Peer Review File

Manuscript Title: The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic

Redactions – Third Party Material

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

A. Summary of key results

The paper presents evidence from small area panel datasets for China, South Korea, Iran, Italy, France and the USA of how the uninhibited exponential daily growth rate of COVID-19 cases in the initial phase of disease progression has been mitigated by the collective impact of the anti-contagion policies adopted in each of these settings. The headline finding is that current policies have prevented or delayed around eighty million infections. This estimate is calculated by comparing actual trends in growth rates with those predicted from the models with all of the policy variables switched off.

B. Originality and significance

The manuscript addresses a question of urgent and substantial significance for global public health and the world economy. It provides evidence from direct measurements of retrospective observational data to complement evidence from epidemiological numerical simulations.

C. Data and methods

Epidemiological data is not my specialism but sources, imputations and assumptions are stated clearly in the Appendix and appear reasonable. A reassuring validation exercise against the Johns Hopkins data is provided.

p.17 notes that the school closure variable is always on (=1) for the South Korea dataset, so that it is effectively dropped from the model (submerged into the constant term).

D. Appropriate use of statistics and treatment of uncertainties

The method adopted is to use separate country (two way fixed effect) panel data regressions for the daily difference in the natural logarithm of infected cases as a function of a set of policy dummies. The fixed effects include time (t), sub-national units (i), day of the week, and dummies for changes in testing regime. Identification of the policy impacts therefore comes from differences across the sub-national units in the variation within these sub-national units over time and for a given day of the week and testing regime (i.e. a differences in differences strategy).

The paper argues that the method adopted accounts for:
- differences in baseline growth rates
- systematic patterns unrelated to policy
- is robust to systematic under-surveillance of cases
- accounts for changes in procedures to diagnose positive cases

The authors should be clear to themselves and to their audience that the definition of z as “exogenous changes in independent policy variables” is not an inherent characteristic of these variables but a crucial identification assumption that they are adopting to allow the association between these variables and the outcome to be attributed to the causal effect of those policies. To assume that the policy being switched on is exogenous and independent is to assume that it is effectively randomized across areas and days. For example it could, in principle, be problematic if the timing of policy adoption was influenced by past changes in the local caseload or by the anticipation of future changes in cases.

p.11 I do not follow the logic of the claim that "usefully equation 2 does not depend on w". From equation 1 it is in general true that the derivative δf(.., w)/δx can be a function of w and it may be the case that these controls may influence the mediator and hence δx/δz may also depend on w. This need not be a problem so long as the confounding effect of w is captured by the fixed effects or any controls are included directly in the regression. Basically the assumption of no unobserved confounders is being made, conditional on the fixed effects (for sub-national units, day effects and changes in testing regime).

p.13 It could be noted that the linear specification in (7) assumes - for simplicity and identification - that contribution of individual policies to growth rate is additive.

I note that a log-linear specification is being used. Predictions of the growth rate based on this model are fine but if were to be used to predict the number of cases (I) then the retransformation problem would have to be taken into account ie. predictions would be a function of exp(ε).

p.14 The authors should be clear that their method should be robust only to time-invariant under-reporting. This appears to be how they interpret "systematic" under surveillance.

p.14 The discussion and rationale for using weekly lags in China is a bit opaque and could be more explicit. It is justified by the incubation period. Wouldn't this apply to other countries as well.

I would urge the authors to include the full regression results in the Appendix. These are currently lacking.

Note that Figure A1 plots the estimated residuals not the (unobserved) errors. These are non-normal but normality is not a requirement of the modeling and inference strategy that is adopted.

Figure A2 would be clearer if the legend for "observations in our data" also appeared in the panel for Hubei.

E. Conclusions: robustness, validity, reliability

A key policy outcome is cases that require hospitalization (or more specifically intensive care) central to the policy of flattening the curve. I wonder if more could be said about this based on the findings of the analysis?

F. Suggested improvements

Going beyond the publication of the current manuscript the authors should be encouraged in their broader project that can be continuously updated with more countries (such as Hong Kong, Singapore, Taiwan, Spain, Germany, UK) and additional time series observations as they become available.

G. References

This is not my field of expertise but the references appear comprehensive. For readers' benefit the incomplete references should be finalized: referenes 2, 3, 4, 13, 14, 17, 22, 26.

H. Clarity and context

The manuscript is presented in a clear and lucid way and brings home the impact of the issue and the findings. As a non-epidemiologist I found the text accessible.

Referee #2 (Remarks to the Author):

I enjoyed the opportunity to read MS 2020-03-04486A: "The Effect of Large-Scale Anti-Contagion Policies on the Coronavirus (COVID-

19) Pan- demic" by Professor Hsiang and co-authors.
Positives
This is a terrific paper and I think it should be published in a top, general interest science journal. My reasons for this are:
1. I don't think i have ever refereed a paper with greater potential impact on human welfare. Nor where the question at hand was such an urgent one time-wise.
2. I very much appreciate the authors' reduced form approach as a com- plement to calibrated epidemiological models. They position their paper very nicely thus.
3. The analysis is clearly described, pedagogical, and well motivated.
4. The data collection effort on such a tight timeframe is nothing short of amazing
5. Despite the above, neither the analysis nor prose appear rushed. It is a careful exercise.
Bravo!
Negative
My main problem is that the authors do not develop any argument for why the anti-contagion policies1 examined are exogenous. Yes, in general, best- practice reduced form methods are great at taking us to causal inference with observational data, which is terrific. But unless i have missed it (i am sorry if so), the authors skip the argument as to why the conditional independence assumption required for causal inference is satisfied in their context.
Meyer [1995] noted: "If one cannot experimentally control the variation one is using, one should understand its source". I see no effort by the authors to explain why some regions adopted some policies and not others, and the timing of these policies. Without this, it is difficult to know whether policies were adopted using information about COVID infection rate trends or future infection rates that is unavailable to the authors.
1 Super minor but dropping Chinese cities where "we cannot find a policy deployment date" seems necessary but should note it could also potentially impart some LATE or bias issues.

The authors draw a parallel in the abstract to existing empirical work on policies that affect economic growth: "We then apply reduced-form econo- metric methods, commonly used to measure the effect of policies on economic growth". In studying volcanic eruptions, Proctor et al. [2018] uses a treat- ment whose timing is certainly exogenous to agricultural yields, so there's not much of a leap required compared to this manuscript. Returning squarely to economic growth, the exogenous case is again much easier to make with temperature and cyclone shocks [Hsiang, 2010].
Fundamentally, why do different places adopt different policies? For ex- ample, if the variation is because it's just really difficult to know impacts and policy makers were just guessing at what might work, that would be a good story to develop for the authors. But I see no case made along these lines. Nor even any assessment of how allowing for different pre-policy trends matters or doesn't matter.
Given the above, i suggest one of two avenues:
1. Develop the argument for causality as Proctor et al. [2018] , Hsiang [2010], most other applied econ papers these days do, rather than merely asserting that reduced form methods "work".
2. Back off the causal language, starting with "Effect" in the manuscript's title.
I think either of these options could lead to a paper that should be published in a journal like Nature. To be clear, i think a paper that's ultimately associa- tional (not causal) still would also be terrific and valuable, and an improvement over one that overstates the causal interpretation.
References
Solomon M. Hsiang. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. Proceedings of the National Academy of Sciences, 107(35):15367–15372, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1009510107. URL https://www.pnas.org/ content/107/35/15367.
Bruce Meyer. Natural and quasi-experiments in economics. Journal of Busi- ness and Economics Statistics, 13(2):151–161, April 1995.
Jonathan Proctor, Solomon Hsiang, Jennifer Burney, Marshall Burke, and Wolfram Schlenker. Estimating global agricultural effects of geoengi- neering using volcanic eruptions. Nature, 560, 08 2018. doi: 10.1038/ s41586-018-0417-3.

Referee #3 (Remarks to the Author):

The premise of the paper is very important: to use the various measures and responses implemented across different sovereignties as a natural experiment for assessing whether these measures were effective and to what extent, for possibly showing which measures were most effective, and for estimating the lives thus far protected. The study has an impressive compilation of data on policy and outcomes compiled from many sources and organized in a, necessarily, ad hoc but reasonable fashion given available resources. That said, my enthusiasm is considerably dampened as a few of the assumptions the authors impose are wrong (i.e. the lag and the mean infection period), and the findings suggest an unreasonably fast doubling time and R0 for the virus in its unimpeded state. Further, there is a problematic lack of sensitivity analyses and cross validation showing whether the findings are robust to withholding of blocks of data, groupings of policies, and adjustments of the model form. Lastly, I have strong concerns that the model system has identifiability issues.

Major Comments

Page 14: The use of a one-week lag is too short—2 weeks would be more appropriate and this needs to be used. Individuals acquire infection and then experience a latent period (3-5 days), a period of pre-symptomatic shedding (if they become symptomatic, which is the population that will be confirmed, 1-3 days), a period of symptom intensity growth (3-5 days), and then testing (the results are not returned for 3 days typically, though some countries could turn it around more quickly). The aggregate is a 10-16-day delay.

Figure 3 and Page 15—how well do your "no-policy" growth rates match estimates of R0 in the literature? This would be a good check on the model fit, particularly in the early stages. Here R0 = g/\gamma +1, so a 3 day 1/gamma yields R0=2.2; however, for the authors' simple SIR, 1/gamma should be more like 7-10 days, which puts R0 around 4. This is too high and indicates that the dynamic model may be mis-specified or that the whole system has identifiability issues. With gamma = 0.052, as the authors suggest (an unreasonable long mean infection period of almost 20 days), R0=8.5, which is an outlier estimate. These numbers have to be reconciled with estimates of R0 from other studies, and strong justification for the discrepancy would have to be provided (one that would mandate re-evaluation of those other prior estimates of R0), if this is to be considered credible. The high R0 suggests the authors are grossly over-estimating the intervention effects (Figure 4). The problem may stem from using an SIR rather than an SEIR construct, or, perhaps more likely, from the use of a one-week lag, which is too short.

Page 5: "We estimate that in the absence of policy, early infection rates of COVID-19 grow 45% per day on average, implying a doubling time of approximately two days." The authors also estimate 55% per day for Wuhan. Some estimates of doubling are that low, but recent detailed, credible estimates for Wuhan prior to quarantine come in at 5.2 days (Wu et al. Nature Medicine, 2020). Note that the Wu et al. study is for the period prior to control. Can you reconcile this discrepancy? At a minimum, it needs to be discussed and qualified. This points to the same problems as mentioned above: the growth rates reported here are high and the R0 estimates seem unreasonable. The short lag (one week) the authors use, may be contributing to this issue, the dynamic model may be too simple/mis-specified, and the overall system may have identifiability issues. Small additional comment: in this same paragraph, the authors present statistical significance—what does this refer to—is this specific to the overall model fit, specific parameters, the growth rate itself?

Figure 4: The SIR simulation assumes that observed cases represent all infections, whereas most infections are actually NOT observed (i.e. most are undocumented). While under-reporting biases do not impact your statistical fitting (provided they are stationary), they do affect the depletion of susceptibles in your dynamic model and must be represented there. For example, for China, the projection of ~75M cases would indicate there were many more total infections, 150-750M, most of which were not documented. Because of the high R0, this simulated outcome is already likely too aggressive, but to the point I'm making here, the model needs to know of the additional undocumented infections as they represent an enormous depletion of susceptibles that will greatly slow the growth of confirmed cases over time. Even for Italy, this may be a factor, as the presented uncontrolled growth scenario would suggest up to 10-12M total infections, which is a sizeable portion of the population. To perform these projections correctly, the authors need to adjust the total infection numbers in their dynamic simulations to account for undocumented infections. Note, rates of undocumented infection varies by country, as testing and reporting practices vary. This doesn't obviate the problem that the authors' no-policy projections are too aggressive (due to the high R0), and when compared with the full model, which is pinned to data, depict too large an effect for the intervention policies. I believe the econometric model has over-estimated the policy effect, possibly due to the short lag (it should be two weeks, not one), or maybe due to the SIR rather than SEIR form, or simply because the system (Eq. 7) solution is not uniquely identifiable.

To address the issue of identifiability, the authors should generate a number of synthetic outbreaks with various prescribed characteristics (thetas, beta, mu, gamma, delta), using the model system (Eqs. 7-10)—including time variable control measures (thetas_t) of variable effectiveness. They should add realistic system noise to the simulation and then feed those mock observations into Equation 7 to see if the system can reliably estimate/reconstruct the prescribed characteristics used to generate the synthetic outbreaks in the first place. And this should be done for a range of different prescribed outbreaks and control policies. By doing this, we can see whether the system can consistently find the correct global solution or is prone to being trapped in errant local minima.

Some sensitivity analyses and cross-validation also has to be presented for the model with actual data. Are the results sensitive to the grouping of policies, to an altered model form (e.g. SEIR instead of SIR), to a change in the lag (2 weeks v. 1.5 weeks)?

Minor Comments

Page 13: “This approach non-parametrically accounts for arbitrary forms of spatial auto-correlation or systematic misreporting in regions of a country on any given day (it generates larger estimates for uncertainty than clustering by i).” Can the authors clarify this for a broader audience?

Page 14—can the authors provide further justification for the difference in policy implementation for Iran requiring a dummy variable. It seems ad hoc. Is it something special about an authoritarian theocracy? How is policy implementation different and what’s the evidence base? Why shouldn’t China differ (they are authoritarian)? Is it cultural compliance? Why not a dummy interactive variable for all countries?

The discussion of how the regression model (Eq. 7) is robust to systematic bias on page 14 is not as important as non-systematic biases arising from difference in testing practices in space and time. The authors address the latter extensively in the supplementary methods, which is great, but some mention of this issue in the main methods section should be provided after the discussion of systematic bias.

Referee #4 (Remarks to the Author):

The authors present an analysis of COVID-19 interventions. Measuring the impact of control measures is an important question, and the study provide a good summary of local-level measures in multiple locations. However, it was not clear that the study appropriately accounts for delays and uncertainty, which may influence the strength of the results.

Main comments:

- The authors note method "is robust to systematic under-surveillance; and it accounts for changes in procedures to diagnose positive cases". However, it wasn't clear how they handled changes in case definition (e.g. https://www.medrxiv.org/content/10.1101/2020.03.23.20041319v1). Such changes could bias transmission estimates upwards - it's likely that the estimated doubling time of two days influenced by this could be an artefact of these changes in testing/definitions.

- p5: "25.23% per day (p< 0.05)" It's not clear what p-values represent here. Surely there is also some uncertainty in the quoted growth estimate?

- "At the time of writing, the minimum susceptible population fraction in any of the administrative units analyzed is 99.4% of the total population (Lodi, Italy: 1,445 infections in a population of 230,000)." China (e.g. PMID 32171059 suggests 95% susceptible at end of Jan). It is also important to distinguish between confirmed cases (which are reported) and infections (which are generally unknown unless there is a serosurvey); the two may be very different.

- Figure 2: It is notable that the lockdown in France seems to have little effect. Could this be because insufficient time has passed for the measures to have an effect? If so, it could be misleading to present evidence showing little effect. This potential bias can be seen in the estimates for China, which appear to become more effective over time - in reality this may just be a transient that the model has not yet accounted for.

