report the full set of specifications that they estimated, not just the ones that "worked". Four decades later, his recommendation continues to be largely ignored. Even-though robustness tests are pervasive in the applied economics literature, we find that it is possible that robustness holds along one margin, while (consciously or unconsciously) ignoring that robustness fails along another dimension.

In light of our findings, instead of reporting robustness tests that work, one possibility could be to report robustness that are increasingly strict until they fail. The relevant question is not whether the results are robust, but instead "*how much* can we depart from the preferred choices before the main findings no longer hold?". In the context of COVID-19 research, this breaking-point threshold is very low, especially when using standard TWFE regressions. Due to the urgency of

[16] For example, additional margins of flexibility include the number of lags and leads in event studies, the definition of the treatment variables, or the choice of the temporal window used in the analysis.

obtaining actionable policy results, these shortcomings were not fully appreciated by previous research, including our own. The econometric work highlighting issues with common specifications developed extremely fast and in parallel with the applied COVID-19 research. Therefore, further developments are still needed to make rigorous causal inference about the effects of COVID-19 policies.

References

Abouk, R. and B. Heydari (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the united states. *Public Health Reports 136*(2), 245–252. PMID: 33400622.

Aiken, A. M., C. Davey, J. R. Hargreaves, and R. J. Hayes (2015, 07). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology 44*(5), 1572–1580.

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Yang (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics 191*, 104254.

Andersen, M. (2020). Early evidence on social distancing in response to COVID-19 in the united states. *Available at SSRN 3569368*.

Angrist, J. D. and J. S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Athey, S. and G. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–497.

Baker, A., D. F. Larcker, and C. C. Y. Wang (2021). How much should we trust staggered difference-in-differences estimates? *Working Paper*.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *Working paper*, 48.

Botvinik-Nezer and many authors (2020, Jun). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature 582*(7810), 84–88.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics In Press*.

Cellini, S., F. Ferreira, and J. Rothstein (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics 1*, 215–261.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

Clemens, M. A. and J. Hunt (2019). The labor market effects of refugee waves: Reconciling conflicting results. *ILR Review 72*(4), 818–857.

Couture, V., J. Dingel, A. Green, J. Handbury, and K. Williams (2020). Exposure indices derived from placeiq movement data. `https://github.com/COVIDExposureIndices/COVIDExposureIndices`. Accessed: 2020-05-01.

Dave, D., A. I. Friedson, K. Matsuzawa, and J. J. Sabia (2021). When do shelter-in-place orders fight COVID-19 best? policy heterogeneity across states and adoption time. *Economic Inquiry 59*(1), 29–52.

de Chaisemartin, C. and X. D'Haultfœuille (2020a). Difference-in-differences estimators of intertemporal treatment effects. *Working Paper*.

de Chaisemartin, C. and X. D'Haultfœuille (2020b, September). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Elenev, V., L. Quintero, A. Rebucci, and E. Simeonova (2021, July). Direct and spillover effects from staggered adoption of health policies: evidence from covid-19 stay-at-home orders. *NBER*, 48.

Foote, C. L. and C. F. Goetz (2008). The impact of legalized abortion on crime: Comment. *The Quarterly Journal of Economics 123*(1), 407–423.

Fullman, N., B. Bang-Jensen, G. Reinke, B. Magistro, R. Castellano, M. Erickson, K. Amano, J. Wilkerson, and C. Adolph (2020). State-level social distancing policies in response to COVID-19 in the US. `http://www.covid19statepolicy.org`. Version 1.98, November 9, 2020.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem,even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics In press*.

Goolsbee, A. and C. Syverson (2021). Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of Public Economics 193*(104311).

Gupta, S., T. D. Nguyen, F. L. Rojas, S. Raman, B. Lee, A. Bento, K. I. Simon, and C. Wing (2020). Tracking public and private responses to the COVID-19 epidemic: evidence from state and local government actions. *National Bureau of Economic Research*.

Holtz, D., M. Zhao, S. G. Benzell, C. Y. Cao, M. A. Rahimian, J. Yang, J. Allen, A. Collis, A. Moehring, T. Sowrirajan, et al. (2020). Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences 117*(33), 19837–19843.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco,

G. Ekpe, T. Pugatch, M. Saavedra, and Y. Stopnitzky (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry n/a*(n/a), 1–17.

Kasy, M. (2021). Of forking paths and tied hands: selective publication of findings, and what economists shoudl do about it. *Journal of Economic Perspectives 35*(3), 175–192.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*, Volume 53. Wiley New York.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*(1), 31–43.

Manski, C. F. (2005). *Social choice with partial knowledge of treatment response*. Princeton University Press.

National Association of Counties (NACo) (2020). COVID-19 pandemic response: County declaration and policies. `https://www.naco.org/resources/featured/counties-and-covid-19-safer-home-orders`. Accessed: 2020-04-15.

Painter, M. and T. Qiu (2021). Political beliefs affect compliance with government mandates. *Journal of Economic Behavior & Organization 185*, 688–701.

Roth, J. (2019). Pre-test with caution: event-study estimates after testing for parallel trends. *Working Paper*.

Roth, J. and P. H. C. Sant'Anna (2021). When Is Parallel Trends Sensitive to Functional Form? pp. 1–28.

Sandler, D. and R. Sandler (2014). Multiple event studies in public finance and labor economics: A simulation study with applications. *Journal of Economic and Social Measurement 39*, 31–57.

Silberzahn, R. and many authors (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science 1*(3), 337–356.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics In Press*.

U.S. Census Bureau, Population Division (2020). Annual Estimates of the Resident Population for Incorporated Places in the United States: April 1, 2010 to July 1, 2019 (SUB-IP-EST2019-ANNRES), Release Date: May 2020. `https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html`. Accessed: 2020-07-02.

Villas-Boas, S. B., J. Sears, M. Villas-Boas, and V. Villas-Boas (2020). Are we# stayinghome to flatten the curve? *conditionally accepted in the American Journal of Health Economics*.

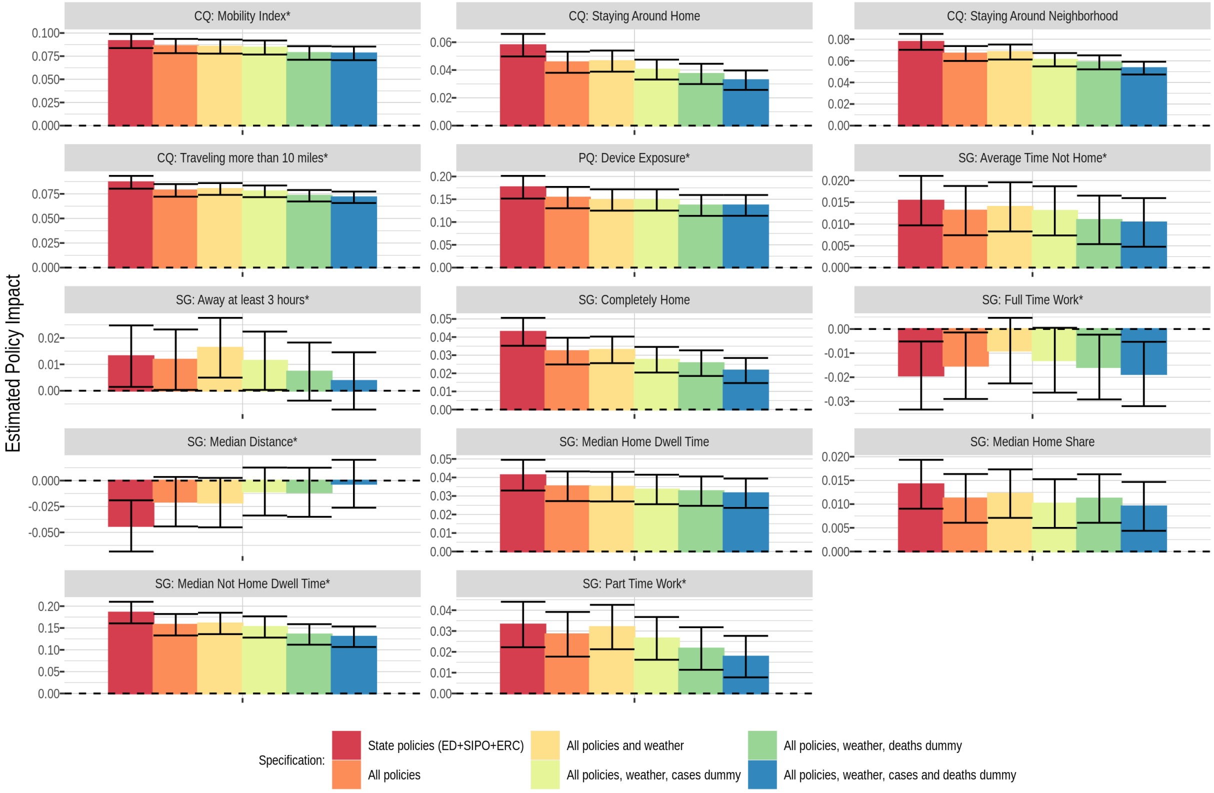
Weill, J. A., M. Stigler, O. Deschenes, and M. R. Springborn (2020). Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences 117*(33), 19658–19660.

A Online Appendix

This Appendix presents additional figures and tables that are described in the main body of the paper. All Appendix figures start with the letter ‘A’, followed by a number to distinguish them from the figures in the main body of the paper. The Appendix also contains a table with detailed information about the mobility variables used in the paper. Finally, we report some additional results based on the decomposition by Goodman-Bacon (2021).