- The authors account for delays resulting from the incubation period by lagging impact by a week, but how could delay from symptom onset to reporting influence results? It seems like it would be better to account for the distribution of delay here, rather than simply lagging by a week, which does not account for the full range of uncertainty.

- "Results are marginally statistically significant for France (p< 0.09)". Terms like "marginally statistically significant" are not rigorous or particularly helpful. It would be better to just report the effect size and p-value, so reader can evaluate the strength of evidence.

- Thank you for providing the underlying data and code for the analysis.

Author Rebuttals to Initial Comments:

Referee #1

Referee comments in italics. **[Author] reply in bold**.

A. *Summary of key results*

The paper presents evidence from small area panel datasets for China, South Korea, Iran, Italy, France and the USA of how the uninhibited exponential daily growth rate of COVID-19 cases in the initial phase of disease progression has been mitigated by the collective impact of the anti-contagion policies adopted in each of these settings. The headline finding is that current policies have prevented or delayed around eighty million infections. This estimate is calculated by comparing actual trends in growth rates with those predicted from the models with all of the policy variables switched off.

Our reply: This is accurate. However, the "eighty million infections" value has been adjusted based on newer data and accounting for unreported cases.

B. *Originality and significance*

The manuscript addresses a question of urgent and substantial significance for global public health and the world economy. It provides evidence from direct measurements of retrospective observational data to complement evidence from epidemiological numerical simulations.

C. *Data and methods*

Epidemiological data is not my specialism but sources, imputations and assumptions are stated clearly in the Appendix and appear reasonable. A reassuring validation exercise against the Johns Hopkins data is provided.

p.17 notes that the school closure variable is always on (=1) for the South Korea dataset, so that it is effectively dropped from the model (submerged into the constant term).

Our reply: This is accurate.

D. *Appropriate use of statistics and treatment of uncertainties*

The method adopted is to use separate country (two way fixed effect) panel data regressions for the daily difference in the natural logarithm of infected cases as a function of a set of policy dummies. The fixed effects include time (t), sub-national units (i), day of the week, and dummies for changes in testing regime. Identification of the policy impacts therefore comes from differences across the sub-national units in the variation within these sub-national units over time and for a given day of the week and testing regime (i.e., a differences in differences strategy).

The paper argues that the method adopted accounts for:

- *differences in baseline growth rates*
- *systematic patterns unrelated to policy*
- *is robust to systematic under-surveillance of cases*
- *accounts for changes in procedures to diagnose positive cases*

The authors should be clear to themselves and to their audience that the definition of z as "exogenous changes in independent policy variables" is not an inherent characteristic of these variables but a crucial identification assumption that they are adopting to allow the association between these variables and the outcome to be attributed to the causal effect of those policies. To assume that the policy being switched on is exogenous and independent is to assume that it is effectively randomized across areas and days. For example it could, in principle, be problematic if the timing of policy adoption was influenced by past changes in the local caseload or by the anticipation of future changes in cases.

Our reply: To address this concern, we have eliminated the description of these changes as "exogenous" (referenced by the reviewer), and introduced a new paragraph in the Discussion to directly address this issue, using language for a general audience. The main text now reads:

> **"Our analysis measures changes in local infection growth rates associated with changes in anti-contagion policies, treating each subnational administrative unit as if it were in a natural experiment. Intuitively, each administrative unit observed just prior to a policy deployment serves as the "control" for the same unit in the days after it receives a policy "treatment". Thus, a necessary condition for our estimates to be interpreted as the plausibly causal effect of these policies is that**

> **the timing of policy deployment is independent of infection growth rates (Angrist & Pischke, 2008). Such an assumption is supported by epidemiological theory, which predicts that infection totals in the absence of policy will be near-perfectly exponential early in the epidemic, (Chowell et al, 2016) implying that pre-policy infection growth rates in this context should be constant. The policies we analyze are unlikely to have been deployed in reaction to or anticipation of changes in growth rates, since epidemiological guidance to decision-makers explicitly projected constant growth rates in the absence of anti-contagion measures (Ferguson et al, 2020; Flaxman et al, 2020; Lourencco et al, 2020; Maier & Brockman, 2020). In practice, decision-makers have tended to deploy policies in response to the count of total infections in their locality, rather than their growth rate (Ferguson et al, 2020), in response to outbreaks in other regions or countries (Tian et al, 2020), or based on other arbitrary and exogenous factors, such as closing schools on a Monday or after Spring Break (Education Week, 2020).”**

p.11 I do not follow the logic of the claim that “usefully equation 2 does not depend on w”. From equation 1 it is in general true that the derivative δf(.., w)/δx can be a function of w and it may be the case that these controls may influence the mediator and hence δx/δz may also depend on w. This need not be a problem so long as the confounding effect of w is captured by the fixed effects or any controls are included directly in the regression. Basically the assumption of no unobserved confounders is being made, conditional on the fixed effects (for sub-national units, day effects and changes in testing regime).

Our reply: The reviewer is correct; we were not sufficiently clear in the original text. The sentence now states:

> **“Usefully, for a fixed population observed over time, empirically estimating an average value of the local derivative on the left-hand-side in Equation 2 does not depend on explicit knowledge of *w*.”**

p.13 It could be noted that the linear specification in (7) assumes - for simplicity and identification - that contribution of individual policies to growth rate isadditive.
I note that a log-linear specification is being used.

Our reply: This is a helpful suggestion. In the paragraph just before Equation 7, we now state:

> **"Given the limited quantity of data currently available, we use a parsimonious model that assumes the effects of policies on infection growth rates are approximately linear and additively separable. However, future work that possesses more data may be able to identify important nonlinearities or interactions between policies."**

Predictions of the growth rate based on this model are fine but if were to be used to predict the number of cases (I) then the retransformation problem would have to be taken into account ie. predictions would be a function of exp(ε).

Our reply: The reviewer is correct that if we were directly exponentiating predictions for the regressand in Equation 7, we would need to adjust predicted values by the factor exp(σ^2/2), where σ is the root mean squared error of the regression. However, we do not do this exponentiation anywhere in the analysis. In Figure 3, we present predictions from Equation 7 without exponentiating them. In Figure 4, we use our estimates of the growth rate in a numerical computation where we integrate growth rates
period-by-period. This differs from simple exponentiation of the regression model predictions because we directly utilize our estimates of the average growth rate itself in the numerical calculation. Uncertainty in these simulations is estimated via re-sampling.

p.14 The authors should be clear that their method should be robust only to time-invariant under-reporting.
This appears to be how they interpret "systematic" undersurveillance.

Our reply: To address this concern, we now state in the Introduction

> **“Our econometric approach ... is robust to systematic under-surveillance specific to each subnational unit;...”**

and we state in the discussion

> **“Our analysis accounts for documented changes in the availability of and procedures for testing for COVID-19 as well as differences in case-detection across locations; however, unobserved trends in case-detection could affect our results (see Methods).”**

where we describe the results of new analysis, shown in new Extended Data Figure 2, where we estimate the potential impact of trends in case-detection. The discussion now states:

> **“[O]ur analysis of estimated case-detection trends (Russell et al., 2020) (Extended Data Fig. 2) suggests that the magnitude of this potential bias is small, elevating our estimated no-policy growth rates by 0.022 (6%) on average.”**

Where these calculations are based on results from a recent study by the Centre for Mathematical Modeling of Infectious Diseases[1] that computes country-specific and time-varying rates of case detection based on the assumption that the case fatality rate of COVID-19 is constant.

Finally, we substantially expand our discussion of this issue in the Methods section (See Lines 566 - 582) to separately and specifically describe time-invariant and time-varying under-reporting, so readers may clearly understand the distinction. This section explains and derives the robustness of our approach to the former but not the latter. To do this, we introduce Equation 8, describing the magnitude of the potential biaswith

[1] Russell, T. W. et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine (2020). https://cmmid.github.io/topics/covid19/371severity/global_cfr_estimates.html. Accessed:

2020-04-09.

time-varying under-reporting, which we try to characterize in the new Extended Data Figure 2.

p.14 The discussion and rationale for using weekly lags in China is a bit opaque and could be more explicit. It is justified by the incubation period. Wouldn't this apply to other countries as well.

Our reply: To address this concern and clarify what transient dynamics we can and cannot characterize, we have introduced a new paragraph to the results section that reads:

> **"In China, only two policies were enacted across 116 cities early in a seven week period, providing us with sufficient data to empirically estimate how the effects of these policies evolve over time without making any assumptions about the timing of these effects (see Methods and Fig. 2b). We estimate that the combined effect of these policies significantly reduced the growth rate of infections by $-$0.14 (SE = 0.031) in the first week immediately following their deployment (also see Extended Data Fig. 5a), with effects doubling in the second week to $-$0.30 (SE = 0.040), and stabilizing in the third week at $-$0.34 (SE = 0.036). In other countries, we lack sufficient data to estimate these temporal dynamics explicitly and only report the average pooled effect of policies across all days following their deployment (see Methods). If other countries were to exhibit transient responses similar to that observed in China, we would expect effects in the first week following deployment to be smaller in magnitude than the average effect for all post-deployment weeks. Below, we explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number of days, but we do not find evidence this improves model fit (Extended Data Fig. 5b)."**

In addition, we have introduced a complete new section on "Transient Dynamics" in the Methods section under Econometric Analysis. This section details (a) why and how we analyze distributed lags only in China, in addition to (b) describing newevent-study

analysis of Chinese cities (Extended Data Figure 5) and (c) "fixed lag" models across all countries (suggested by other reviewers) which explore the sensitivity of our results to assuming that policies do not affect infection growth rates in the first week following their deployment (Extended Data Figure 5 and Supplementary Table 5.).

I would urge the authors to include the full regression results in the Appendix. These are currently lacking.

Our reply: Our full regression results are now presented in Supplementary Table 3.

Note that Figure A1 plots the estimated residuals not the (unobserved) errors. These are non-normal but normality is not a requirement of the modeling and inference strategy that is adopted.

Our reply: We have changed the terminology in the title and the description of Extended Data Figure 10 (formerly Figure A1) from "errors" to "residuals" to reflect this distinction. We also modified the language in the main text to read:

> **"We display the estimated residuals ε_{cit} in Extended Data Fig. 10, which are mean zero but not strictly normal (normality is not a requirement of our modeling and inference strategy), and we estimate uncertainty over all parameters by clustering our standard errors at the day level."**

Figure A2 would be clearer if the legend for "observations in our data" also appeared in the panel for Hubei.

Our reply: We have moved the legend for "observations in our data" into the Hubei panel in Extended Data Figure 1 (formerly Figure A2).

E. Conclusions: robustness, validity, reliability

A key policy outcome is cases that require hospitalization (or more specifically intensive care) central to the policy of flattening the curve. I wonder if more could be said about this based on the findings of the analysis?

Our reply: To address this concern, we have added the following text to the Discussion section:

> **"While our analysis has focused on changes in the growth rate of infections, other outcomes, such as hospitalizations or deaths, are also of policy interest. Because these outcomes are more context- and state-dependent than infection growth rates, their analysis in future work may require additional modeling approaches. Nonetheless, we experimentally implement our approach on the daily growth rate of hospitalizations in France, the only country in our sample where we were able to obtain hospitalization data at the granularity of this study, and we find that the total estimated effect of anti-contagion policies is similar to our reported effect on the infection growth rate (Extended Data Fig. 6c)."**

F. Suggested improvements

Going beyond the publication of the current manuscript the authors should be encouraged in their broader project that can be continuously updated with more countries (such as Hong Kong, Singapore, Taiwan, Spain, Germany, UK) and additional time series observations as they become available.

Our reply: We fully updated the current manuscript to include all available data up to April 6, 2020 (the original manuscript data ended on March 18, 2020), except for France and Iran, where the subnational health data are available only through March 25, 2020 and March 22, 2020, respectively. The number of policies in our sample has increased from 936 to 1,659. We are now applying for funding to expand this data collection effort.

G. References

This is not my field of expertise but the references appear comprehensive. For readers' benefit the incomplete references should be finalized: references 2, 3, 4, 13, 14, 17, 22, 26.

Our reply: We thank the reviewer for checking the references so thoroughly. We have fixed these references.

H. Clarity and context

The manuscript is presented in a clear and lucid way and brings home the impact of the issue and the findings. As a non-epidemiologist I found the text accessible.

Our reply: We thank the referee for their careful and thoughtful review of the manuscript. We believe many of these comments led to additions to the manuscript that have significantly improved its quality.

Referee #2

Referee comments in italics. **[Author] reply in bold**.

I enjoyed the opportunity to read MS 2020-03-04486A: "The Effect of Large-Scale

Anti-Contagion Policies on the Coronavirus (COVID-19) Pandemic" by Professor Hsiang and co-authors.

Positives

This is a terrific paper and I think it should be published in a top, general interest science journal. My reasons for this are:

1. *I don't think I have ever refereed a paper with greater potential impact on human welfare. Nor where the question at hand was such an urgent one time-wise.*
2. *I very much appreciate the authors' reduced form approach as a com- plement to calibrated epidemiological models. They position their paper very nicely thus.*
3. *The analysis is clearly described, pedagogical, and well motivated.*
4. *The data collection effort on such a tight timeframe is nothing short of amazing*
5. *Despite the above, neither the analysis nor prose appear rushed. It is a careful exercise. Bravo!*

Our reply: We sincerely appreciate the referee's kind comments.

Negative

Super minor but dropping Chinese cities where "we cannot find a policy deployment date" seems necessary but should note it could also potentially impart some LATE or bias issues.

Our reply: To address this point, at the end of the section on Chinese Policy Data in the Methods section, we now state, "Thus our results are only representative for the sample of 116 cities where we could obtain policy data."

My main problem is that the authors do not develop any argument for why the anti-contagion policies examined are exogenous. Yes, in general, best- practice reduced form methods are great at taking us to causal inference with observational data, which is terrific. But unless i have *missed it (i am sorry if so), the authors skip the argument as to why the conditional independence assumption required for causal inference is satisfied in their context.*

Meyer [1995] noted: "If one cannot experimentally control the variation one is using, one should understand its source". I see no effort by the authors to explain why some regions adopted some policies and not others, and the timing of these policies. Without this, it is difficult to know whether policies were adopted using information about COVID infection rate trends or future infection rates that is unavailable to the authors.

The authors draw a parallel in the abstract to existing empirical work on policies that affect economic growth: "We then apply reduced-form econo- metric methods, commonly used to measure the effect of policies on economic growth". In studying volcanic eruptions, Proctor et al. [2018] uses a treatment whose timing is certainly exogenous to agricultural yields, so there's not much of a leap required compared to this manuscript. Returning squarely to economic growth, the exogenous case is again much easier to make with temperature and cyclone shocks [Hsiang, 2010].

Fundamentally, why do different places adopt different policies? For example, if the variation is because it's just really difficult to know impacts and policy makers were just guessing at what might work, that would be a good story to develop for the authors. But I see no case made along these lines. Nor even any assessment of how allowing for different pre-policy trends matters or doesn't matter.