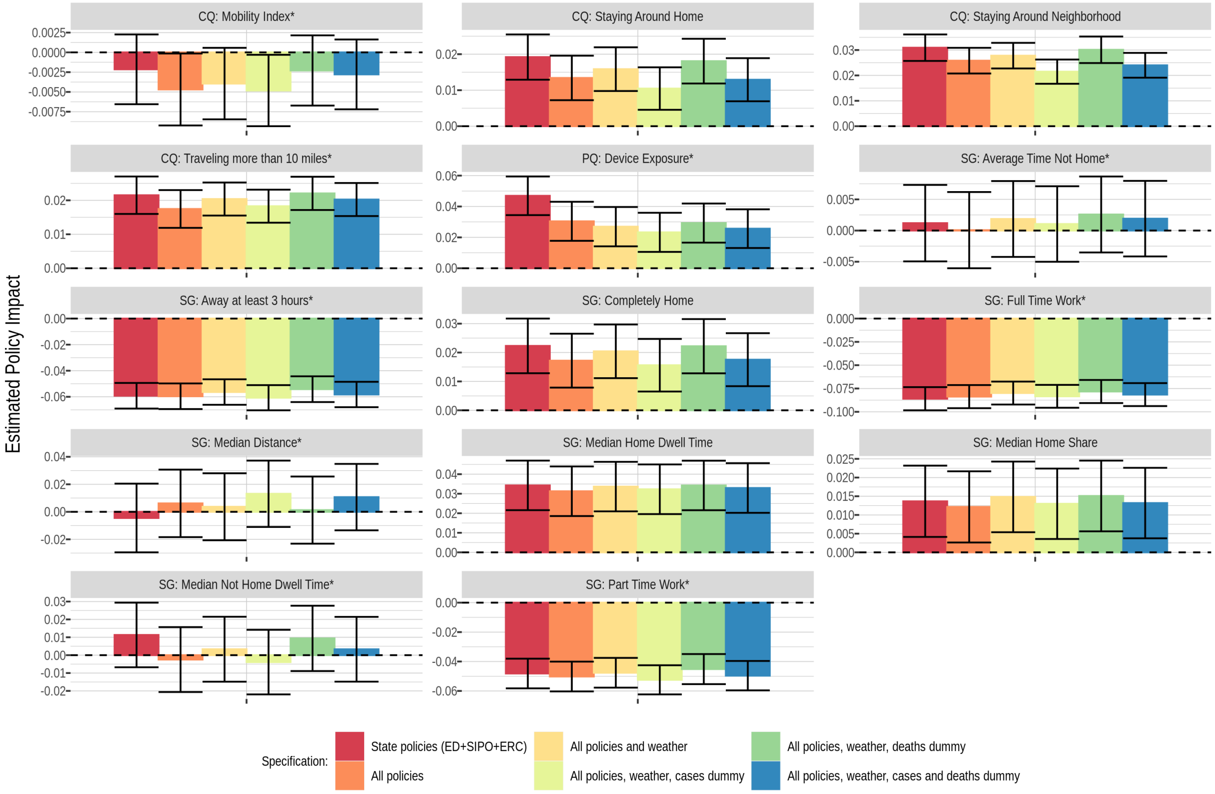
A.1 Additional estimates of the impacts of MRPs on mobility

Figure A.1: Estimated Impact of State SIPO on Various Log Mobility and Activity Outcomes



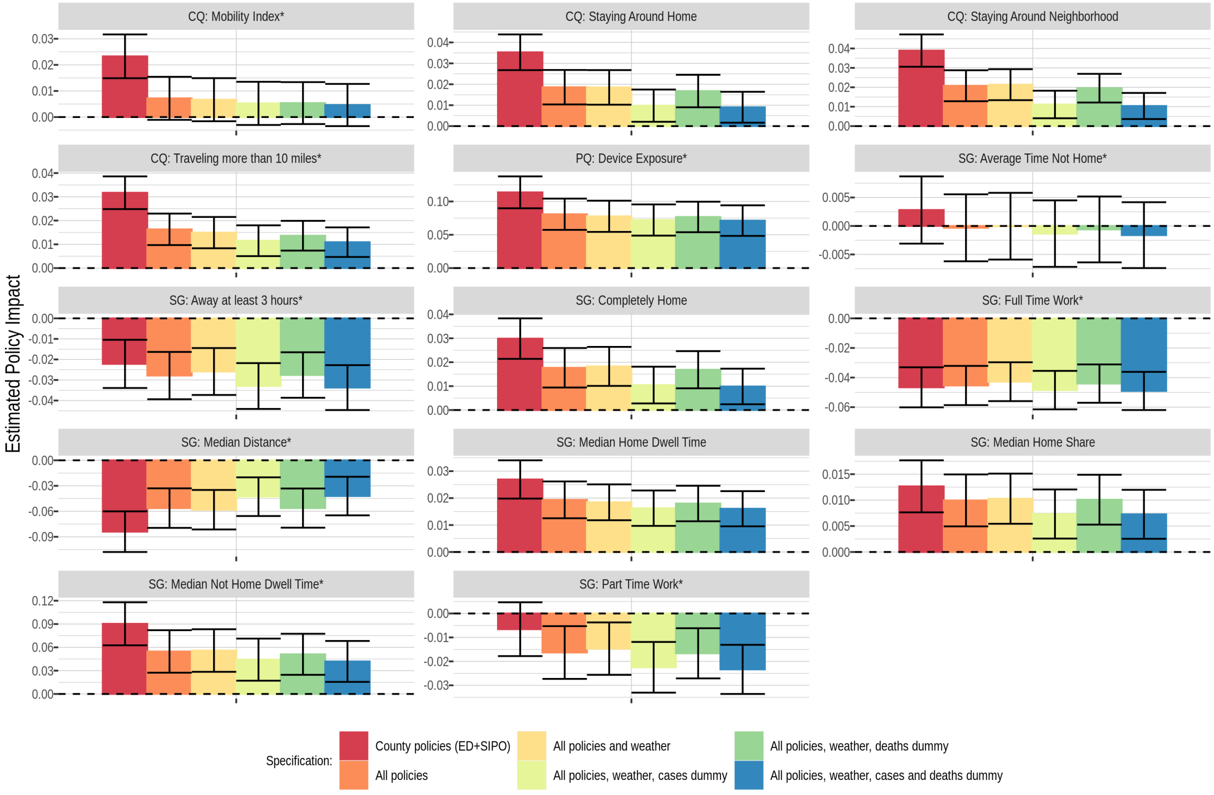
Notes: Figure A.1 reports the TWFE estimates of the impact of State SIPO on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.2: Estimated Impact of State ERC on Various Log Mobility and Activity Outcomes



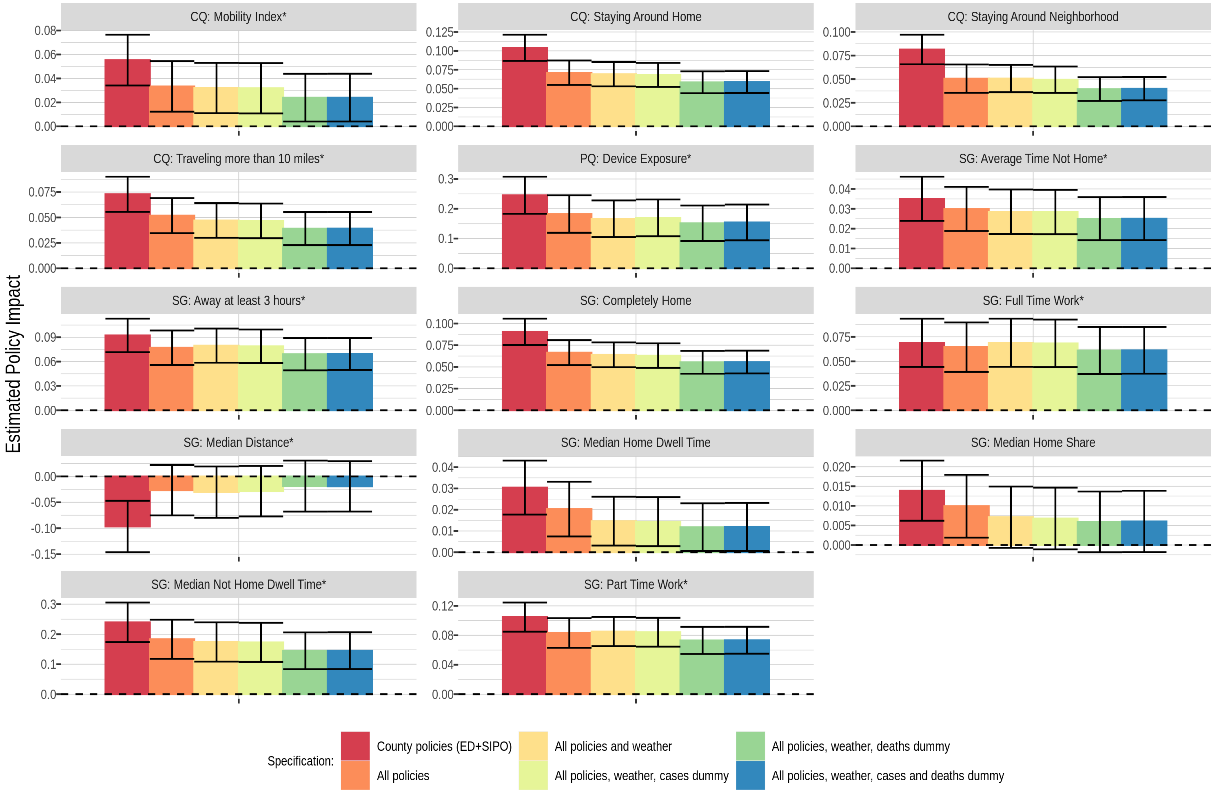
Notes: Figure A.2 reports the TWFE estimates of the impact of State ERC on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.3: Estimated Impact of County ED on Various Log Mobility and Activity Outcomes



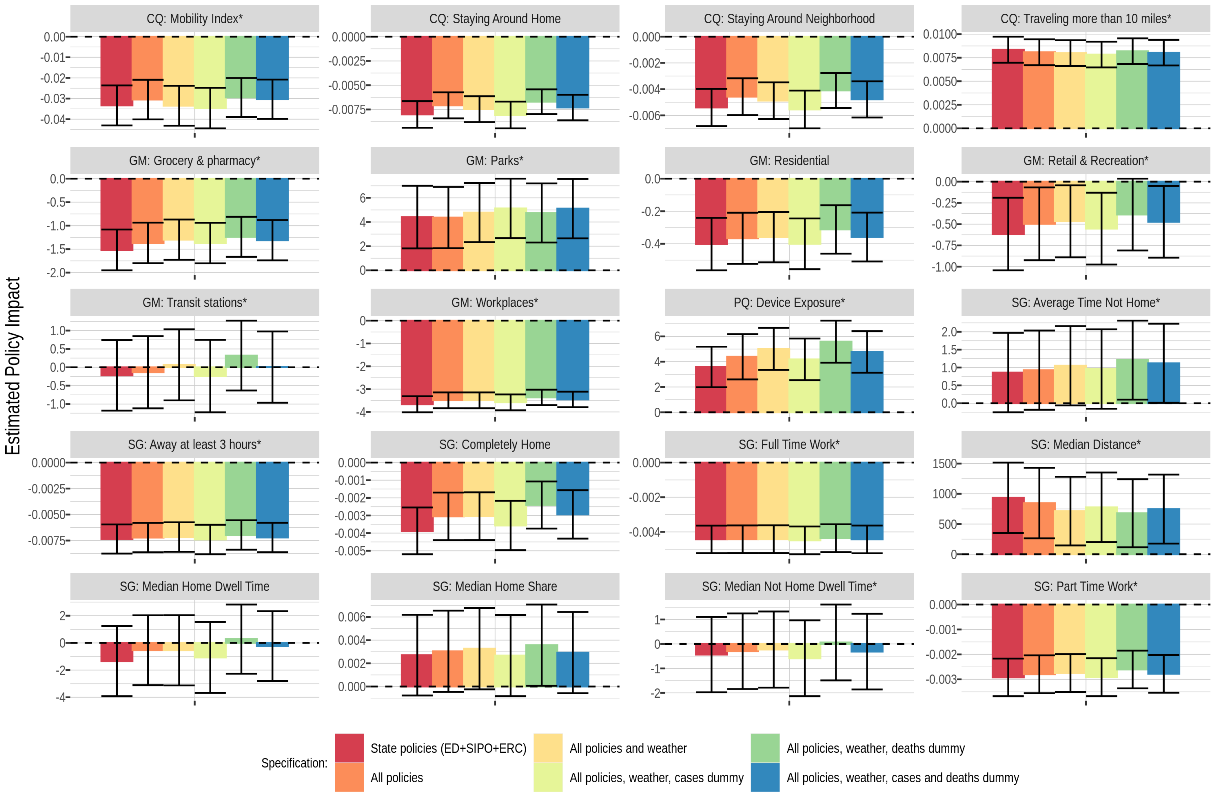
Notes: Figure A.3 reports the TWFE estimates of the impact of County ED on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.4: Estimated Impact of County SIPO on Various Log Mobility and Activity Outcomes



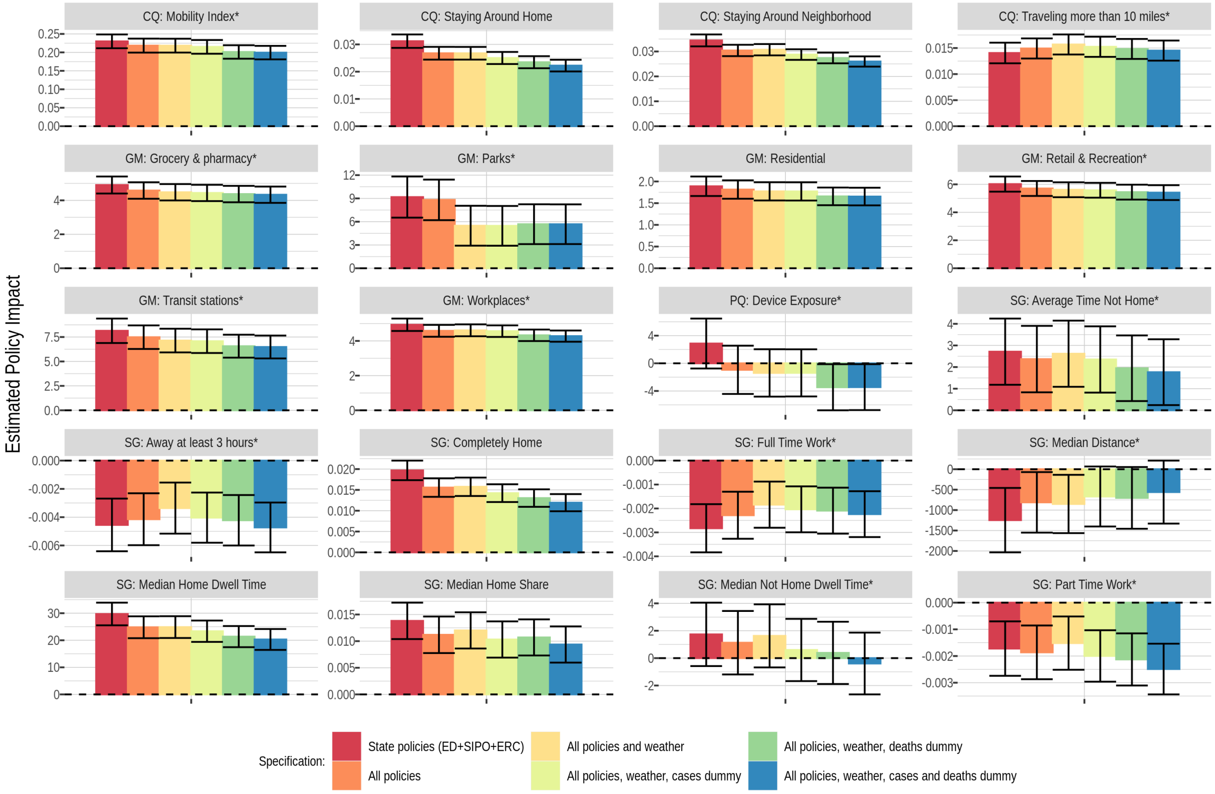
Notes: Figure A.4 reports the TWFE estimates of the impact of County SIPO on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.5: Estimated Impact of State ED on Mobility and Activity Outcomes (in levels)



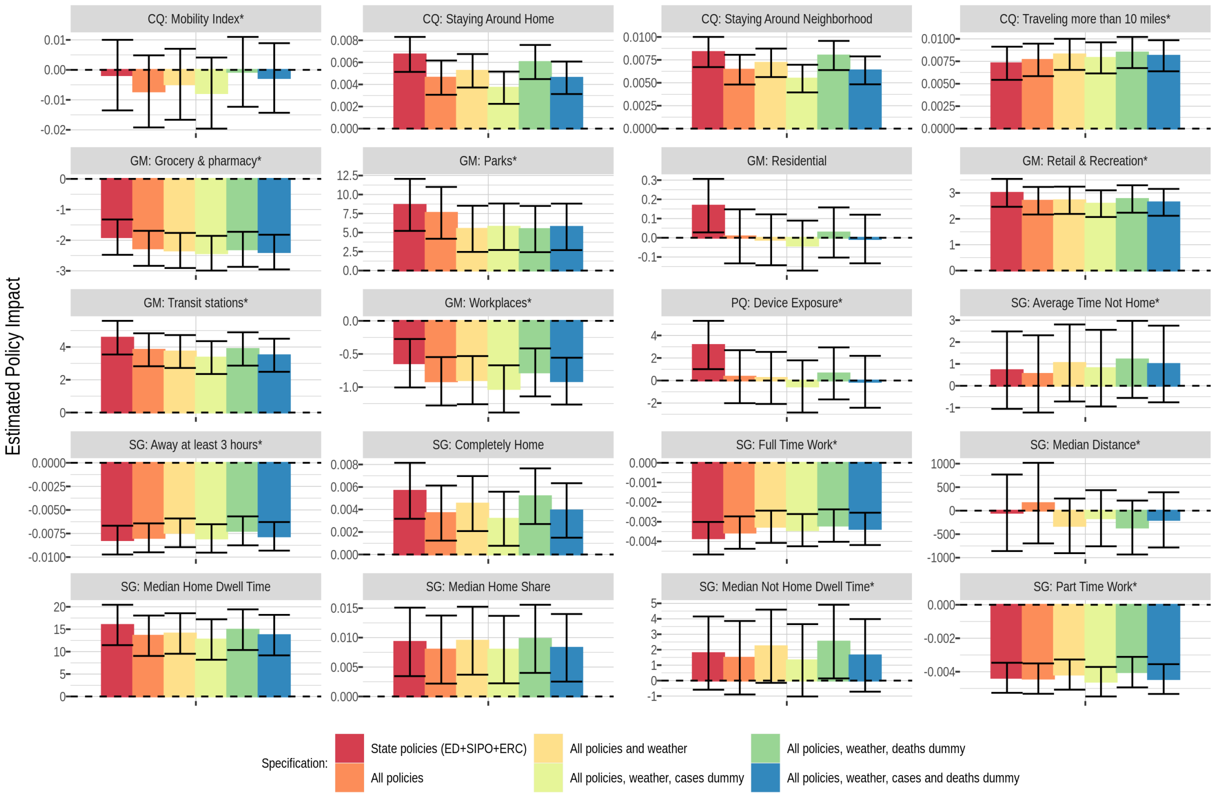
Notes: Figure A.5 reports the TWFE estimates of the impact of State ED on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.6: Estimated Impact of State SIPO on Mobility and Activity Outcomes (in levels)