Given the above, i suggest one of two avenues:

1. *Develop the argument for causality as Proctor et al. [2018] , Hsiang [2010], most other applied econ papers these days do, rather than merely asserting that reduced form methods "work".*
2. *Back off the causal language, starting with "Effect" in the manuscript's title.*

I think either of these options could lead to a paper that should be published in a journal like Nature. To be clear, I think a paper that's ultimately associa- tional (not causal) still would also be terrific and valuable, and an improvement over one that overstates the causal interpretation. References

Solomon M. Hsiang. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. Proceedings of the National Academy of Sciences, 107(35):15367–15372, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1009510107. URL

https://www.pnas.org/ content/107/35/15367.

Bruce Meyer. Natural and quasi-experiments in economics. Journal of Busi- ness and Economics Statistics, 13(2):151–161, April 1995.

Jonathan Proctor, Solomon Hsiang, Jennifer Burney, Marshall Burke, and Wolfram Schlenker. Estimating global agricultural effects of geoengi- neering using volcanic eruptions. Nature, 560, 08 2018. doi: 10.1038/ s41586-018-0417-3.

Our reply: To address this concern, we have adopted the first strategy suggested by the reviewer. We now include a paragraph in the Discussion of the main text explaining why policy deployments are likely independent of infection growth rates:

> **"Our analysis measures changes in local infection growth rates associated with changes in anti-contagion policies, treating each subnational administrative unit as if it were in a natural experiment. Intuitively, each administrative unit observed just prior to a policy deployment serves as the "control" for the same unit in the days after it receives a policy "treatment". Thus, a necessary condition for our estimates to be interpreted as the plausibly causal effect of these policies is that the timing of policy deployment is independent of infection growth rates (Angrist & Pischke, 2008). Such an assumption is supported by epidemiological theory, which predicts that infection totals in the absence of policy will be near-perfectly exponential early in the epidemic, (Chowell et al, 2016) implying that pre-policy infection growth rates in this context should be constant. The policies we analyze are unlikely to have been deployed in reaction to or anticipation of changes in growth rates, since epidemiological guidance to decision-makers explicitly projected constant growth rates in the absence of anti-contagion measures (Ferguson et al, 2020; Flaxman et al, 2020; Lourencco et al, 2020; Maier & Brockman, 2020). In practice, decision-makers have tended to deploy policies in response to the count of total infections in their locality, rather than their growth rate (Ferguson et al, 2020), in response to outbreaks in other regions or countries**

> **(Tian et al, 2020), or based on other arbitrary and exogenous factors, such as closing schools on a Monday or after Spring Break (Education Week, 2020)."**

We thank the referee for their careful and thoughtful review of the manuscript. We believe these comments led to additions to the manuscript that have significantly improved its quality.

Referee #3

Referee comments in italics. **[Author] reply in bold**.

The premise of the paper is very important: to use the various measures and responses implemented across different sovereignties as a natural experiment for assessing whether these measures were effective and to what extent, for possibly showing which measures were most effective, and for estimating the lives thus far protected. The study has an impressive compilation of data on policy and outcomes compiled from many sources and organized in a, necessarily, ad hoc but reasonable fashion given available resources. That said, my enthusiasm is considerably dampened as a few of the assumptions the authors impose are wrong (i.e. the lag and the mean infection period), and the findings suggest an unreasonably fast doubling time and R0 for the virus in its unimpeded state.

Our reply: We address these concerns below, following the reviewer's main comments. Although, to summarize, we believe our exposition of how lags were used in the analysis was unclear in the original manuscript and have worked to substantially clarify these issues in the revised text (e.g. we believe the 1-week lag interpretation resulted from some unclear language that has since been removed). In response to these concerns, we have also conducted additional sensitivity analyses using different lags to understand if they substantially affect our main results and find they do not. We also conduct an event study in China to understand whether policy effects can be detected in the days immediately following policy deployment.

In response to concerns regarding comparisons with prior findings: we systematically compare our doubling time to results from the previous literature referenced by the reviewer. We find that previously published longer doubling times (e.g. 5-7 days) depended on very early data from Wuhan prior to standard diagnosis procedures (instituted on January 15, 2020 by the Chinese government), data which also includes multiple major irregularities that we now document in the Appendix. We are able to replicate earlier findings of long doubling times (>5 days) using these data and our original estimating procedure. However, when we impose the data quality-control

measures used in our original analysis on the exact same sample and use the same estimating procedure, we obtain doubling times that are much shorter (2 days) and in agreement with results from other countries (South Korea, France, Italy, USA). We conclude that some published doubling times are likely an artifact of data collection challenges during the very first weeks of the COVID-19 outbreak in Wuhan, when there was limited surveillance of cases and active censorship of case reporting.

Further, there is a problematic lack of sensitivity analyses and cross validation showing whether the findings are robust to withholding of blocks of data, groupings of policies, and adjustments of the model form.

Our reply: To address these concerns, we now conduct multiple new analyses, each of which is now detailed in the manuscript and summarized here:

(1) **We conduct sensitivity analyses to withholding blocks of data at thefirst administrative level (Extended Data Figure 3-4) and find our mainresults (no-policy growth rate and total effect of policies) are unaffected. We find that individual policies are somewhat more sensitive, consistent with these values being more challenging to estimate with limited data. These results are now reported in the main text.**

(2) **We also re-estimate our model without grouping any policies and recompute the effect of all policies combined (Supplementary Table 4) and the total effect of actual policies (Extended Data Figure 6), finding that neither are substantively or significantly affected by this change.**

(3) **We explore the effect of altering the lag structure of the model, assuming various different fixed lag lengths (Supplementary Table 5 and Extended Data Figure 5). We find that model fit is not improved (and often deteriorates) by using lags larger than a week and that our point estimates for policy impact are largely unaffected by these adjustments to model form.**

Lastly, I have strong concerns that the model system has identifiability issues.

Our reply: To address this concern, we implement the simulation tests suggested by the reviewer to demonstrate the identifiability of the model (detailed below, results summarized in Extended Data Figures 8-9).

Major Comments

Page 14: The use of a one-week lag is too short—

Our reply: We believe that our exposition in the original manuscript was unclear. Our main model does not use a one-week lag and we have revised the text to provide greater clarity on this issue (we believe the 1-week lag interpretation resulted from some unclear language that has since been removed). We now clarify this in the Methods section when describing the parameters of interest, immediately after presenting the estimating equation:

> **"… θ is the average effect of the policy on growth rate *g* over all periods subsequent to the policy's introduction, thereby encompassing any lagged effects of policies."**

2 weeks would be more appropriate and this needs to be used. Individuals acquire infection and then experience a latent period (3-5 days), a period of pre-symptomatic shedding (if they become symptomatic, which is the population that will be confirmed, 1-3 days), a period of symptom intensity growth (3-5 days), and then testing (the results are not returned for 3 days typically, though some countries could turn it around more quickly). The aggregate is a

10-16-day delay.

Our reply: We do not assume a fixed lag length, as the reviewer was led to believe by our unclear language. We apologize for the confusion. As described in the sentence above, our main estimates simply report the Average Treatment Effect (ATE) for all days during which a policy is deployed, a standard measure in policy impact evaluations[2] . As the reviewer correctly intuits, the time-scale of the transient response to the policy,

[2] Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2008.

combined with the length of time over which a policy is evaluated, will influence the ATE. Thus, the ATE is not a "deep parameter" describing a fundamental biological relationship that is unchanging (and which might require lag structures similar to what the reviewer suggests). Rather, the ATE simply describes the unconditional average change in Y due to a change in X that is observable in current real world contexts, given actual heterogeneity in populations, lags, treatment conditions, etc. The distinction between the ATE and fundamental parameters often measured in epidemiological studies was high-lighted in the original introduction:

> "The reduced-form statistical techniques we use are designed to measure the total magnitude of the effect of *changes in policy*, without attempting to explain the origin of baseline growth rates or the specific epidemiological mechanisms linking policy changes to infection growth rates (seeMethods). Thus, this approach does not provide the important mechanistic insights generated by process-based models; however, it does effectively quantify the key policy-relevant relationships of interest using recent real-world data, when fundamental epidemiological parameters are still uncertain."

However, to clarify this further, we have added the sentence described in response to the previous comment (above).

Nonetheless, the time-scale of the infection-growth response to policy is important, as the reviewer rightly points out. Thus, to fully address the reviewer's concern, we have implemented three additional changes:

(1) In the main text we have introduced a new paragraph (fifth in the Results section) breaking down what we can and cannot observe about the transient response, given currently available data. We also explain how the transient effect might influence our estimates. Because we cannot know *ex ante* what the correct lag structure is (since no prior study has measured the effects we characterize), we simply "allow the data to speak" when examining data from China, imposing no assumptions on the structure of the lag. Unfortunately, we do not have sufficient data to implement this flexible approach

in other countries, and thus simply report an estimated average treatment effect of each policy (not assuming any lag structure). The paragraph now reads:

> **"In China, only two policies were enacted across 116 cities early in a seven week period, providing us with sufficient data to empirically estimate how the effects of these policies evolve over time without making any assumptions about the timing of these effects (see Methods and Fig. 2b). We estimate that the combined effect of these policies significantly reduced the growth rate of infections by -0.14 (SE=0.031) in the first week immediately following their deployment (also see Extended Data Fig. 5a), with effects doubling in the second week to -0.30 (SE=0.040), and stabilizing in the third week at -0.34 (SE=0.036). In other countries, we lack sufficient data to estimate these temporal dynamics explicitly and only report the average pooled effect of policies for all days following their deployment (see Methods). If other countries were to exhibit transient responses similar to that observed in China, we would expect effects in the first week following deployment to be smaller in magnitude than the average effect for all post-deployment weeks. Below we explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number of days, but we do not find evidence that this systematically improves model fit (Extended Data Fig. 5b)."**

And we have further explained this approach and its interpretation in a new section of the Methods titled "Transient Dynamics" which reads:

> **"In China, we are able to examine the transient response of infection growth rates following policy deployment because only two policies were deployed early in a long seven-week sample period during which we observe many cities simultaneously. This provides us with sufficient data to estimate the temporal structure of policy effects without imposing assumptions regarding this structure. To do this, we estimate a distributed-lag model that encodes policy parameters using weekly lags based on the date that each policy is first implemented in locality *i*. This means the effect of a policy implemented one**

> **week ago is allowed to differ arbitrarily from the effect of that same policy in the following week, etc. These effects are then estimated simultaneously and are displayed in Fig. 2 (also Supplementary Table 3). Such a distributed lag approach did not provide statistically meaningful insight in other countries using currently available data because there were fewer administrative units and shorter periods of observation (i.e. smaller samples), and more policies (i.e. more parameters to estimate) in all other countries. Future work may be able to successfully explore these dynamics outside of China."**

(2) Because we observe changes in infection growth rates within the first week of policy deployment in China, we also conduct a new "event study" analysis in a sample of Chinese cities, aimed at inspecting whether an infection growth rate response actually manifests in the individual days immediately following policy deployment (at time-scales less than one week). The results of this new analysis are shown in Extended Figure 5a, and the analysis is described in the new Transient Dynamics section:

> **"We also explore the day-by-day response to the first anti-contagion policies in a limited number Chinese cities using an event study approach. We examine the 36 cities in which five days of infection growth data immediately before and after deployment of the first anti-contagion policy (home isolation) are available (similar samples were unavailable in the other countries we study). Pooling these data, we then estimate average rates of infection growth five days before deployment, four days before, etc., shown in Extended Data Fig. 5a. In this limited sample of cities, we find that infection growth rates separate from the average pre-policy growth rate within the first three days following deployment of the policy."**

(3) We explore how our results would change if we impose the assumption that policies cannot affect infection growth rate for a fixed number of days immediately after they are deployed. This is, we believe, similar to what the reviewer suggests. Results from this analysis are displayed in Extended Data Figure 5 and Supplementary

Table 5. Essentially, the estimated effects of policies do not change substantially under this adjustment, although we do find some evidence of attenuation bias (shrinkage of coefficients towards zero due to mismeasurement of the treatment variable). This new analysis is now described in the Transient Dynamics section:

> **"As a robustness check, we examine whether excluding the transient response from the estimated effects of policy substantially alters our results. We do this by estimating a "fixed lag'" model, where we assume that policies cannot influence infection growth rates for *L* days, recoding a policy variable at time *t* as zero if a policy was implemented fewer than *L* days before *t*. We re-estimate Equation 7 for each value of *L* and present results in Extended Data Fig. 5 and Supplementary Table 5."**

The results of this analysis are reported in the main Results section, after describing the results of the distributed lag analysis, and now states:

> **"We explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number of days, but we do not find evidence this improves model fit (Extended Data Fig. 5b)."**

We also discuss these results when we report the estimated effects of individual policies:

> **"The estimated effects of individual policies are broadly robust to assuming a delayed effect of all policies (Extended Data Fig. 5c)."**

Figure 3 and Page 15—how well do your "no-policy" growth rates match estimates of R0 in the literature? This would be a good check on the model fit, particularly in the early stages. Here R0 = g/\gamma +1, so a 3 day 1/gamma yields R0=2.2; however, for the authors' simple SIR, 1/gamma should be more like 7-10 days, which puts R0 around 4. This is too high and indicates that the dynamic model may be mis-specified or that the whole system has identifiability issues. With gamma = 0.052, as the authors suggest (an unreasonable long mean *infection period of almost 20 days), R0=8.5, which is an outlier estimate. These numbers have to be reconciled with estimates of R0 from other studies, and strong justification for the discrepancy would have to be provided (one that would mandate re-evaluation of those other prior estimates of R0), if this is to be considered credible.*

Our reply: Our updated data now implies an R_0 ≈ 5.5, for a growth rate of 0.36 (our estimate for Wuhan). This value is within the range of previously estimated values for R_0 in the literature using a variety of estimation methods. Published estimates of no-policy or pre-policy R_0 values range from 2.5 in Singapore[3] and 2.6 in South Korea,[4] to 3.1 in Italy,[5] 3.8 in mainland China,[6] 4.4 in Iran,[7] 5.25 in Brazil,[8] 5.7[9] in Wuhan, and 6.49 in the Hubei province.[10] Our estimate is at the higher end of this range, but it is not an outlier. Further, we independently replicate similar values for the growth rate in South Korea (0.31), Italy (0.37), France (0.34), and the US (0.30) using an identical and transparent empirical procedure.

We also note that we reconcile our estimates with earlier published estimates of lower growth rates in Wuhan (e.g. Wu et al, Nature Medicine, 2020) below and in the main text. In that analysis (now documented in Supplementary Table 2), we find that previously reported low growth rates in Wuhan are obtained when relying on unstandardized data that contains many obvious irregularities. Basic cleaning of that data, as was implemented in our original analysis but not in several prior published studies, produces results consistent with our reported values.

The high R0 suggests the authors are grossly over-estimating the intervention effects (Figure 4). The problem may stem from using an SIR rather than an SEIR construct, or, perhaps more likely, from the use of a one-week lag, which is too short.