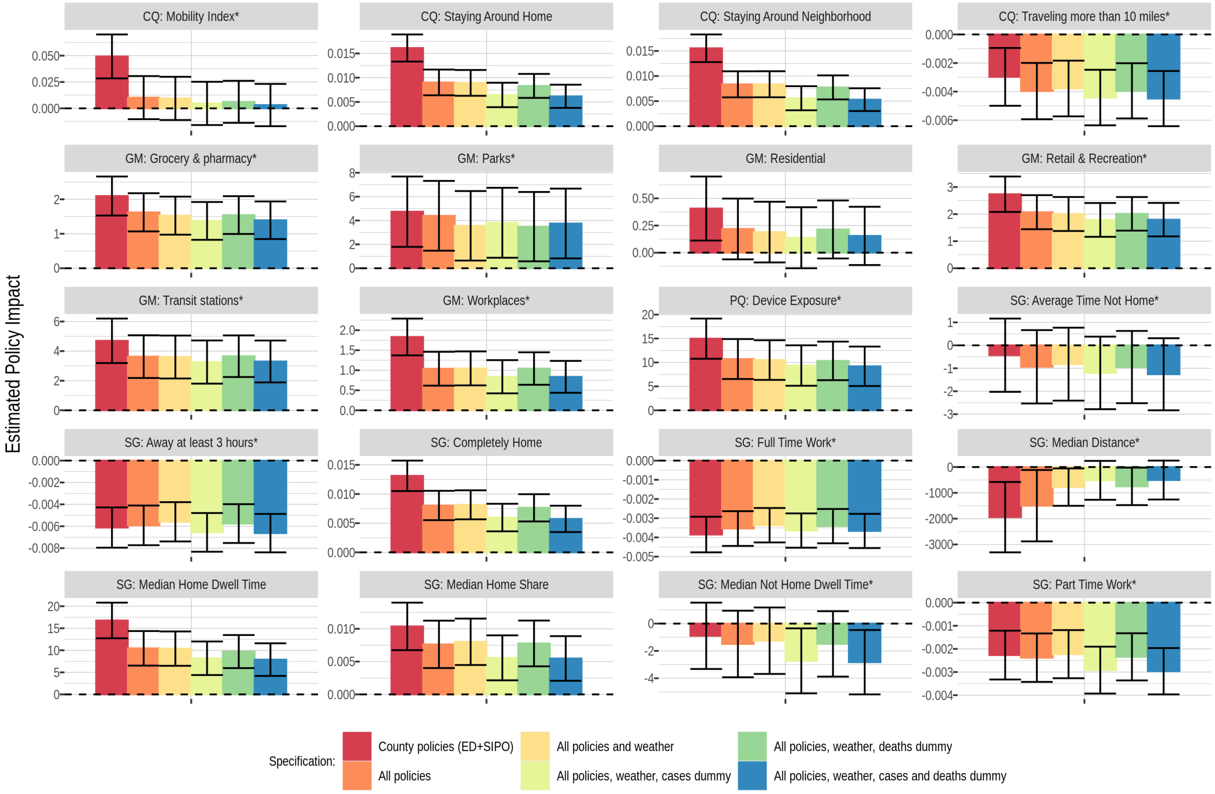
Notes: Figure A.6 reports the TWFE estimates of the impact of State SIPO on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.7: Estimated Impact of State ERC on Mobility and Activity Outcomes (in levels)



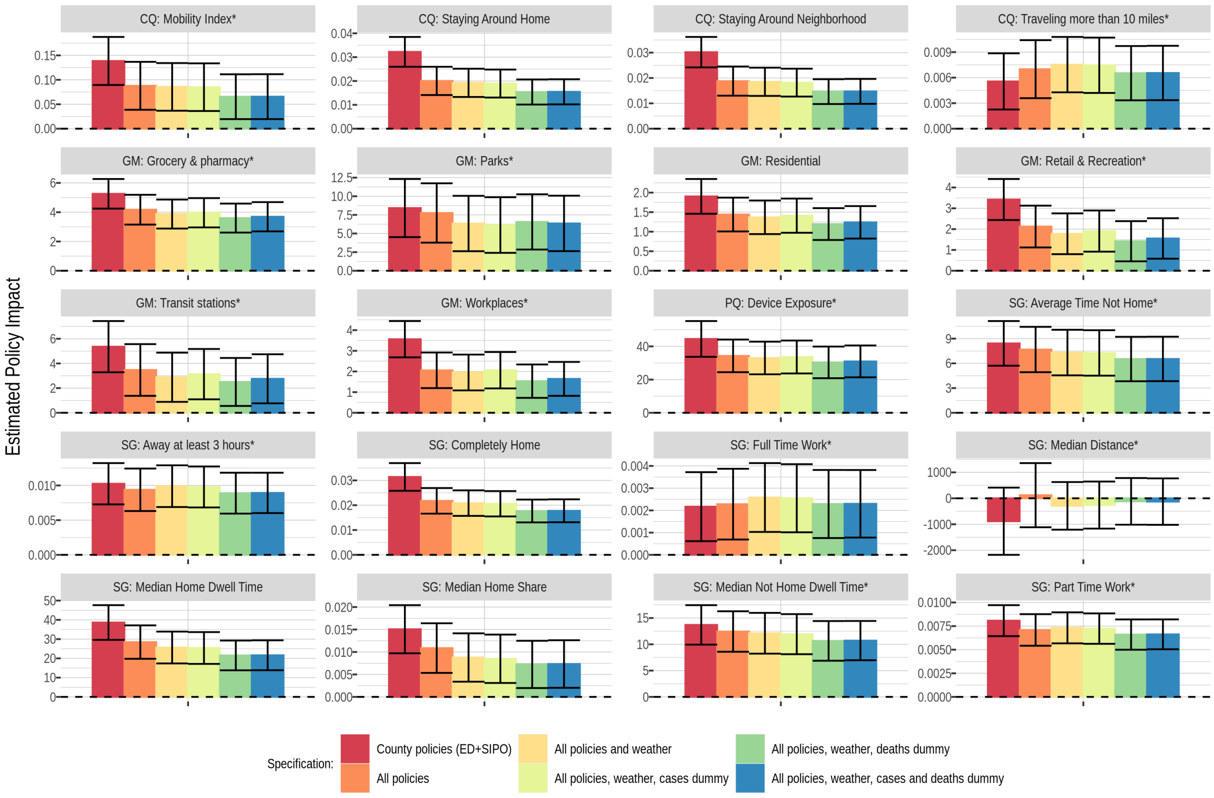
Notes: Figure A.7 reports the TWFE estimates of the impact of State ERC on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.8: Estimated Impact of County ED on Mobility and Activity Outcomes (in levels)



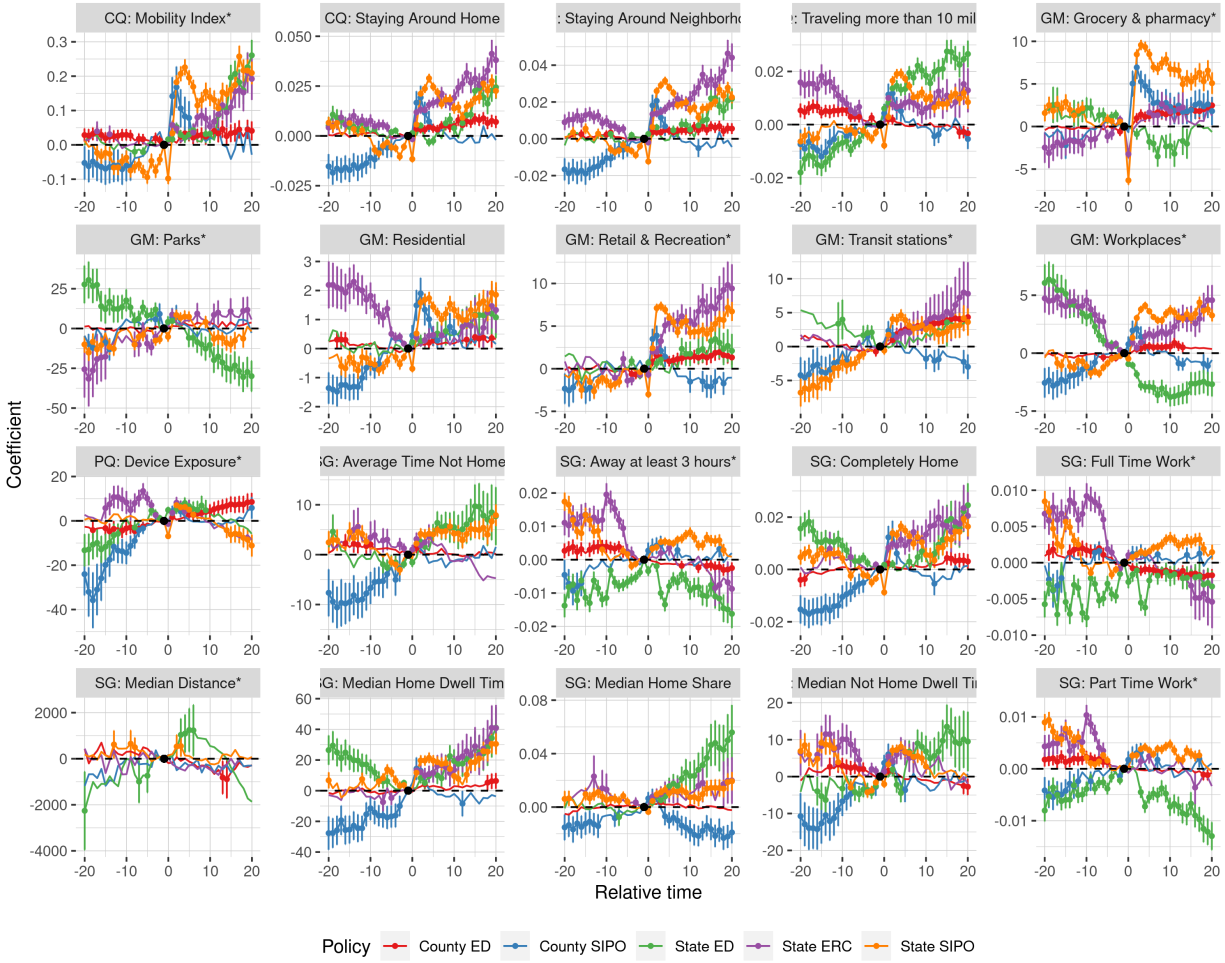
Notes: Figure A.8 reports the TWFE estimates of the impact of County ED on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.9: Estimated Impact of County SIPO on Mobility and Activity Outcomes (in levels)



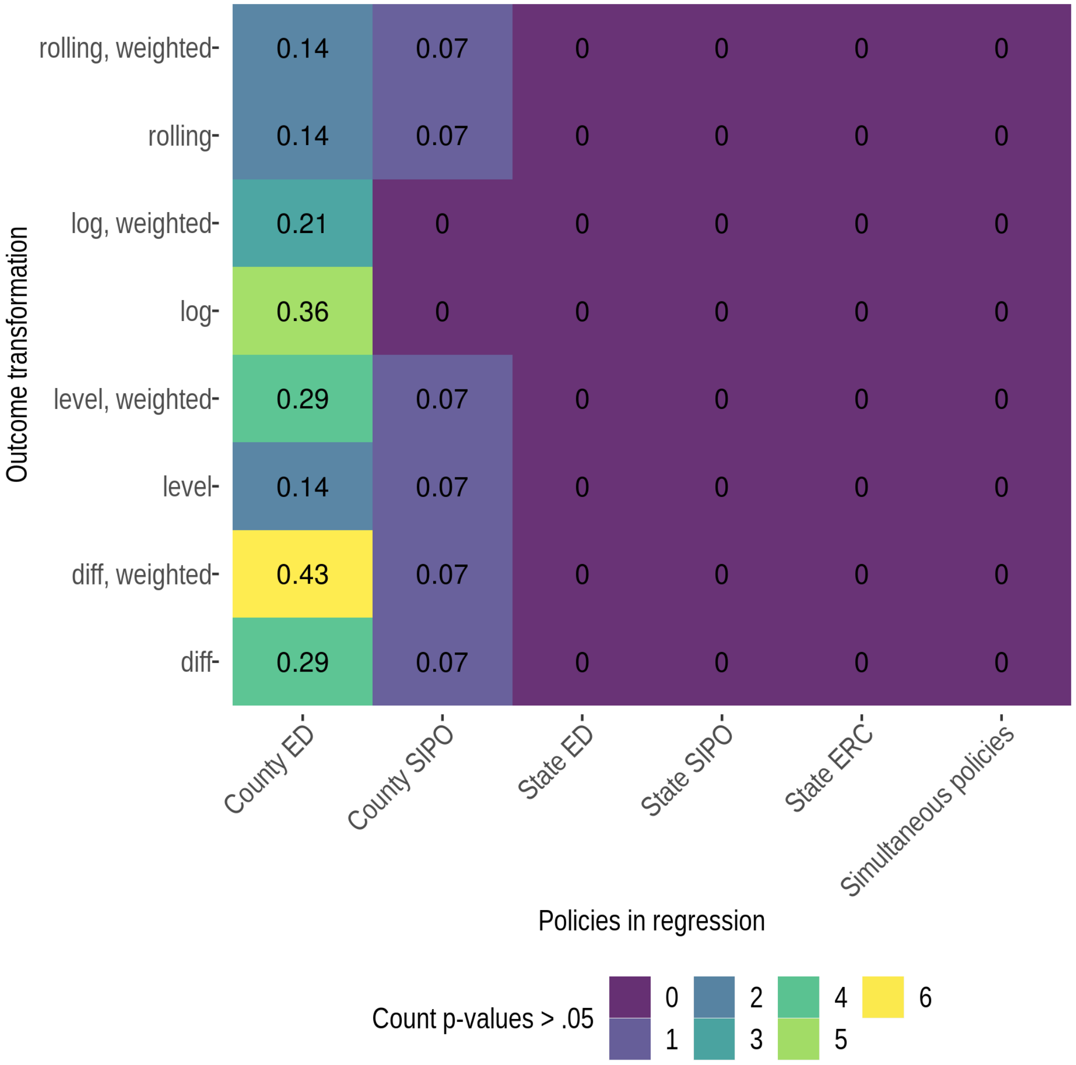
Notes: Figure A.9 reports the TWFE estimates of the impact of County SIPO on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.10: Estimated Event-Study Coefficients for Mobility and Activity Outcomes (in levels)



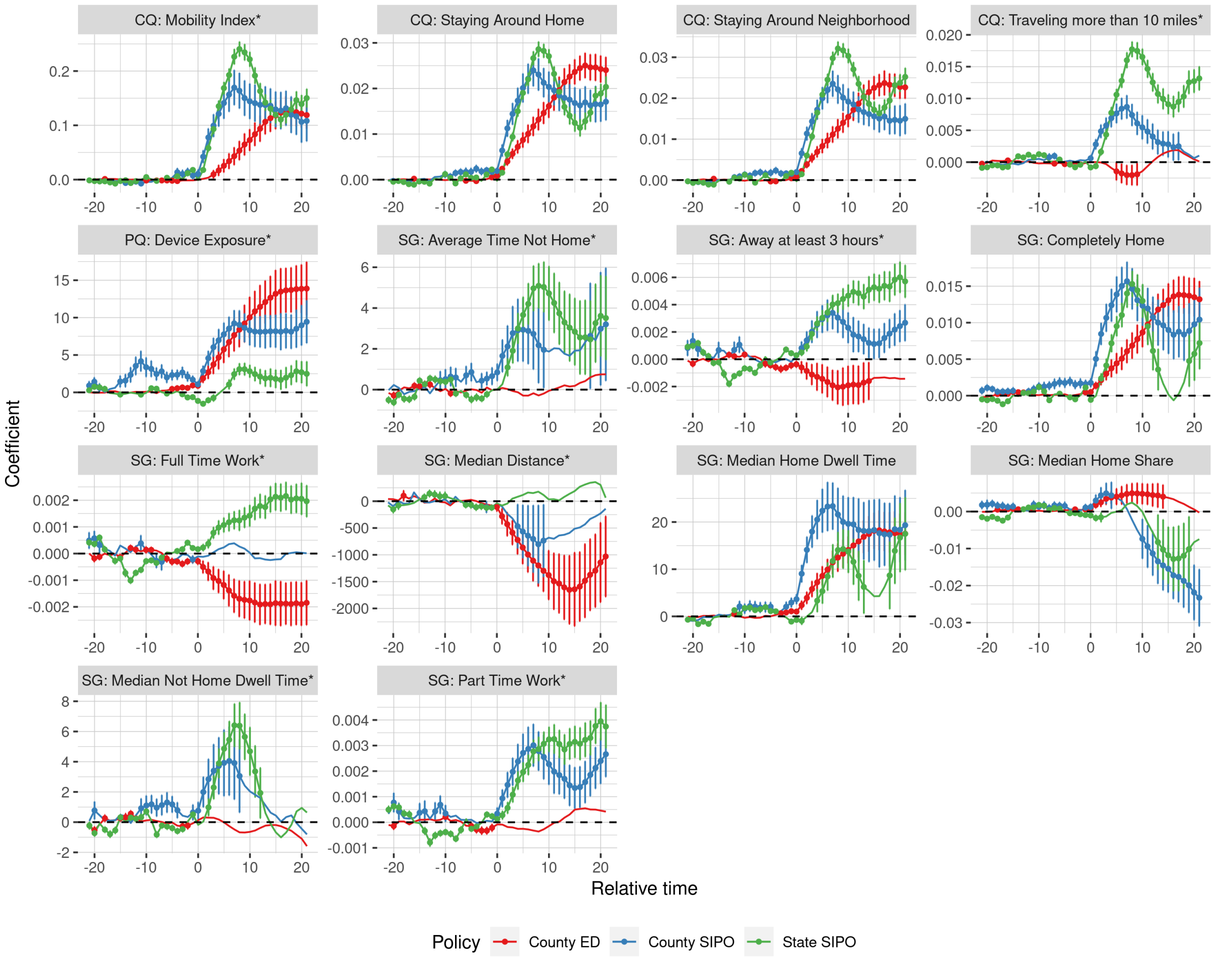
Notes: Figure A.10 reports the TWFE estimates of the event-study coefficients as specified in Equation (2). MRPs under consideration are represented by the color of each line. The whiskers show the 95% confidence intervals based on standard errors clustered at the county-level.