[3] Sugishita et al. 2020

[4] Mizumoto et al. 2020

[5] Cereda et al. 2020

[6] Tang et al. 2020

[7] Muniz-Rodriguez et al. 2020

[8] Crokidakis 2020

~~[9] Sanche et al. 2020~~

[10] Shen et al. 2020

Our reply: To address this concern, we have conducted new analysis using an SEIR modeling approach, rather than an SIR approach (concerns about R0 and the lag structure are addressed above). We re-run our projections (Figure 4) using an SEIR model parameterized with multiple plausible values for γ and σ. For each of these simulations, we compare the projected cumulative cases averted across all six countries to that from our main analysis, which uses γ = .08 and σ = ∞ (a latent period of 0, equivalent to an SIR model).

These results are now displayed in Extended Data Figure 7. Panels (a) and (b) show that γ and σ largely influence our final estimate of cases avoided or delayed though the estimate of "no-policy" cases. Panel (c) corroborates the referee's intuition, in that a higher value of γ does yield estimates of cases delayed/avoided that are up to 10% lower than our preferred γ if disease dynamics are SIR, though this reverses for SEIR. However, Panel (d) shows that the choices for γ or σ (and, by implication, SIR or SEIR disease dynamics) are relatively inconsequential to determining the order of magnitude of our estimates of cases avoided or delayed. The statistical uncertainty in our parameter estimates (shown in Figure 4) is much greater than this range.

This new SEIR analysis is described in the Methods under the Projections section, and we now report in the main text (at the end of the Results section):

> **"Sensitivity tests that assume a range of plausible alternative parameter values and disease dynamics, such as incorporating a Susceptible-Exposed-Infected- Removed (SEIR) model, suggest that policies may have reduced the number of infections by 57--65 million confirmed cases over the dates in our sample (central estimates)."**

Page 5: "We estimate that in the absence of policy, early infection rates of COVID-19 grow 45% per day on average, implying a doubling time of approximately two days." The authors also estimate 55% per day for Wuhan. Some estimates of doubling are that low, but recent detailed, credible estimates for Wuhan prior to quarantine come in at 5.2 days (Wu et al. Nature

Medicine, 2020). Note that the Wu et al. study is for the period prior to control. Can you reconcile this discrepancy? At a minimum, it needs to be discussed and qualified.

Our reply: Since submission, we have revised our pre-policy growth estimates based on additional data. We now compute an average 42% pre-policy growth rate (across all six countries) and a 42.8% pre-policy growth rate in Wuhan (doubling time = 1.9 days).

We have systematically compared our data, analysis, and results to

- **Wu et al., "Estimating clinical severity of COVID-19 from thetransmission dynamics in Wuhan, China."** ***Nature Medicine*** **(2020)**
- **Li et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia."** ***New England Journal of Medicine*** **(2020)**

both of which report pre-policy doubling times in Wuhan similar to the 5 day value referenced by the reviewer. We found that both of the studies relied heavily on early case data collected in Wuhan, during a period when procedures for case diagnosis had not yet been standardized, when there were questions about whether the information about the disease was actively suppressed, and when the data show many obvious irregularities. These studies use data starting on 12/10/19 (see Supplementary Table 2). As described in our original manuscript, we had intentionally not used data prior to 1/16/20 because national standards for case diagnosis were not released by the Chinese government until 1/15/20, and because data before this date appeared unreliable to us.
Specifically,

(i) the official record of cumulative cases declined from 59 on 1/8/20 to 41on 1/9/20, when cumulative cases should not decline by definition, and

(ii) no new cases were reported for seven consecutive days between 1/9/20 and 1/15/20, when at least 5 new cases per day should have been reported (more than doubling the case count) if the growth rates in Wu et al. (2020) were indeed correct.

We obtain the original data from Wu et al. and implement our approach for the sample (12/10/19-1/22/20, ending with the lockdown), recovering a case doubling time of 5.7 days, replicating the Wu et al. estimate. We then gradually shorten the sample one day at a time, re-estimating the doubling time for each sample, until we are using our sample based on our data quality control standards (1/16/20-1/21/20). We see that data irregularities, particularly the two mentioned above, bias estimates towards longer doubling times until 1/15/20, when government case definitions were issued. Estimates stabilize near a 2 day doubling time once data quality is standardized. This analysis leads us to conclude that the long doubling time reported in early studies of Wuhan was largely an artifact of poor data quality obtained during the earliest and most chaotic period of the outbreak.

This new analysis is now documented in Supplementary Table 2.

To discuss this issue for readers, as requested by the reviewer, we have heavily revised the first paragraphs of the Results section to compare our results against these other estimates and to describe our understanding for the source of this discrepancy. The main article now says:

> **"We estimate that in the absence of policy, early infection rates of COVID-19 grow 42% per day on average (Standard Error [SE] = 7%), implying a doubling time of approximately 2 days. Country-specific estimates range from 24% per day in China (SE = 9%) to 69% per day in Iran (SE = 5%). Growth rates in South Korea, Italy, France, and the US are very near the 42% average value (Fig. 2a). These estimated values differ from the observed growth rates because the latter are confounded by the effects of policy. These growth rates are not driven by the expansion of testing or increasing rates of case detection (see Methods and Extended Data Fig. 2) and are not dependent on data from any particular region of any country (Extended Data Fig. 3).**

> **“Some prior analyses of pre-intervention infections in Wuhan suggest slower growth rates (doubling every 5-7 days) using data collected before national standards for diagnosis and case definitions were first issued by the Chinese government on January 15, 2020. However, case data in Wuhan from before this date contains multiple irregularities: the cumulative case count decreased on January 9; no new cases were reported between January 9 and January 15; and there were concerns over whether information about the outbreak was actively suppressed (see Supplementary Table 2). When we remove these problematic data, utilizing a shorter but more reliable pre-intervention time series from Wuhan (January 16-21, 2020), we recover a growth rate of 43% per day (SE = 3%, doubling every 2 days) consistent with results from all other countries (Fig. 2a and Supplementary Table 3), except Iran.”**

This points to the same problems as mentioned above: the growth rates reported here are high and the R0 estimates seem unreasonable. The short lag (one week) the authors use, may be contributing to this issue, the dynamic model may be too simple/mis-specified, and the overall system may have identifiability issues.

Our reply: Our efforts to address concerns related to R_0, lag structure, and the disease model are described in response to comments above. We address the concern of identifiability in response to a comment below.

Small additional comment: in this same paragraph, the authors present statistical significance—what does this refer to—is this specific to the overall model fit, specific parameters, the growth rate itself?

Our reply: To clarify this for readers, we have replaced the p-values in this paragraph with standard errors for the estimated parameters.

Figure 4: The SIR simulation assumes that observed cases represent all infections, whereas most infections are actually NOT observed (i.e. most are undocumented). While

under-reporting biases do not impact your statistical fitting (provided they are stationary), they *do affect the depletion of susceptibles in your dynamic model and must be represented there. For example, for China, the projection of ~75M cases would indicate there were many more total infections, 150-750M, most of which were not documented. Because of the high R0, this simulated outcome is already likely too aggressive, but to the point I'm making here, the model needs to know of the additional undocumented infections as they represent an enormous depletion of susceptibles that will greatly slow the growth of confirmed cases over time. Even for Italy, this may be a factor, as the presented uncontrolled growth scenario would suggest up to 10-12M total infections, which is a sizeable portion of the population.*

To perform these projections correctly, the authors need to adjust the total infection numbers in their dynamic simulations to account for undocumented infections. Note, rates of undocumented infection varies by country, as testing and reporting practices vary.

Our reply: The reviewer is correct that it is important to distinguish between confirmed cases and total infections (which are unobserved). Much new evidence has emerged on the gap between these two values since our original submission. To address this issue, we obtained country-specific and time-varying estimates of case detection rates from the Centre for Mathematical Modeling of Infectious Diseases[11]. Throughout our analysis, whenever computing susceptible populations, we now rescale daily confirmed cases by $1/P_{ct}$, where P_{ct} is the estimated probability that a true case is detected and confirmed in country *c* on day *t*. This adjustment is now detailed in the Methods section, and it substantially affects the results presented in Figure 4 because it alters transmission dynamics when the susceptible population declines. However, we note that we continue to report the fraction of cases that one would expect to be "confirmed" as we believe it is important to maintain consistent units throughout the paper.

These adjustments altered our total averted "confirmed" cases by more than a factor of two in China, and by large values in other countries as well. We view this correction as tremendously important and thank the reviewer for focusing our attention on this critical issue.

[11] Russell, T. W. et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine (2020). https://cmmid.github.io/topics/covid19/371severity/global_cfr_estimates.html. Accessed:

2020-04-09.

This doesn't obviate the problem that the authors' no-policy projections are too aggressive (due to the high R0), and when compared with the full model, which is pinned to data, depict too large an effect for the intervention policies. I believe the econometric model has over-estimated the policy effect, possibly due to the short lag (it should be two weeks, not one), or maybe due to the SIR rather than SEIR form, or simply because the system (Eq. 7) solution is not uniquely identifiable.

Our reply: Our efforts to address concerns related to R_0, lag structure, and the disease model are described in response to comments above.

Regarding the concern about the unique identifiability of Equation 7: This equation is linear and has a unique solution for each country-sample given that our matrices of data for each country have full rank (Wooldridge 2010)[12]. We solve for the unique solution via convex optimization (ordinary least squares).

To address the issue of identifiability, the authors should generate a number of synthetic outbreaks with various prescribed characteristics (thetas, beta, mu, gamma, delta), using the model system (Eqs. 7-10)—including time variable control measures (thetas_t) of variable effectiveness. They should add realistic system noise to the simulation and then feed those mock observations into Equation 7 to see if the system can reliably estimate/reconstruct the prescribed characteristics used to generate the synthetic outbreaks in the first place. And this should be done for a range of different prescribed outbreaks and control policies. By doing this, we can see whether the system can consistently find the correct global solution or is prone to being trapped in errant local minima.

Our reply: We thank the referee for this helpful suggestion. As mentioned above, Equation 7 describes a convex optimization problem, so there cannot be any errant local minima. Nonetheless, the simulation suggested by the reviewer is helpful in allowing us to test whether our reduced form econometric approach (Equation 7) is able to recover

[12] Wooldridge 2010

the steady state effects of policy on an SIR- or SEIR-governed outbreak under a variety of epidemiological conditions. This simulation allows us to additionally test the performance of our approach when we are constrained to using cumulative, rather than active, infections.

The reviewer requests that we generate a suite of simulations varying the following parameters of Eq. 7: θ (policy effect coefficients), β (transmission rate), γ (removal rate), μ (testing regime changes), and δ (country-level day-of-week fixed effects). We follow these instructions, with a few caveats. First, we additionally vary the population size of the simulated administrative unit, to understand the sensitivity of our results to increasing susceptible population shares. Second, we do not include day-of-week or testing regime effects in our simulation. Given that these variables primarily control for measurement effects, they would simply be absorbed by the δ and μ terms and cannot affect the results of the simulation.

The structure of our simulation is described in full in the new Supplementary Methods Section 1. The results of our simulations, conducted for a range of values of γ and σ, are shown in Extended Data Figures 8 and 9 and further discussed in Supplementary Methods Section 1.

ED Figure 8 assesses the sensitivity of our estimated no-policy growth rate to changing epidemiological parameters (γ and σ)[13], while ED Figure 9 assesses the sensitivity of our estimated policy effects (total across all policies). Together, these estimates are the primary determinants of our projected "cases averted or delayed" due to policy. In both figures, we summarize potential biases in both near-ideal (a) and non-ideal (b) data environments. In the near-ideal case, a large population means that the susceptible fraction of the population remains large throughout the sample, and asymptotically exponential growth is a near-perfect assumption. In addition, we are able to directly observe active cases (as we do in our China and Korea panels). In the non-ideal

[13] **We derive the third parameter, β,** at any given time point by plugging **γ, σ, and the asymptotic growth rate *g* into the equation for the positive eigenvalue of [E, I] in a SEIR system (Ma, 2020 and Supplementary Methods Section 1). The growth rate is calculated as the assumed no-policy growth rate plus the effects of any policies that are implemented at the given time.**

scenario, a smaller population leads to observations of instantaneous growth in the late sample for which the susceptible fraction is low, and we run our regression with cumulative cases on the left-hand-side.

ED Figure 8 suggests that our estimates of no-policy growth rates are almost perfectly unbiased (largest bias = 2.7%) and essentially unaffected by disease parameters or the data environment. ED Figure 9 suggests that our estimates of policy effects may have a near-zero (min= +0.5%) or modest (max = -20.8%) bias, depending on the unknown disease parameters, although biases tend to be below 10% in magnitude. All biases greater than 1% cause us to under-estimate the effectiveness of anti-contagion policies. This leads us to hypothesize that our reported results may be on the slightly conservative side, although in practice the level of statistical uncertainty in our estimates, shown as histograms in ED Figures 8-9, dominate any potential bias.

The source of these small to modest biases can be understood as follows. The model is estimating the average growth rate effect *within our sample*. Longer latency periods cause a substantial lag between the change to transmission rate caused by policy and the resulting change to the asymptotic growth rate. Because of this, a greater portion of the post-policy observations will be of growth rates that fall between the no-policy growth rate and the steady state with-policy growth rate. In these cases, the estimated average effect of policy will generally be a lower bound on the estimated steady state policy effect. This is consistent with the results displayed in ED Figure 9. When that delay is shorter (or nonexistent, as with SIR dynamics) we see that our estimates of policy effect converge to the true steady-state effect.

Together, ED Figures 8 and 9 suggest that our empirical model (Eq. 7) estimating policy effects and the no-policy growth rate is well-specified and generally robust to underlying disease dynamics described by either an SIR or SEIR model. We demonstrate this for a range of γ and σ values, when estimating growth effects on either active or cumulative infections, and in both small and large populations. We find that our empirical estimates of no-policy growth rate are unbiased in all scenarios, and we document how much our estimates of policy impacts might be understated for longer disease latency periods.

We have added a Jupyter Notebook (in Python) to our data repository for users to further explore the behavior of our exponential regression model in synthetic outbreaks.

The Methods section in the main text now points readers towards this simulation:

> **"Our main empirical specification is motivated with an SIR model of disease contagion, which assumes zero latent period between exposure to COVID-19 and infectiousness. If we relax this assumption to allow for a latent period of infection, as in a Susceptible-Exposed-Infected-Recovered (SEIR) model, the growth of the outbreak is only asymptotically exponential (Ma, 2020).**
> **Nonetheless, we demonstrate that SEIR dynamics have only a minor potential impact on the coefficients recovered by using our empirical approach in this context. In Extended Data Figs. 8 and 9 we present results from a simulation exercise which uses Equations 9-11, along with a generalization to the SEIR model (Ma, 2020) to generate synthetic outbreaks (see Supplementary Methods Section 2). We use these simulated data to test the ability of our statistical model (Equation 7) to recover both the unimpeded growth rate (Extended Data Fig. 8) as well as the impact of simulated policies on growth rates (Extended DataFig.**
> **9) when applied to data generated by SIR or SEIR dynamics over a wide range of epidemiological conditions."**

Some sensitivity analyses and cross-validation also has to be presented for the model with actual data. Are the results sensitive to the grouping of policies, to an altered model form (e.g. SEIR instead of SIR), to a change in the lag (2 weeks v. 1.5 weeks)?