Figure A.11: Heatmap of F-tests for Pre-Trend Tests in Event-Study Regressions



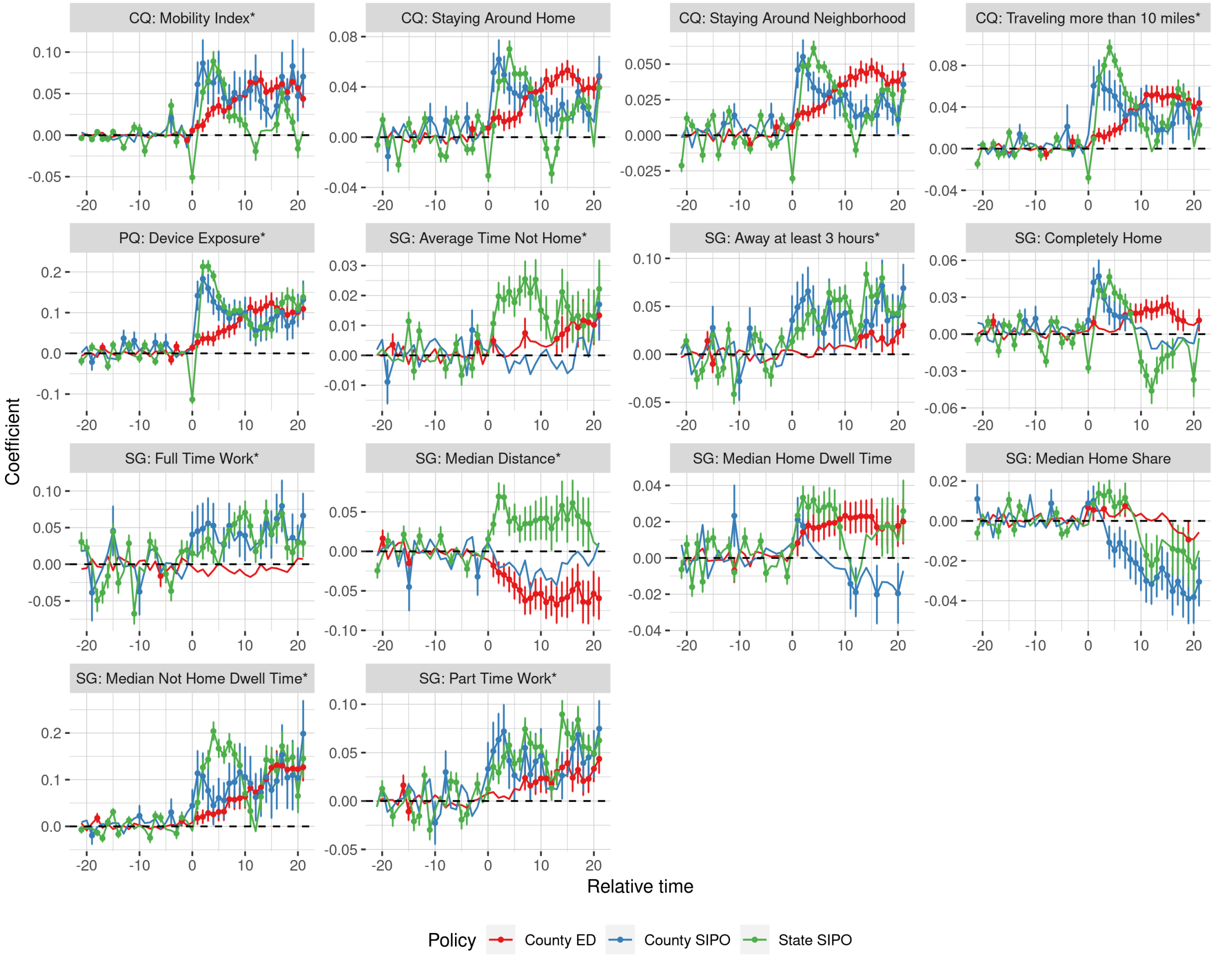
Notes: Figure A.11 is a heat map based on the p-values from F-tests testing the null hypothesis that the pre-MRP-adoption coefficients are jointly equal to zero in Event-study regressions. Each cell summarizes the results of 14 F-tests on pre-trends coefficients, corresponding to the 14 non-negative mobility outcomes from Safegraph, Cuebiq, and PlaceIQ (Google Mobility outcomes are ignored here since we cannot apply the log transformation on them). The number in each cell corresponds to the fraction of tests (out of 14) where the p-value on the null hypothesis exceeds 0.05, indicating that the null hypothesis would be rejected at the usual 5% significance level. The rows indicate the transformation of the dependent variable while the columns indicate the MRP considered (estimated individually in columns 1-5, and estimated jointly in column 6).

Figure A.12: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes Transformed with a 7-Day Moving Average



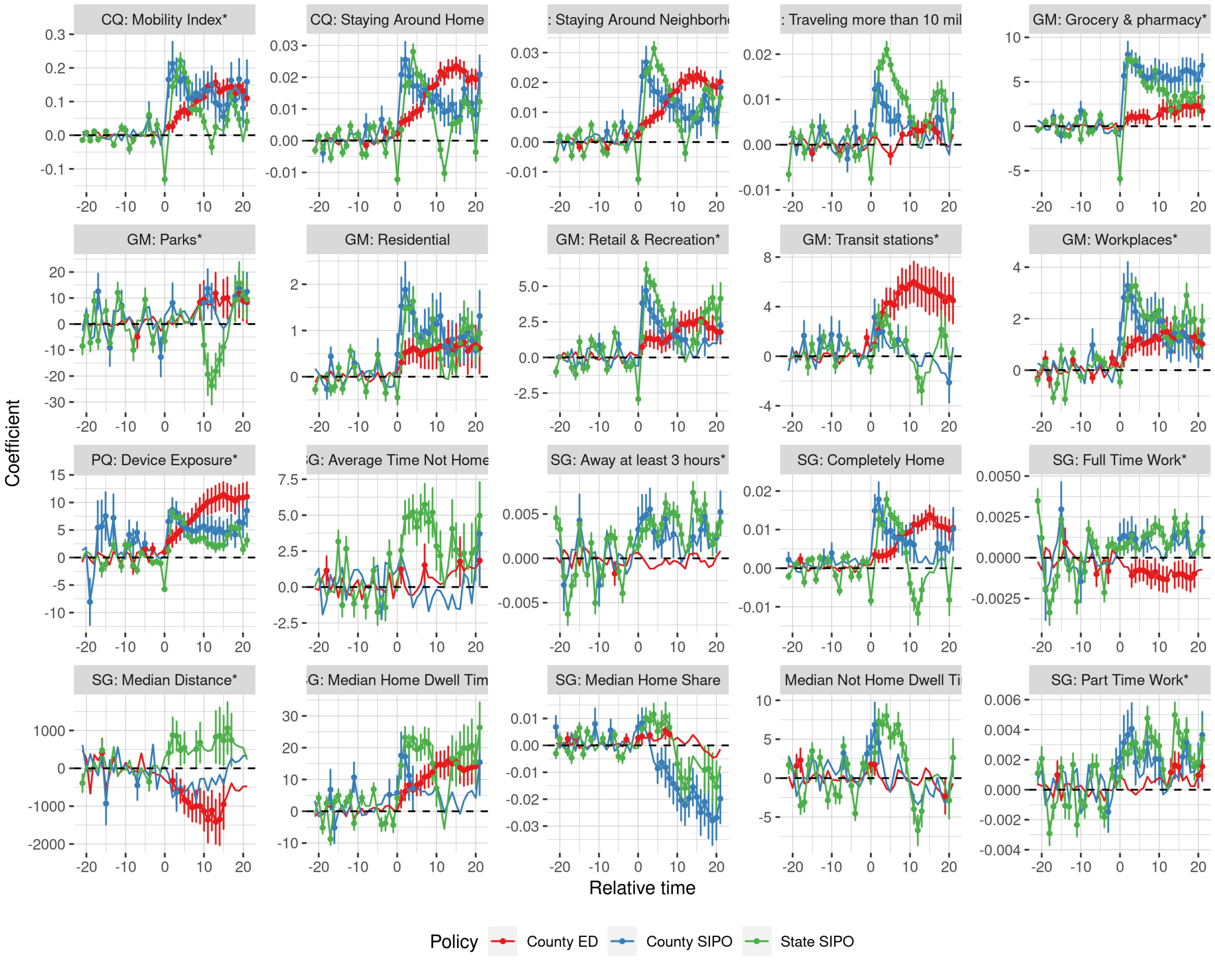
Notes: Figure A.12 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure A.13: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes in Logs



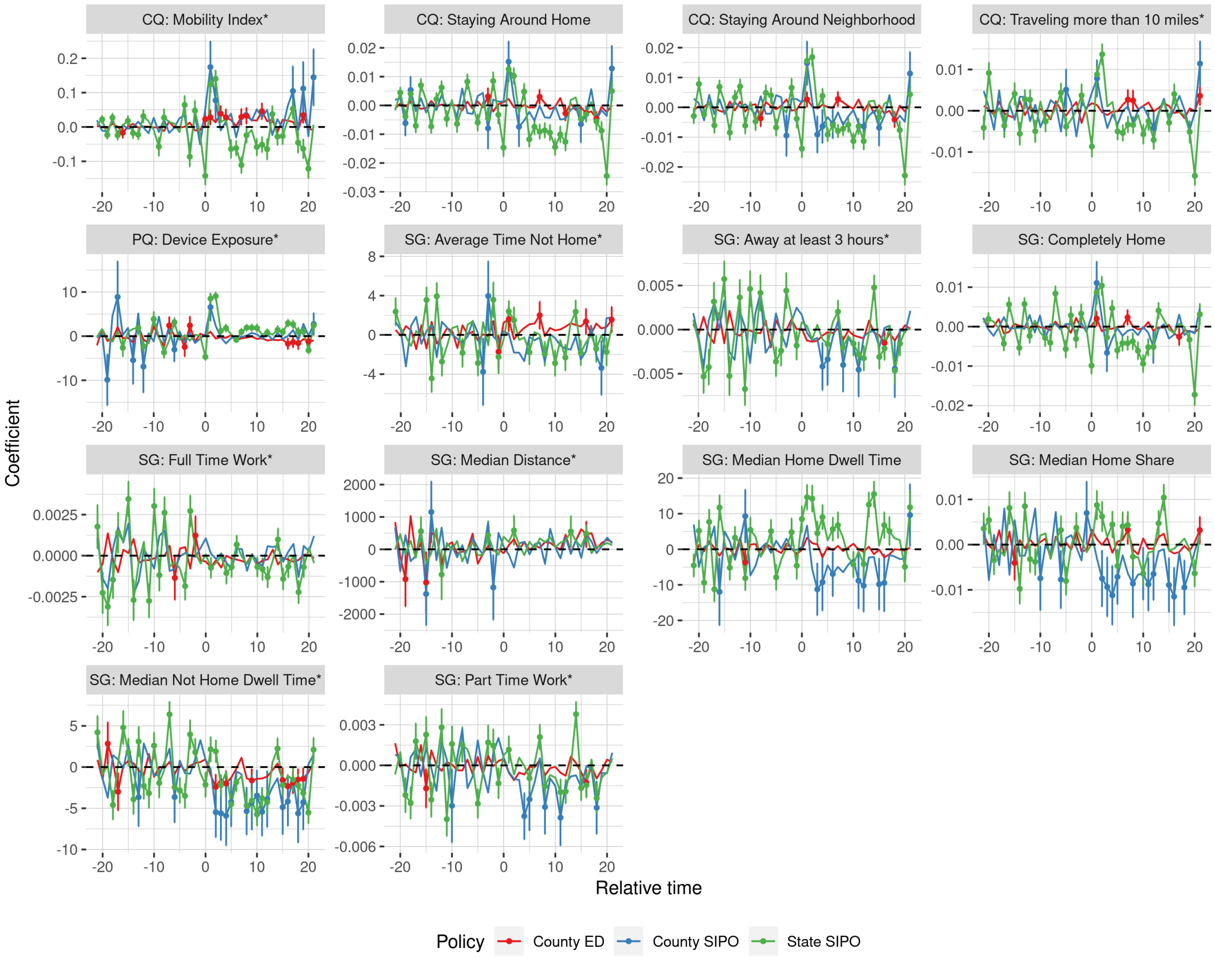
Notes: Figure A.13 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure A.14: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes Transformed with a First-Difference



Notes: Figure A.14 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure A.15: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes in Levels



Notes: Figure A.15 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

A.2 Variables description

Table A.1: Details about the Mobility Indicators Used in the Paper

Variable	Source	Description	Units
Cuebiq Mobility Index	Cuebiq	Calculated using a derivative factor indicating the distance between opposite corners of a box drawn around the locations observed for users on each day. The index for each county is the median of the aggregated movements of all users within a county. Values can be interpreted as: 5 - 100 km, 4 - 10 km, 3 - 1 km,2 - 100m,1 - 10m. An index of 2.5 for a county, would mean the median user in that county is traveling 250m	No unit
Staying Around Home	Cuebiq	Percentage of users staying at home in any given state/county. It is calculated by measuring how many users moved less than 330 feet from home	%
Staying Around Neighborhood	Cuebiq	Percentage of users traveling less than one mile from home	%
Traveling more than ten miles	Cuebiq	Percentage of users traveling more than ten miles from home	%
Grocery and Pharmacy	Google Mobility	Changes in the number of visits to grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies, relative to baseline. The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020	%
Parks	Google Mobility	Changes in the number of visits to places like national parks,public beaches, marinas, dog parks, plazas, and public gardens, relative to baseline.	%
Retail and Recreation	Google Mobility	Changes in the number of visits to places like restaurants,cafes, shopping centers, theme parks,museums, libraries, and movie theaters, relative to baseline	%
Residential	Google Mobility	Changes in the number of visits to places of residence relative to baseline.	%
Transit Stations	Google Mobility	Changes in the number of visits places like public transport hubs such as subway, bus, and train stations, relative to baseline	%

Table A.1: Details about the Mobility Indicators Used in the Paper

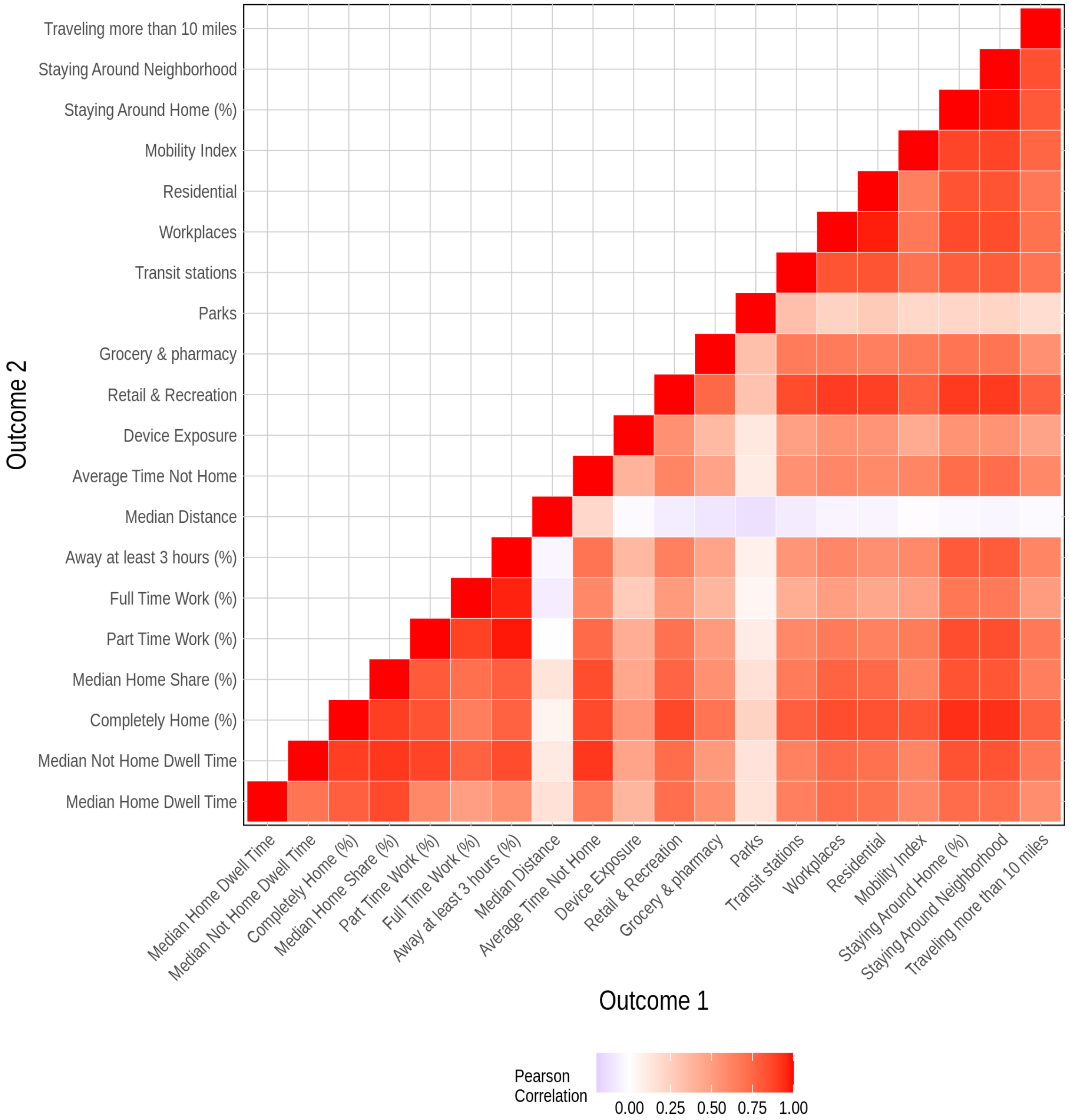
Variable	Source	Description	Units
Workplaces	Google Mobility	Changes in the number of visits to places of work relative to baseline	%
Average Time Not Home	Safegraph	Average time a device is recorded as away from home within a county, reconstructed from ”bucketed” time away from home	min-utes
Full Time Work	Safegraph	County share of devices that spent greater than 6 hours at a location other than their home geohash-7 during the period of 8 am - 6 pm in local time	%
Median Distance	Safegraph	Median distance traveled from the geohash-7 of the home by the devices during the time period (excluding any distances of 0). The median is provided at the census block group level and we take the weighted mean of these medians at the county level (using the number of devices as weights)	me-ters
Median Not Home Dwell Time	Safegraph	Median dwell time at places outside of geohash-7 home for all observed devices during the time period. For each device, the observed minutes outside of home across the day (whether or not these were contiguous) are summed to get the total minutes for each device. The median is first calculated across all these devices at the census block group level, and then at the county level	min-utes
Part Time Work	Safegraph	County share of devices that spent one period of between 3 and 6 hours at one location other than their geohash-7 home during the period of 8 am - 6 pm in local time. This does not include any device that spent 6 or more hours at a location other than home	%
Away at least three hours	Safegraph	County share of devices that spent at least 3 hours at one location other than their geohash-7 home during the period of 8 am - 6 pm in local time	%
Completely Home	Safegraph	County share of devices which did not leave the geohash-7 in which their home is located during the time period.	%
Median Home Dwell Time	Safegraph	Median dwell time at home geohash-7 (”home”) in minutes for all devices during the time period. For each device, the observed minutes at home across the day are summed (whether or not these were contiguous) to get the total minutes for each device. The median is provided at the census block group level and we take the weighted mean of these medians at the county level	min-utes