Our reply: To address this concern, we have introduced new analyses to explore the sensitivity of the model to an altered model form (SEIR instead of SIR), a change in the lag, the groupings of policies, and withholding of data. All of these new robustness checks are described in response to comments above (testing the effects of policy grouping and withholding blocks of data are described in reply to the reviewer's second comment).

Minor Comments

Page 13: "This approach non-parametrically accounts for arbitrary forms of spatial auto-correlation or systematic misreporting in regions of a country on any given day (it

generates larger estimates for uncertainty than clustering by i)." Can the authors clarify this for a broader audience?

Our reply: To address this concern, we have added additional text explaining both the mechanics and importance of the procedure we implement. The text now states:

> **"This approach allows the covariance in ε_{cit} across different locations within a country, observed on the same day, to be nonzero. Such clustering is important in this context because idiosyncratic events within a country, such as a holiday or a backlog in testing laboratories, could generate nonuniform country-wide changes in infection growth for individual days not explicitly captured in our model. Thus, this approach non-parametrically accounts for both arbitrary forms of spatial auto-correlation or systematic misreporting in regions of a country on any given day (we note that it generates larger estimates for uncertainty than clustering by *i*)."**

Page 14—can the authors provide further justification for the difference in policy implementation for Iran requiring a dummy variable. It seems ad hoc. Is it something special about an authoritarian theocracy? How is policy implementation different and what's the evidence base? Why shouldn't China differ (they are authoritarian)? Is it cultural compliance? Why not a dummy interactive variable for all countries?

Our reply: We apologize for any confusion due to our exposition. Throughout all of our analysis, the effect of policies on infection growth rates are estimated separately (i.e., using a separate regression) in each country. This is mathematically equivalent to interacting a dummy variable for each country with the model for that country, and

estimating a single regression using all of the data simultaneously. Thus we think our current approach agrees with the reviewer's suggestion.

However, it is possible that the reviewer is referring to an interaction term applied to policies in Tehran within the original Iran model. We originally created Tehran-specific policy variables because the national policies enacted in early March 2020 appeared to have a heterogenous effect in Tehran. Per the reviewer's suggestion, and updated data, we no longer treat Tehran separately.

The discussion of how the regression model (Eq. 7) is robust to systematic bias on page 14 is not as important as non-systematic biases arising from differences in testing practices in space and time. The authors address the latter extensively in the supplementary methods, which is great, but some mention of this issue in the main methods section should be provided after the discussion of systematic bias.

Our reply: To address this concern, we now state in the introduction

> **"Our econometric approach ... is robust to systematic under-surveillance specific to each subnational unit..."**

to note robustness to non-systematic biases across space. We then also state in the discussion

> **"Our analysis accounts for documented changes in the availability of and procedures for testing for COVID-19 as well as differences in case-detection across locations; however, unobserved trends in case-detection could affect our results (see Methods)"**

where we also describe the results of a new analysis, shown in new Extended Data Figure 2, in which we estimate the potential impact of trends in case-detection. The discussion now states:

> **“[O]ur analysis of estimated case-detection trends (Extended Data Fig. 2) suggests that the magnitude of this potential bias is small, elevating our estimated no-policy growth rates by 0.022 (6%) on average.”**

Where these calculations are based on results from a recent study by the Centre for Mathematical Modeling of Infectious Diseases[14] that computes country-specific and time-varying rates of case detection.

Finally, we substantially expand our discussion of these issues in the Methods section (See Lines 566 - 582) to separately and specifically describe time-invariant and time-varying under-reporting, so readers may clearly understand the distinction. This section explains and derives the robustness of our approach to the former but not the latter. To do this, we introduce Equation 8, describing the magnitude of the potential bias with time-varying under-reporting, which we try to characterize in the new Extended Data Figure 2.

We thank the referee for their careful and thoughtful review of the manuscript. We believe many of these comments led to additions to the manuscript that have significantly improved its quality.

[14] Russell, T. W. et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine (2020). https://cmmid.github.io/topics/covid19/371severity/global_cfr_estimates.html. Accessed:

2020-04-09.

Referee #4

Referee comments in italics. **[Author] reply in bold**.

The authors present an analysis of COVID-19 interventions. Measuring the impact of control measures is an important question, and the study provides a good summary of local-level measures in multiple locations. However, it was not clear that the study appropriately accounts for delays and uncertainty, which may influence the strength of the results.

Main comments:

- *The authors note method "is robust to systematic under-surveillance; and it accounts for changes in procedures to diagnose positive cases". However, it wasn't clear how they handled changes in case definition (e.g. https://www.medrxiv.org/content/10.1101/2020.03.23.20041319v1). Such changes could bias transmission estimates upwards - it's likely that the estimated doubling time of two days influenced by this could be an artefact of these changes in testing/definitions.*

Our reply: Changes in case definition generally appear as discontinuous increases in cases (e.g. the February 5 change in China is clear in Figure 1). Because our independent variable is the first difference in log(cases), these discontinuities appear as solitary spikes in the time series due to the one-time change in the level of cases. To address this issue, we compiled data for all documented changes in case definition and encode these events as dummy variables. This allows the magnitude of each spike to be estimated flexibly, absorbing this variation in a nuisance parameter, so that it does not bias our parameters of interest. This is a standard procedure for accounting for breaks in time-series when the time of the break is known but the magnitude of the break is unknown. This was described in the original text of the Methods under the "Estimation" section:

> **"We also include a separate single-day dummy variable each time there is an abrupt change in the availability of COVID-19 testing or a change in the procedure to diagnose positive cases. Such changes generally manifest as a discontinuous**

> **jump in infections and a re-scaling of subsequent infection rates (e.g., See China in Figure 1), effects that are flexibly absorbed by a single-day dummy variable because the dependent variable is the first-difference of the logarithm of infections. We denote the vector of these testing dummies μ."**

However, the reviewer's comment makes us aware that we did not explain that we had collected data on changes in case definitions anywhere in the main text, which may have made our accounting for these jumps less clear to readers. Thus, to address this, the introduction now explains

> **"We compile publicly available subnational data on daily infection rates, changes in case definitions, and the timing of policy deployments, ..."**

And the "Data Collection and Processing" section of the Methods section of the main text now states:

> **Epidemiological, case definition/testing regime, and policy data for each of the six countries in our sample were collected from a variety of in-country data sources,**
> **…"**

- *p5: "25.23% per day (p< 0.05)" It's not clear what p-values represent here. Surely there is also some uncertainty in the quoted growth estimate?*

Our reply: To clarify, we have replaced the p-values in this paragraph with standard errors for the estimated value.

- *"At the time of writing, the minimum susceptible population fraction in any of the administrative units analyzed is 99.4% of the total population (Lodi, Italy: 1,445 infections in a population of 230,000)." China (e.g. PMID 32171059 suggests 95% susceptible at the end of Jan). It is also important to distinguish between confirmed cases (which are reported) and infections (which are generally unknown unless there is a serosurvey); the two may be very different.*

Our reply: The reviewer is correct that it is important to distinguish between confirmed cases and total infections (which are unobserved). Much new evidence has emerged on the gap between these two values since our original submission. To address this issue, we obtained country-specific and time-varying estimates of case detection rates from the Centre for Mathematical Modeling of Infectious Diseases[15]. Throughout our analysis, whenever computing susceptible populations, we now rescale daily confirmed cases by $1/P_{ct}$, where P_{ct} is the estimated probability that a true case is detected and confirmed in country *c* on day *t*. This adjustment is now detailed in the Methods and Supplementary Methods, and it substantially affects the results presented in Figure 4 because it alters transmission dynamics when the susceptible population declines. However, we note that we continue to report the fraction of cases that one would expect to be "confirmed" as we believe it is important to maintain consistent units throughout the paper.

Also, we have updated (based on new data) and corrected the sentence above referenced by the reviewer to reflect these adjustments. The sentence now reads:

> **"After correcting for estimated rates of case-detection, we compute that the minimum susceptible population in any of the administrative units in our sample is approximately 78.0% of the total population (Cremona, Italy: roughly 79,000 total infections in a population of 360,000) and 86% of administrative units across all six countries would likely be in a regime of uninhibited exponential growth (susceptible fraction of population>95%) if policies were removed on the last date of our sample."**

We are grateful to the reviewer for drawing our attention to this critical issue.

- *Figure 2: It is notable that the lockdown in France seems to have little effect. Could this be because insufficient time has passed for the measures to have an effect? If so, it could be misleading to present evidence showing little effect. This potential bias can be seen in the*

[15] Russell, T. W. et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine (2020). https://cmmid.github.io/topics/covid19/371severity/global_cfr_estimates.html. Accessed:

2020-04-09.

estimates for China, which appear to become more effective over time - in reality this may just be a transient that the model has not yet accounted for.

Our reply: The referee is correct that in the original submission, insufficient time had passed since the lockdown in France for the policy to have a clearly detectable effect. In the revised manuscript, we include the latest available data in France (up to March 25, 2020). These updated data do indicate that the lockdown in France has slowed infection growth rates by a statistically significant -0.05 per day.

To explicitly address the concern about transient effects, and to avoid confusion among readers, we have expanded our discussion of this issue in the main text. The fifth paragraph of the Results section now reads:

> **In China, only two policies were enacted across 116 cities early in a seven week period, providing us with sufficient data to empirically estimate how the effects of these policies evolve over time without making any assumptions about the timing of these effects (see Methods and Fig. 2b). We estimate that the combined effect of these policies significantly reduced the growth rate of infections by -0.14 (SE=0.031) in the first week immediately following their deployment (also see Extended Data Fig. 5a), with effects doubling in the second week to -0.30 (SE=0.040), and stabilizing in the third week at -0.34 (SE=0.036). In other countries, we lack sufficient data to estimate these temporal dynamics explicitly and only report the average pooled effect of policies for all days following their deployment (see Methods). If other countries were to exhibit transient responses similar to that observed in China, we would expect effects in the first week following deployment to be smaller in magnitude than the average effect for all post-deployment weeks. Below we explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number of days, but we do not find evidence that this systematically improves model fit (Extended Data Fig. 5b).**

- *The authors account for delays resulting from the incubation period by lagging impact by a week, but how could delay from symptom onset to reporting influence results? It seems like it would be better to account for the distribution of delay here, rather than simply lagging by a week, which does not account for the full range of uncertainty.*

Our reply: We believe that our original analysis was in line with the reviewer's suggestion (i.e. not assuming a specific lag length), and that our original exposition was unclear. We have revised the text to provide greater clarity on this issue (e.g. see paragraph above; also, we believe the 1-week lag interpretation resulted from some unclear language that has since been removed). Since we do not know what the correct lag structure is, we "allow the data to speak" when examining data from China, imposing no assumptions on the structure of the lag. Unfortunately, we do not have sufficient data to implement this flexible approach in other countries, and thus simply report an estimated average treatment effect of each policy (not assuming any lag).

We have also clarified this in the Methods section when describing the parameters of interest, immediately after presenting the estimating equation:

> **"… θ is the average effect of the policy on growth rate *g* over all periods subsequent to the policy's introduction, thereby encompassing any lagged effects of policies."**

Finally, to help clarify multiple issues related to the timing of policy effects, we have created a new subsection on Transient Dynamics at the end of the Econometric Analysis section of the Methods. This section details the distributed lag model estimated in China, an events study examining 36 Chinese cities where we have a substantial quantity of pre-policy data, and a robustness check (suggested by Reviewer #3) where we re-estimate the model using various fixed lag lengths.

"Results are marginally statistically significant for France (p< 0.09)". Terms like "marginally statistically significant" are not rigorous or particularly helpful. It would be better to just report the effect size and p-value, so reader can evaluate the strength of evidence.

Our reply: We have removed this statement from the caption of Figure 2. As in the original submission, effect sizes and 95% confidence intervals for all estimates are reported in Figure 2 of the main text. We now also report standard errors and p-value thresholds in Supplementary Table 3. (We also note that using updated data, the offending p-value is now estimated to be <0.001).

- *Thank you for providing the underlying data and code for the analysis.*

Our reply: We thank the referee for their careful and thoughtful review of the manuscript. We believe many of these comments led to additions to the manuscript that have significantly improved its quality.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

I am satisfied with the thorough response to my original comments.

Referee #2 (Remarks to the Author):

Thank you for the opportunity to read the revision of manuscript 2020- 03-04486B, entitled “The Effect of Large-Scale Anti-Contagion Policies on the Coronavirus (COVID-19) Pandemic” by Hsiang and co-authors.

I am surprised by the authors’ revision and response. They have asserted they analyze natural experiments without providing a developed, substantive argument for why they are exogenous. I am more familiar with the standard for causal inference with observational data in economics journals. I doubt this assertion of exogeneity would fly at a top 5 economics journal. Publishing the manuscript in its current form will lead many of the economists who read it to infer that top science journals have a lower bar for causal inference than economics journals, which seems backwards to me.

The authors’ assertion also seems needless to me. I think a more cautiouslyinterpreted analysis should also be published in a top science (or economics) journal (as stated in my previous report), particularly given the importance of the question and timeliness of empirical results.

1. The authors’ causal argument rests primarily on functional form assumptions regarding infection growth rates. The authors now state in the main text: “a necessary condition for our estimates to be treated as the causal effect of these policies is that the timing of policy deployment is independent of infection growth rates” and that “decision-makers have tended to deploy policies in response to the count of total infections in their locality, rather than their growth rate”.

Insofar as infections are concerned, the authors seem to believe policy makers are uniquely obsessed with the level rather than the change in levels. This does not seem to be how, for example, Governor Cuomo describes the progression of COVID in NewYork. For example, on April 19 he stated:

“If the data hold and if this trend holds, we are past the high point and all indications at this point are that we are on the descent,” Cuomo added. “Whether or not the descent continues depends on what we do, but right

now are on the descent."
-The Hill

2. I had trouble with the following statement on page 10 of manuscript:

The policies we analyze are unlikely to have been deployed in reaction to or anticipation of changes in growth rates, since epidemiological guidance to decision-makers explicitly projected constant growth rates in the absence of anti-contagion measures.

(a) If policy makers really knew that growth was exponential absent policy, why are they obsessed with levels and ignoring the growth rate or recent changes to it?
(b) If policymakers did know that growth was exponential absent policy, wouldn't they have started policies earlier than they actually did? Why did President Trump wait to act if he thought growth would be exponential?

3. The authors also note that: "pre-policy infection growth rates in this context should be constant". Why not test this prominently?

4. Erik Brynjolfsson (MIT Sloan) posted this event study figure on Twitter:

[Figure redacted]

This suggests that social-distancing behaviors in the US were changing markedly just *before* closures were ordered by US states. As changes in social distancing would impact changes in infection rates and these were occurring just before closures, assessing the ordered closure's impact on subsequent growth in cases just isn't so straightforward. If the above figure is correct, it also argues against authors' claim that "pre-policy infection growth rates in this context should be constant".