Table A.1: Details about the Mobility Indicators Used in the Paper

Variable	Source	Description	Units
Median Home Share	Safegraph	Median percentage of time we observed devices home versus observed at all during the time period. The median is provided at the census block group level and we take the weighted mean of these medians at the county level	%
Device Exposure	PlaceIQ	For a device, number of distinct devices that also visited any of the commercial venues that this device visited that day. The county-level DEX reports the county-level average of this number across all devices residing in the county that day. We use the adjusted version of this variable, which accounts for devices not seen living their home (further details provided by the authors.	Count

A.3 Correlation between mobility measures

Figure A.16: Pearson correlations between mobility outcome variables



A.4 Results from Bacon-Goodman decomposition

The recent econometric literature highlights that two-way fixed effect estimates do not generally identify an average treatment effect when treatment is staggered over time (de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021). In particular, Goodman-Bacon (2021) shows that the two-way fixed effect estimator can be decomposed as a weighted sum of all possible 2x2 difference-in-differences estimators generated by a staggered policy adoption design. Formally, let $k = 1...K$ be different groups of units ordered by the time at which they receive a binary treatment, and U is a group that never receives treatment. Let $\bar{y}_b^{POST(a)}$ denote the sample mean of y_{it} in group b during group a post period, and define $\bar{y}_b^{PRE(a)}$ similarly. Finally, let $\bar{y}_a^{MID(a,b)}$ be the sample average of units in group a after group a becomes treated, but before group b becomes treated. Then Theorem 1 of Goodman-Bacon (2021) shows that in the two-way fixed effect model with a unique binary policy and no covariates, the DD estimator can be written as:

$$\widehat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\beta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{\ell > k} \left[s_{k\ell}^{k} \widehat{\beta}_{k\ell}^{2x2,k} + s_{k\ell}^{\ell} \widehat{\beta}_{k\ell}^{2x2,\ell} \right]$$

where the 2 × 2 diff-in-diffs estimators are:

$$\begin{aligned}
\widehat{\beta}_{kU}^{2x2} &\equiv \left(\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)} \right) - \left(\bar{y}_U^{POST(j)} - \bar{y}_U^{PRE(j)} \right) \\
\widehat{\beta}_{k\ell}^{2x2,k} &\equiv \left(\bar{y}_k^{MID(k,\ell)} - \bar{y}_k^{PRE(k)} \right) - \left(\bar{y}_\ell^{MID(k,\ell)} - \bar{y}_\ell^{PRE(k)} \right) \\
\widehat{\beta}_{k\ell}^{2x2,\ell} &\equiv \left(\bar{y}_\ell^{POST(\ell)} - \bar{y}_\ell^{MID(k,\ell)} \right) - \left(\bar{y}_k^{POST(\ell)} - \bar{y}_k^{MID(k,\ell)} \right)
\end{aligned}$$

and the weights s_{kj} are functions of group sizes and variance of treatment within groups.

This decomposition highlights two important potential pitfalls of the TWFE estimator. First, due to the weights, any amount of treatment effect heterogeneity between groups means that the $\beta^{\hat{DD}}$ estimator can be quite different from the average treatment effect on the treated. Furthermore, the $\hat{\beta}^{DD}$ estimator implicitly compares units that are treated early to units that are treated "later" (the $\hat{\beta}_{kl}^{2x2,\ell}$ components). Thus if the treatment effect is time-varying, these comparisons will not be informative about the (time-varying) average treatment effect on the treated.

To assess whether identification issues related to treatment effect heterogeneity in bias our baseline TWFE estimates, we first estimate a modified version of (1) above including each policy in isolation and no covariates. We then implement the Bacon-Goodman decomposition above, but remove all $\hat{\beta}_{kl}^{2x2,\ell}$ components and normalize the remaining weights so that they sum to one. We finally compute the ratios of these two estimators: the standard TWFE and the adjusted TWFE that excludes the "treated earlier vs. later" comparisons.

$$\frac{\sum_{k \neq U} \widetilde{s}_{kU} \widehat{\beta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{\ell > k} \left[\widetilde{s}_{k\ell}^{k} \widehat{\beta}_{k\ell}^{2x2,k} \right]}{(\sum_i X_i' X_i)^{-1} (\sum_i X_i' Y_i)}$$

Figure A.17: Ratio of weighted coefficients with ”Later vs Earlier” comparisons removed to regular 2-way fixed effects coefficients

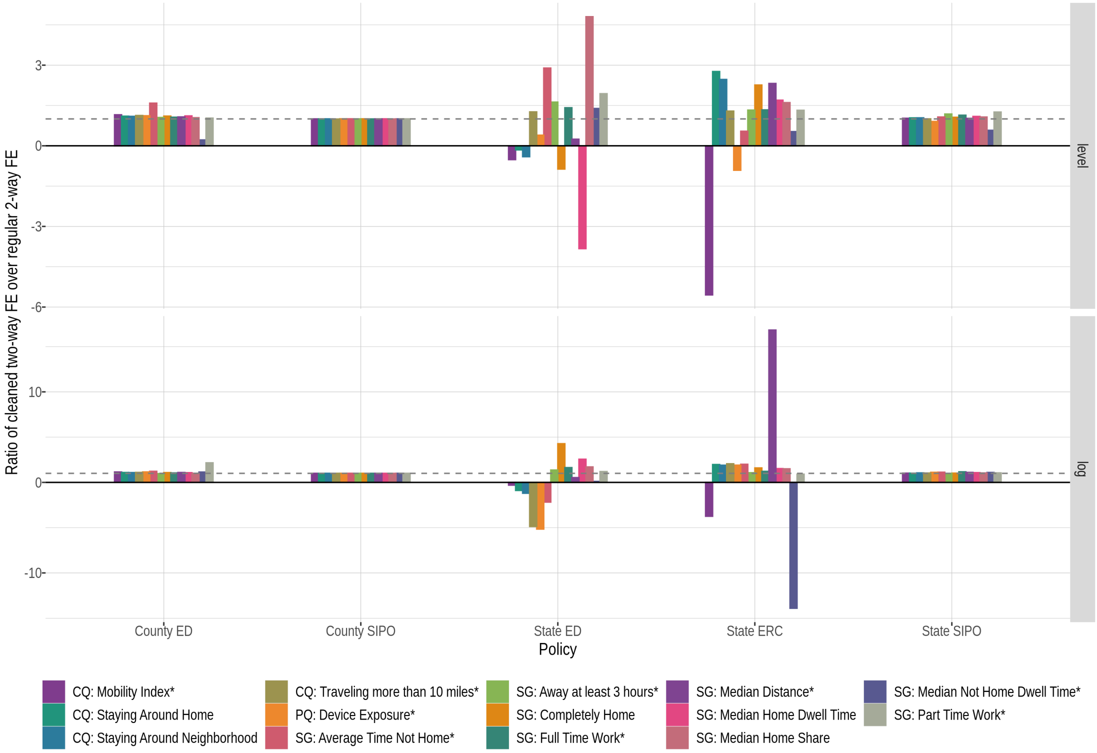


Figure A.17 presents the ratios for 14 outcome variables in levels and in logs (Google Mobility outcomes are excluded since they include a lot of zeros). We note that for the county policies and the State Shelter-in-place orders, the ratios are very close to one for most outcomes: removing the “treated later vs treated earlier” comparisons does not substantially impact our estimates. However, for the State ED and State ERC policies the discrepancy can be very large, with some estimates having opposite signs depending on the outcome variable considered. This result is somewhat expected from the Bacon Goodman decompositions: for State ED and State ERC there is no “true” control group to estimate the impact of the policy. All units are subject to a state ED and ERC at some point. As a result, removing the “treated later vs treated earlier” comparisons can have a large impact on the overall estimate. These discrepancies lead us to suspect dynamic treatment effects, which we further investigate with event studies in section 5.