5. The new paragraph in the manuscript also mentions policies may respond to "outbreaks in other regions or countries" or "based on other arbitrary and exogenous factors, such as closing schools on a Monday or after Spring Break". This seems quite a promising avenue but is not developed. For example, can we use neighboring policies as an instrumental variable for local COVID policies? Or their tendency to begin on a Monday and not a Friday? How much of the variation in COVID policies do these more exogenous factors explain? Merely "some" or 2 "most"? Without knowing it's most (or better still all), causal inference seems premature.

6. Policy decisions are not analogous to volcanic eruptions [Proctor et al., 2018] or typhoons [Hsiang, 2010] because these policy decisions are based in large part on the local progression of COVID. In the manuscript, this local progression is both a right-hand side variable (through policy) and the regression outcome variable. Finding a relationship here isn't super surprising or informative absent a natural experiment in policies. Yes, we do think the timing and location of a typhoon or volcanic eruption is also orthogonal to first differences in agricultural yields, economic growth, etc. Why? In large part because volcanoes, etc. are not appreciably affected by agricultural yields, etc. which is not the case for policy decisions and local COVID infections here.

"In the language of the natural experiment, the treatment and control populations in these analyses are not comparable units, so we cannot infer whether a climatic treatment has a causal effect or not" [Hsiang et al., 2013]. It remains unclear to me why we should believe here that early-adopting localities are comparable to late-adopting ones.

References

Solomon M. Hsiang. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of Sciences*, 107(35):15367–15372, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1009510107. URL https://www.pnas.org/ content/107/35/15367.

Solomon M. Hsiang, Marshall Burke, and Edward Miguel. Quantifying the influence of climate on human conflict. *Science*, 341(6151), 2013. ISSN 0036- 8075. doi: 10.1126/science.1235367. URL https://science.sciencemag. org/content/341/6151/1235367.

Jonathan Proctor, Solomon Hsiang, Jennifer Burney, Marshall Burke, and Wolfram Schlenker. Estimating global agricultural effects of geoengineering using volcanic eruptions. *Nature*, 560, 08 2018. doi: 10.1038/ s41586-018-0417-3.

Referee #3 (Remarks to the Author):

This is a very nice revision--comprehensive, well articulated. I commend the authors. I think the manuscript should be accepted. Below, I append a few small comments.

The caption for Extended figure 8 mislabels latency as 1/gamma and infectious period as 1/sigma—these labels are reversed (i.e. latency is 1/sigma, infectious period is 1/gamma, sigma◊infinity is SIR)

Extended figure 7 is helpful. However, I think a \gamma of 0.18 is more sensible for use in Figure 4 of the main text. As the authors note: 'individuals are coded as "recovered" when they no longer test positive for COVID-19, whereas in the classical SIR model this occurs when they are no longer infectious.' The higher \gamma (shorter infectious period) is more realistic.

I appreciate the authors adjustment for undocumented infections, but it is likely incomplete. Russell et al. work from symptomatic cases only; they assume a CFR of 1.4% and they then use lag-to-death rates in known countries to estimate how deviations from 1.4% are indicative of under-reporting. What about the asymptomatic infections? The IFR is estimated as 0.4-0.7%--see Verity et al.—that would provide a better estimate of susceptible depletion. However, for the purposes of your presentation, which rightfully doesn't emphasize precise numbers, this likely doesn't matter.

I greatly appreciate the synthetic tests and exploration of the effects of differing \gamma and \sigma. This, to my mind, strengthens the paper, as there are non-identifiable features (e.g. g=\beta - \gamma) embedded in the Equation 7 model.

Jeffrey Shaman

Referee #4 (Remarks to the Author):

Thanks to the authors for addressing my comments satisfactorily.

ADDITIONAL PRODUCTION NOTES:

1. Please provide the article file as a doc./docx. file
2. The title is over budget at 88 characters with spaces. Please shorten so that it does not exceed 75 characters with spaces
3. Please resupply the main figures at individual editable vector files separate from the main text in one of the following formats: AI, PS, EPS, PSD, PPT, PDF, or CDR (up to version 8). The legends may however remain in the article file
4. The text size in figures 1 and 2 need to be increased to 5pt when sized to the dimensions in the stats
5. Figure 4 is slightly over the 17cm height budget. Please ask the authors to ensure that their main figures do not exceed 17cm (18cm in the stats) in height
6. Please ask the authors to resupply their ED figures as individual files in one of the following formats: JPEG, EPS, or TIF. The legends should be added to the article file
7. Authors need to confirm permissions for the maps used in their main figures

Author Rebuttals to First Revision:

Referee in italics. **[Author} reply in bold**.

Referee #2

I am surprised by the authors' revision and response. They have asserted they analyze natural experiments without providing a developed, substantive argument for why they are exogenous. I am more familiar with the standard for causal inference with observational data in economics journals. I doubt this assertion of exogeneity would fly at a top 5 economics journal. Publishing the manuscript in its current form will lead many of the economists who read it to infer that top science journals have a lower bar for causal inference than economics journals, which seems backwards to me.

The authors' assertion also seems needless to me. I think a more cautiously-interpreted analysis should also be published in a top science (or economics) journal (as stated in my previous report), particularly given the importance of the question and timeliness of empirical results.

<u>Our reply:</u> To address this concern, we have gone back through our text and softened the language throughout. For example, we have removed reference to the events we analyze as "natural experiments", deleting the clause "treating each subnational administrative unit as if it were in a natural experiment."

We also revised the language in the paragraph in the main article to more precisely describe the basis for our identification strategy:

> **"Our analysis measures changes in local infection growth rates associated with changes in anti-contagion policies. A necessary condition for this association to be interpreted as the plausibly causal effect of these policies is that the timing of policy deployment is independent of infection growth rates (Angrist & Pischke, 2008). This assumption is supported by established epidemiological theory (Anderson & May, 1992; Chowell et al, 2016; Ma, 2020) and evidence (Nishiura et al, 2010; WHO 2014), which indicate that infections in the absence of policy will grow exponentially early in the epidemic, implying that pre-policy infection growth rates should be constant over time and therefore uncorrelated with the timing of policy deployment. Further, scientific guidance to decision-makers early in the current epidemic explicitly projected constant growth rates in the absence of anti-contagion measures, limiting the possibility that anticipated changes in natural growth rates affected decision-making (Ferguson et al, 2020; Flaxman et al, 2020; Lourencco et al, 2020; Maier & Brockman, 2020). In practice, policies tended to be deployed in response to high total numbers of cases (e.g. in France) (Ministère de la santé, 2020), in**

> response to outbreaks in other regions (e.g. in China, South Korea, and Iran) (Tian et al, 2020), after delays due to political constraints (e.g. in the US and Italy), and often with timing that coincided with arbitrary events, like weekends or holidays (see Supplementary Notes for detailed chronologies)."

Additionally, the paragraph in our Discussion on potential sources of confounding now states:

> "It is also possible that changing public knowledge during the period of our study affects our results. If individuals alter behavior in response to new information unrelated to anti-contagion policies, such as seeking out online resources, this could alter the growth rate of infections and thus affect our estimates. If increasing availability of information reduces infection growth rates, it would cause us to overstate the effectiveness of anti-contagion policies. We note, however, that if public knowledge is increasing in response to policy actions, such as through news reports, then it should be considered a pathway through which policies alter infection growth, not a form of bias.
> Investigating these potential effects is beyond the scope of this analysis, but it is an important topic for future investigations. "

We have also introduced a new paragraph in the Methods / Estimation section to familiarize readers from economics with the epidemiological background on exponential growth of infections, a concept that is important for our identification strategy (discussed below). The new paragraph states:

> "A necessary condition for unbiased estimates is that the timing of policy deployment is independent of natural infection growth rates (Angrist & Pischke, 2008) a mathematical condition that should be true in the context of a new epidemic. In established epidemiological models, including the standard SIR model above, early rates of infection within a susceptible population are characterized by constant exponential growth. This phenomenon is well understood theoretically, (Chowell et al, 2016; Anderson & May, 1992; Kermack, & McKendrick, 1927) has been repeatedly documented in past epidemics (WHO Ebola Response Team, 2014; Nishiura et al, 2010; Mills et al, 2004) as well as the current COVID-19 pandemic (Ma, 2020; Muniz et al, 2020) and implies constant infection growth rates in the absence of policy intervention. Thus, we treat changes in infection growth rates as conditionally independent of policy deployments since the correlation between a constant variable and any other variable is zero in expectation."

We have also developed a new section of the Supplementary Notes titled "Chronology of Policy Implementation", which provides fully referenced accounts and context for policy deployments in each of the six countries, as is the standard in economics journals, so that readers may understand and independently evaluate the extent to which identifying assumptions might be violated. The first two paragraphs of this nine page-long section summarize its discussion:

> **"In the main article, we analyze how infection growth rates change when anti-contagion policies are deployed. In order for these estimated associations to be interpreted as plausibly causal, it must be the case that changes in policy are not correlated with changes in infection growth rates that would have occured in the absence of policy actions (Angrist & Pischke, 2008).[1] For locations and periods where policy was actually deployed, it is not possible to directly observe how growth rates would have changed in the absenceof policy actions (Holland, 1986).[2] In general, it is well-established that in the absence of policy, early rates of infection within a susceptible population follow almost-perfect exponential growth (Mena-Lorcat, 1992[3]; Brauer & Chowell, 2012 [4]; WHO Ebola Response Team, 2014 [5]; Mills et al, 2004[6]; Anreason et al, 2008[7]; Nishiura et al, 2010[8]; Towers et al, 2014[9]; Anderson & May, 1992[10]). This fact would suggest that in the absence of policy, infection growth rates would be constant (Ma, 2020[11]; Muniz-Rodriguez et al, 2020[12]) and, therefore, changes in the infection growth rate could not be correlated with the timing of policy deployments (since correlation with a constant variable is zero). However, it is nonetheless also worthwhile to consider how the timing of policy actions were determined in practice. In this section, we provide a basic accounting of policy deployment decisions in each of the six countries we analyze, focusing on the timing of major events and reported motivations for notable policy deployments. These accounts are intended to familiarize readers with the general decision-making context of each country, although they are not exhaustive and are not intended to capture the full complexity of decision-making in the ongoing pandemic.**

[1] Angrist & Pischke, 2008

[2] Holland, 1986

[3] Mena-Lorcat, 1992

[4] Brauer & Chowell, 2012

[5] WHO Ebola Response Team

[6] Mills et al, 2004

[7] Andreason et al, 2008

[8] Nishiura et al, 2010

[9] Towers et al, 2014

[10] Anderson & May, 1992

[11] Ma, 2020

[12] Muniz-Rodriguez et al, 2020

> **As stated in the main text, policies were generally not initially deployed with any reference to observed or anticipated natural changes in growth rates. In general, policies tended to be deployed in response to high total numbers of cases or to outbreaks in other regions, after long delays due to political constraints, and often with timing that coincided with arbitrary events, like the start and end of the week, or holidays. For example, epidemic planning in France explicitly tied policy actions to case-counts in specific regions, and epidemiological guidance in the US explicitly recommended policy actions depending on when thresholds in total case counts were passed. In South Korea, policy actions across the country were initially triggered by an idiosyncratic outbreak in a Shincheonji Church in Daegu, and in Iran they were triggered by an outbreak in Qom associated with a religious pilgrimage. In China and Italy, nation-wide policy actions were deployed relatively swiftly by the central government in response to a regional outbreak, with limited variation across locations that could be correlated with changes in local growth rates. In China, Italy and the US, policies were tended to be deployed during the weekend or on a Monday or Friday, with the preferred day of the week varying across these countries."**

1. *The authors' causal argument rests primarily on functional form assumptions regarding infection growth rates. The authors now state in the main text: "a necessary condition for our estimates to be treated as the causal effect of these policies is that the timing of policy deployment is independent of infection growth rates"*

Our reply: **The reviewer is correct that a functional form assumption provides us with confidence that the timing of policy deployments is likely to be independent of changes in the growth rate of infections (early on in the epidemic). Unlike many other potential functional form assumptions, the particular phenomenon of constant infection growth rates, which we rely on, is a fundamental and well-established result in modern epidemiology -- one that has been widely understood and repeatedly tested**[13, 14, 15, 16, 17, 18, 19, 20, 21].

[13] Mena-Lorcat, 1992

[14] Brauer & Chowell, 2012

[15] WHO Ebola Response Team

[16] Mills et al, 2004

[17] Andreason et al, 2008

[18] Nishiura et al, 2010

[19] Towers et al, 2014

[20] WHO Ebola Response Team

[21] Anderson & May, 1992

As an example of how established this concept is in epidemiology, the widespread use of *R_0* (i.e. the "basic reproductive number") to summarize the infectiousness of different diseases relies on the assumption that infections from all of these diseases grow exponentially during their epidemic phase.

Such unmitigated exponential growth is also well-documented by epidemiologists in recent epidemics. For example, as documented by Nishiura et al. (2010)[22], the 2009 H1N1 epidemic followed two different exponential functions during two separate periods of unmitigated contagion:

[Figure redacted]

Similarly, infections during the 2014 Ebola outbreak in West Africa followed exponential growth in each country, such that the WHO Ebola Response Team wrote (2014, *New England Journal of Medicine*):

> **"As of September 14, the doubling time of the epidemic was 15.7 days in Guinea,**
>
> **23.6 days in Liberia, and 30.2 days inSierra Leone .... We estimate that, at the**
>
> **current rate of increase, assuming no changes in control efforts, the cumulative number of confirmed and probable cases by November 2 (the end of week 44 of the epidemic) will be 5740 in Guinea, 9890 in Liberia, and 5000 in Sierra Leone, exceeding 20,000 cases in total (Figure 4)"**

[22] Nishiuraetal,2010

[23] Nishiuraetal,2010

[Figure Redacted]

Similar descriptions characterizing constant exponential growth during the early outbreak of COVID-19 are present throughout the literature, appearing in many studies that we cite in our manuscript[25, 26, 27, 28] and many epidemiological studies of COVID-19 make this assumption. For example, this exponential behavior is embedded in the major models currently used by teams around the world to forecast the growth of COVID-19 infection growth.[29]

Thus, we do not believe the functional form assumption that we use is a strong assumption. Rather, it is the established standard in the literature that has been widely validated, and any alternative assumption would be inconsistent with existing knowledge in the field of epidemiology.

...and that "decision-makers have tended to deploy policies in response to the count of total infections in their locality, rather than their growth rate".

Insofar as infections are concerned, the authors seem to believe policy makers are uniquely

[24] WHO Ebola Response Team

[25] Maier & Brockmann, 2020

[26] Ferguson et al., 2020

[27] Ma, 2020

[28] Muniz-Rodriguez et al, 2020

[29] COVID-19 Forecasts | CDC

obsessed with the level rather than the change in levels. This does not seem to be how, for example, Governor Cuomo describes the progression of COVID in NewYork. For example, on April 19 he stated:

"If the data hold and if this trend holds, we are past the high point and all indications at this point are that we are on the descent," Cuomo added. "Whether or not the descent continues depends on what we do, but right now are on the descent." -The Hill

<u>Our reply:</u> To address this concern, we have now developed an entirely new section in the Supplementary Notes on the "Chronology of Policy Implementation", as mentioned above. This section aims to provide unfamiliar readers with a basic understanding of the policy context in each of the six countries we study. In particular, we focus on elements of the decision-making structure that influenced the timing of policy deployment.

For example, we explain that policy deployments in France were determined by the ORSAN (*organisation de la réponse du système de santé en situations sanitaires exceptionnelles)* plan. Originally designed to deal with the 2009 H1N1 epidemic, we now explain that the ORSAN plan "is a predefined, structured emergency plan used to manage epidemic responses in France" The plan explicitly triggers policy actions when a certain number and/or distribution of cases are recorded. For example, "Stage 2" of the plan was triggered when the number of confirmed cases within the country reached 100 cases, and "Stage 3" was triggered when the virus was detected in every region.

As another example, in this new section we now also provide a direct quote by Governor Cuomo in which he explains that he deploys policies in response to the levels of cases. The manuscript now states:

> **On March 9, 2020, during an interview with Katy Tur of MSNBC, when asked "What about schools? Are you considering closing them? What will it take to close them?," Cuomo responded:**
>
>> ***[w]e close them on a case-by-case basis...You have higher numbers in certain areas [of the state], lower in others. If the numbers are low, God bless. If the numbers are high, take action. As I mentioned before where we have a cluster in Westchester that has more cases than New York City. So how do you handle that hotbed, that hot spot, as they call it?***
>>
>> ***Closing schools, closing gatherings, etcetera, extraordinary efforts where you have higher density of cases.***

2. *I had trouble with the following statement on page 10 of manuscript:*

> *The policies we analyze are unlikely to have been deployed in reaction to or anticipation of changes in growth rates, since epidemiological guidance to decision-makers explicitly projected constant growth rates in the absence of anti-contagion measures.*

(a) *If policy makers really knew that growth was exponential absent policy, why are they obsessed with levels and ignoring the growth rate or recent changes to it?*

(b) *If policymakers did know that growth rate was exponential absent policy, wouldn't they have started policies earlier than they actually did? Why did President Trump wait to act if he thought growth would be exponential?*

Our reply: **We cannot know for certain why many leaders take the course of action they ultimately chose. However, we do know that many, if not all, key policy-makers were informed by experts that infections would follow an exponential trajectory in the absence of policy action. For example, widely read studies by the Imperial College COVID-19 Response Team (Ferguson et al, 2020) clearly stated that early infection growth rates would be exponential in the absence of policy action. Perhaps surprising to the reviewer, the same report advised that policy actions be triggered when certain infection *levels* were reached, stating that "suppression policies are best triggered early in the epidemic, with a cumulative total of 200 ICU cases per week being the latest point at which policies can be triggered [and still keep peak ICU demand below surge limits]." We agree with the reviewer that such advice may seem counterintuitive in some dimensions, and perhaps sub-optimal in an economic cost-benefit framework that maximizes social surplus. Nonetheless, these accounts indicate that decisions to deploy policy actions were very likely independent of expected changes to natural infection growth rates, since in the absence of policy, there essentially were no expectations that growth rates would change in the short term, and it was advised that policy actions should be triggered by total case counts.**

The above account is now included in the new section of the Appendix regarding "Chronology of Policy Implementation".

3. *The authors also note that: "pre-policy infection growth rates in this context should be constant". Why not test this prominently?*

<u>Our reply:</u> As described above, constant growth rates in the early pre-policy stages of a new epidemic is a known phenomenon in epidemiology[30, 31, 32, 33, 34, 35, 36, 37, 38] that has been tested so regularly that it is currently embedded in all the main models being used to project the COVID-19 outbreak[39]. However, we understand that our prior discussion of this issue did not convey this and have now added these citations to the main text in the Methods / Estimation paragraph quoted above.

Yet, unlike the H1N1 and Ebola outbreaks referenced above, there is currently insufficient data available to test this fundamental behavior in our sample -- largely as a result of (1) how quickly policies were deployed and (2) our quality control condition that observational units are tracked only after ten confirmed cases (both for data-reliability and because of integer-based issues when first differences are applied to the logarithm of small numbers). However, as discussed extensively in our prior review, other authors[40, 41] have fit exponential functions to early pre-policy data that we removed due to concerns regarding data reliability (e.g. Supplementary Table 2).

Below are all of the observations in our sample that satisfy our original quality control criteria and for which there are no policy treatments (after covariates are partialled-out). As can be seen, there is not a clear general trend in the growth rate, and the sample is so short that trends would be almost meaningless to estimate.
Thus, in this analysis we focus on estimating the simple average growth rate for each country during this period.

[Figure redacted]

[30] Mena-Lorcat, 1992

[31] Brauer & Chowell, 2012

[32] WHO Ebola Response Team

[33] Mills et al, 2004

[34] Andreason et al, 2008

[35] Nishiura et al, 2010

[36] Towers et al, 2014

[37] WHO Ebola Response Team

[38] Anderson&May,1992

[39] Best, R. and Boice, J. *Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree*. FiveThirtyEight (2020).

[40] Wu et al, 2020

[41] Li et al, 2020

4. *Erik Brynjolfsson (MIT Sloan) posted this event study figure on Twitter:*

[Figure redacted]

This suggests that social-distancing behaviors in the US were changing markedly just before closures were ordered by US states. As changes in social distancing would impact changes in infection rates and these were occurring just before closures, assessing the ordered closure's impact on subsequent growth in cases just isn't so straightforward. If the above figure is correct, it also argues against authors' claim that "pre-policy infection growth rates in this context should be constant".

Our reply: **We agree with the reviewer that if this figure were both accurate and broadly representative of all forms of citizen behavior and all policy deployments that we analyze, it would suggest that private actions might partially confound our analysis. However, we have found that this figure is not representative and the analysis presented in the figure is strongly confounded by policy actions that occurred before the restaurant closure date presented (notably, these earlier policies are present and fully accounted for in our analysis). What we understand about this figure is as follows.**

Prof. Erik Brynjolfsson tweeted this figure on April 16, although he did not create it.

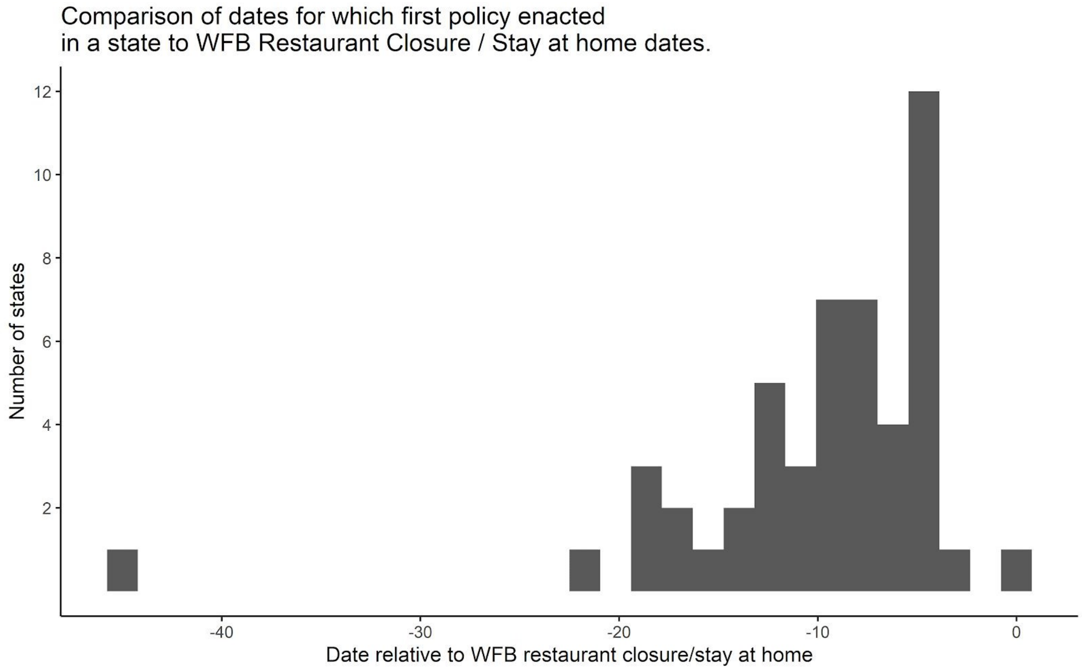
Charles Fein Lehman created the figure for an article posted April 15 on the Washington Free Beacon website. When Lehman tweeted about the figure that same day, he explained that he created it to show that government policies were not needed to get people to social distance.

Lehman did not explain how he created the figure or provide enough information to recreate it in the originally published article. That article did not describe what data source he used for policy deployments. All that was stated in the online article was:

> "[e]ven before state governments shut down their economies, Americans were shrinking from commerce .... One major piece of evidence comesfrom
>
> data about restaurants, which were some of the first institutions targeted by coronavirus-related restrictions. Many states have ordered restaurants to close their dine-in service, with the timing of those closures varying between states. OpenTable, a popular online reservation service, has published
>
> state-level data on daily restaurant reservations as a percentage of reservations on the same dayof 2019 .... Those data show thatrestaurant
>
> reservations had declined precipitously in most states before restaurants were officially closed. On the day before closure orders, the median state had seen reservations fall by a whopping 73 percent. In some states, like Michigan and Georgia, reservations had already fully stopped before restaurants were officially closed. In short, patrons did not need the government to tell them to before they stopped eating out."

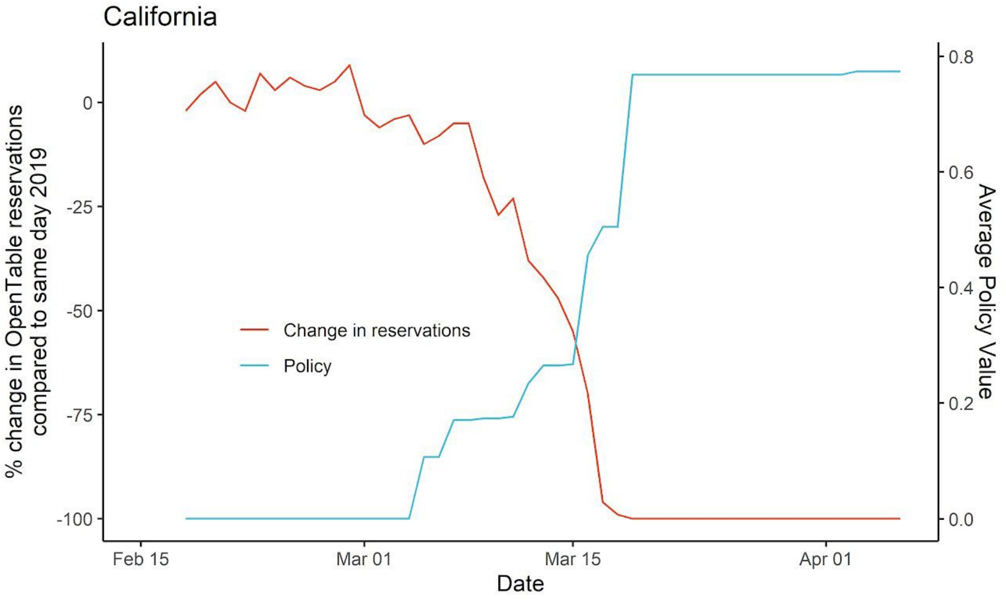
However Lehman's statement that restaurants were "some of the first institutions targeted by coronavirus-related restrictions" is not consistent with a comprehensive view of anti-contagion policies in the US. Many policy actions were widely deployed before the restaurant closures shown in the figure, and these earlier policy actions appear to confound the analysis presented in Lehman's figure.

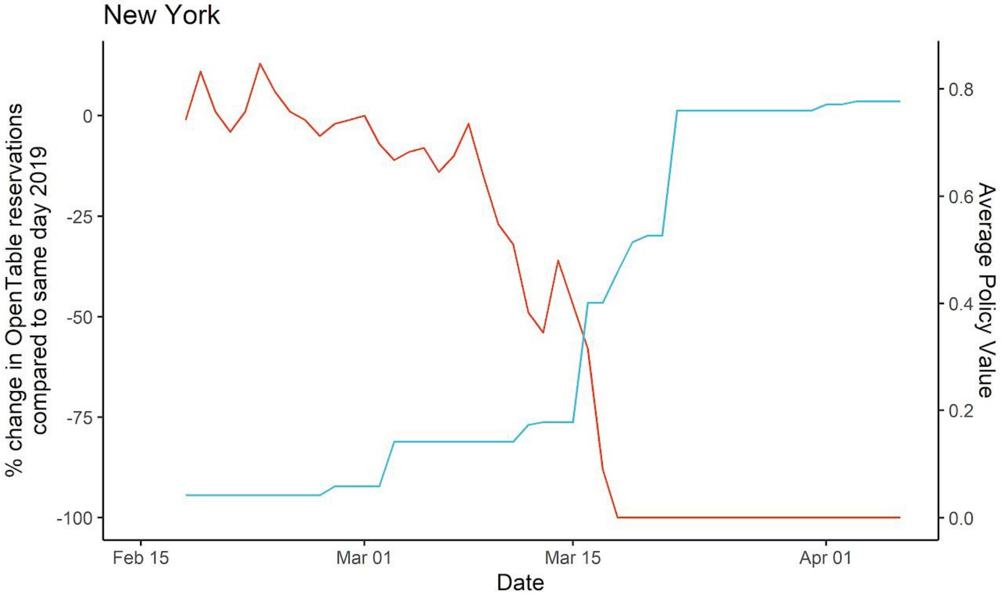
We contacted Lehman privately by email, and he shared the data and code he used to make the figure directly with us. This allowed us to compare Lehman's restaurant closure variable with our 1,194 policies in our US sample. We found that in 35 of the 36 states shown in Lehman's figure, there were substantial policy actions deployed *prior* to the restaurant closure event depicted in the figure (Vermont is the only exception, where closures on March 13 coincided with emergency declarations and prohibition on gatherings). The histogram below shows the average difference, in days, between the first policy in our sample is deployed and the date used by Lehman to create his figure:



Thus, the decline in reservations depicted was not occurring in the absence of policy actions, as suggested. Rather, it appears that the analysis in the figure is showing the effect of earlier policy actions.

To test this further, we obtained the OpenTable data cited in Lehman's article and linked it to our data. It appears that cancellation of dinner reservations corresponded with earlier policy enactments. For example, when we plot the average value of the 12 policies in our analysis with the OpenTable data in California and New York, we find that the strong drop in restaurant attendance is preceded by policy changes in our dataset:





Thus, we conclude that the analysis presented in the Tweeted figure should not be interpreted as an event study, that the figure does not provide reliable evidence that individuals dramatically altered their behavior before policies were deployed, and that the figure does not threaten the validity of our analysis. Instead, we interpret this figure as an example of why it was crucial in our analysis that we fully account for the large number of diverse and overlapping policies that have been deployed in each country, since failing to account for early policy actions would confound an analysis that focused only on a single type of policy action.

5. *The new paragraph in the manuscript also mentions policies may respond to "outbreaks in other regions or countries" or "based on other arbitrary and exogenous factors, such as closing schools on a Monday or after Spring Break". This seems quite a promising avenue but is not developed. For example, can we use neighboring policies as an instrumental variable for local COVID policies? Or their tendency to begin on a Monday and not a Friday? How much of the variation in COVID policies do these more exogenous factors explain? Merely "some" or "most"? Without knowing it's most (or better still all), causal inference seems premature.*

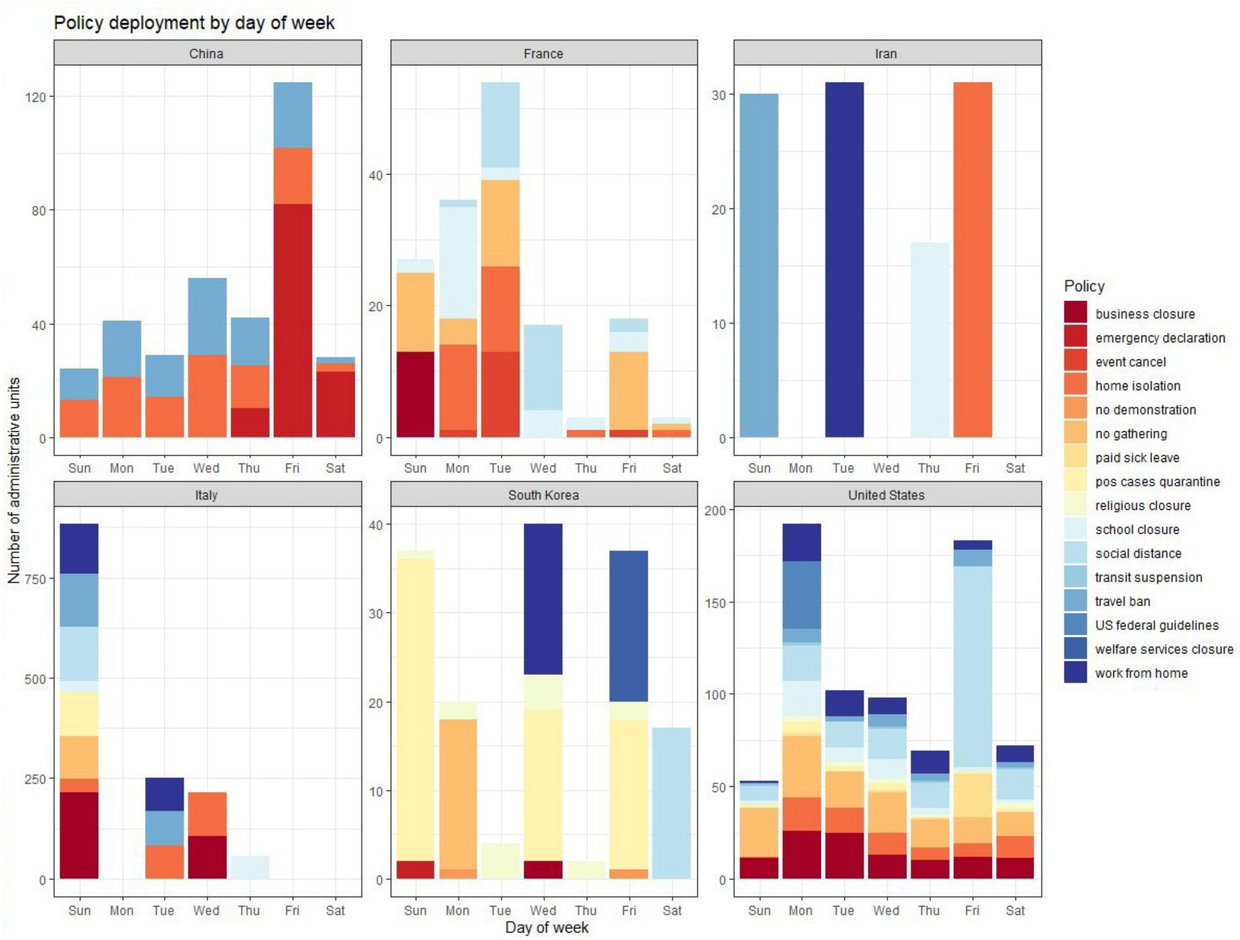
Our reply: To address this, we now report the fraction of policies deployed on weekends or weekend-adjacent days of the week. The new Appendix section on "Chronology of Policy Implementation" now notes in its introduction that:

> **"In China, Italy and the US, policies tended to be deployed during the weekend or on a Monday or Friday, with the preferred day of the week varying across these countries."**

And we explicitly report fractional values for these countries in their individual sub-sections, stating:

> **"A majority of all policies in Italy were deployed on a Sunday (62.7%)." "Policies across the US**
>
> **were deployed on all days of the week, but a**
>
> **disproportionate number were deployed on Mondays (25.0%) and Fridays (23.8%)**
>
> **relative to other days of the week (10.2% on average)."**
>
> **"A disproportionate fraction of policies were deployed on Fridays in China (36.2%)"**

For reference, the figure below displays the distribution of policies deployed in each country-specific sample at the administrative unit level by day-of-week. A chi-squared test on each country's observed distribution of policy deployments indicate that the observed distribution of policies across days of the week are statistically different from a uniform distribution (the six chi-squared test statistics ranged from 68 in South Korea to 3,028 in Italy, all with a p-values less than 0.0001).



We agree with the reviewer that a day-of-week instrumental variables analysis might be a useful avenue to explore in future work that tries to understand whether these results are affected by policy-independent information. While such an exploration is beyond the scope of this analysis, we point readers towards this question for future work in the Discussion section of the main article.

6. *Policy decisions are not analogous to volcanic eruptions [Proctor et al., 2018] or typhoons [Hsiang, 2010] because these policy decisions are based in large part on the local progression of COVID. In the manuscript, this local progression is both a right-hand side variable (through policy) and the regression outcome variable. Finding a relationship here isn't su- per surprising or informative absent a natural experiment in policies. Yes, we do think the timing and location of a typhoon or volcanic eruption is also orthogonal to first differences in agricultural yields, economic growth, etc. Why? In large part because volcanoes, etc. are not appreciably affected by agricultural yields, etc. which is not the case for policy decisions and local COVID infections here. "In the language of the natural experiment, the treatment and control populations in these analyses are not comparable units, so we cannot infer whether a climatic treatment has a causal effect or not" [Hsiang et*

> *al., 2013]. It remains unclear to me why we should believe here that early-adopting localities are comparable to late-adopting ones.*

<u>Our reply:</u> In the present analysis, inferences are not drawn from comparing early-adopting localities to late-adopting localities. Rather, consistent with the earlier studies referenced by the reviewer (Proctor et al 2018, Hsiang 2010, Hsiang et al 2013) location-specific fixed effects (intercept terms) are included in this analysis.

This term absorbs the effect of all constant location-specific factors that might otherwise make early- and late-adopting localities incomparable on average. Similar to these earlier studies, inference in the present analysis is drawn from comparing infection growth rates within a single location before and after policies are deployed. Thus, each population is only compared to itself over time, serving as a "control" (pre-policy) for itself after a policy "treatment" is deployed (notably, each location generally experiences a sequence of policy deployments). Since we only study the period for which policies are deployed, ending the sample before policies are lifted, there is no opportunity for changes in growth rate within a locality (resulting from the policy) to alter the status of treatment (which would be a source of reverse causality).

To address the reviewer's concern, we have made the source of comparisons more prominent for readers. The Introduction now states

> ***"We compare the growth rate of infections within hundreds of sub-national regions before and after each of these policies is implemented locally.***
>
> ***Intuitively, each administrative unit observed just prior to a policy deployment serves as the 'control' for the same unit in the days after it receives a policy 'treatment'."***

where the second sentence was previously in the Discussion.

Referee in italics. **[Author] reply in bold**.

Referee #3

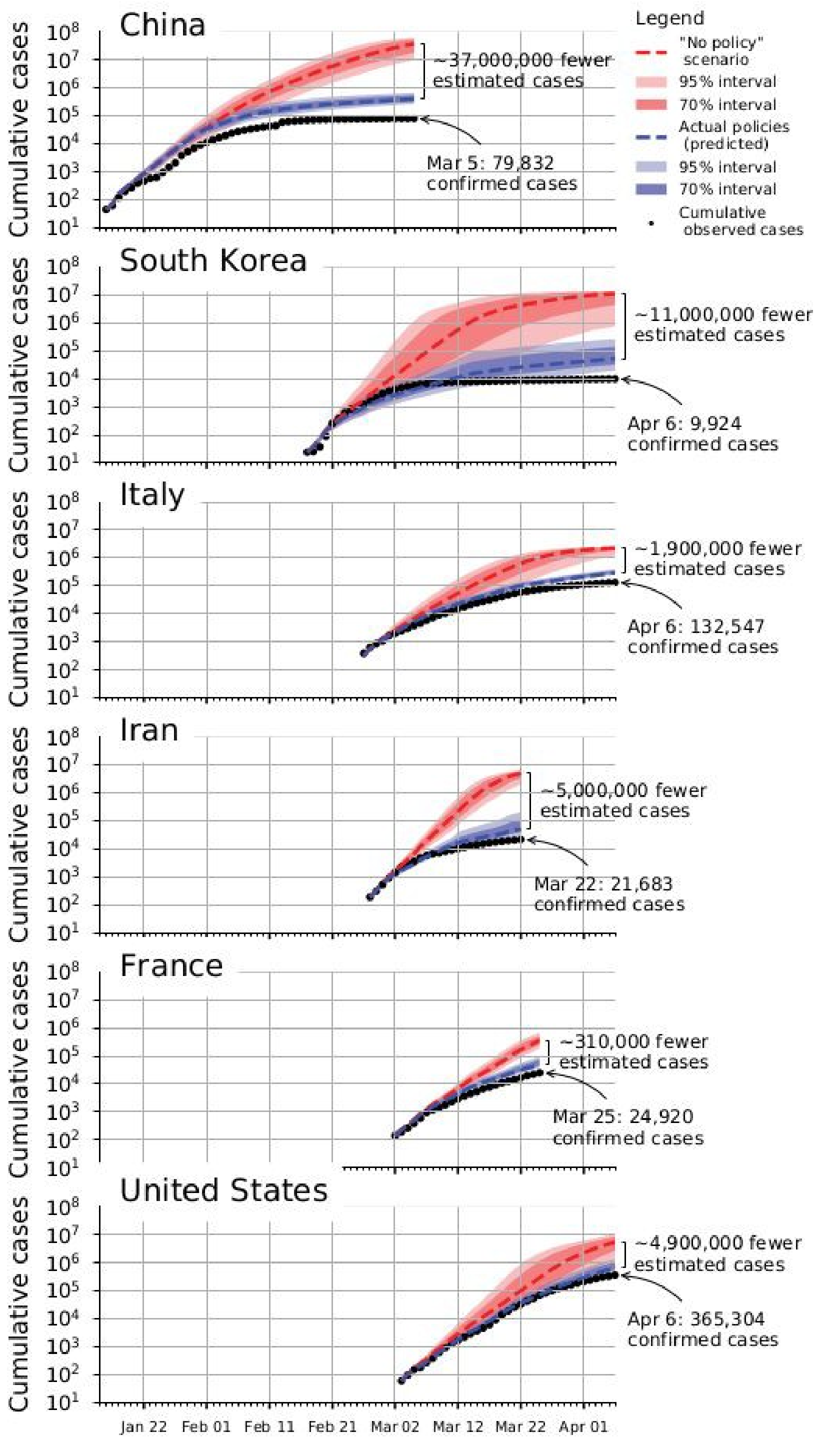
This is a very nice revision--comprehensive, well articulated. I commend the authors. I think the manuscript should be accepted. Below, I append a few small comments.

The caption for Extended figure 8 mislabels latency as 1/gamma and infectious period as 1/sigma—these labels are reversed (i.e. latency is 1/sigma, infectious period is 1/gamma, sigmainfinity is SIR)

Our reply: **This has been corrected, thank you for catching the error.**

Extended figure 7 is helpful. However, I think a \gamma of 0.18 is more sensible for use in Figure 4 of the main text. As the authors note: 'individuals are coded as "recovered" when they no longer test positive for COVID-19, whereas in the classical SIR model this occurs when they are no longer infectious.' The higher \gamma (shorter infectious period) is more realistic.

Our reply: **Following this suggestion, we computed a version of Fig. 4 using γ = 0.18:**



However we find that this simulation exhibits a poorer fit to the observed data (comparing dotted-black and dashed-blue lines). A higher value of γ causes the simulation to further overestimate cumulative confirmed cases at the end of our sample, as a higher value of γ implies higher rates of transmission. Given the lower value of γ

appears to more accurately represent the available data, we elect to retain the use of γ =

0.079 for the main results.

I appreciate the authors adjustment for undocumented infections, but it is likely incomplete. Russell et al. work from symptomatic cases only; they assume a CFR of 1.4% and they then use lag-to-death rates in known countries to estimate how deviations from 1.4% are indicative of under-reporting. What about the asymptomatic infections? The IFR is estimated as

0.4-0.7%--see Verity et al.—that would provide a better estimate of susceptible depletion. However, for the purposes of your presentation, which rightfully doesn't emphasize precise numbers, this likely doesn't matter.

Our reply: We thank the reviewer for pointing out that the CFR does not capture asymptomatic cases. We have now changed our main results to assume an IFR of 0.75% based on the central estimate from a recent meta-analysis[42] that includes the Verity et al. analysis. We also now show sensitivity to this parameter in a new Supplementary Table 6.

In the main text we now describe thisadjustment:

> **"To compute estimates of the proportion of infections confirmed, we adjust existing estimates to assume an infection-fatality ratio of 0.075% (Meyerowitz-Katz, G. & Merone, L., 2020)."**

We also describe the sensitivity analysis in Supplementary Methods section 1:

> **We also compute country-specific estimates of infection underreporting to improve the outbreak simulations in Figure 4. To produce these estimates, we use code from Russell et al. (2020) and substitute the assumed case-fatality ratio (a key parameter in their model) for an infection-fatality ratio (IFR) of 0.75% (Meyerowitz-Katz & Merone 2020). Their analysis produces country-specific estimates of infection underreporting, which we use to scale our estimates of confirmed cases to estimate total infections. We test the sensitivity of our main results to this IFR assumption in Supplementary Table 6, where we also show results for the total number of confirmed cases and infections avoided/delayed using IFR assumptions of 0.5% and 1%. The table shows approximately a 70% increase in the estimated number of confirmed cases avoided/delayed for a doubling in the IFR. However, the estimated number of infections is relatively stable with approximately a 15% decline in the estimated number of infections for a doubling in the IFR.**

[42] A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates

We also note that, after reflecting on the reviewer's comment that these unreported cases are important to account for, we now describe our implied estimates for *total* infections avoided/delayed alongside our originally reported estimates of avoided/delayed confirmed cases throughout the text, including in the abstract.

I greatly appreciate the synthetic tests and exploration of the effects of differing \gamma and

\sigma. This, to my mind, strengthens the paper, as there are non-identifiable features (e.g. g=\beta - \gamma) embedded in the Equation 7 model.

Jeffrey Shaman

Our reply: **We agree that the synthetic tests and the exploration of the effects of differing epidemiological parameters improved the paper and we thank the reviewer for suggesting these analyses.**