We thank the reviewers for their constructive feedback. We appreciate the comments that our "careful empirical study" **[R4]** and "sensitivity analyses [...] are of extreme importance" **[R2]**. Further, they are "critical to understand[ing] whether assumptions, data or both are providing evidence" **[R3]** about the effectiveness of different nonpharmaceutical interventions (NPIs) against COVID19 transmission. Given the importance and time-sensitivity of these results, and the minor criticisms raised by the reviewers, we hope that our clarifications below will allow the reviewers to increase their scores. Additionally, in line with reviewer comments, **we have run a number of additional experiments that will also be included in the camera-ready version**.

[R2] [R3] [R4] Contribution. Our work is the first that performs *structural* sensitivity analysis and compares the robustness of data-driven NPI effectiveness models. Our findings are policy relevant; the high sensitivity of the model used in [8], subsequently published in Nature, raises concerns (though the authors do not claim to distinguish individual NPI effects). A recent preprint (concurrent work to us) also finds [8] has high sensitivity [*Soltesz et al, On the sensitivity of non-pharmaceutical intervention models for SARS-CoV-2 spread estimation, 2020*]. We highlight that neither [2] nor [8] test structural assumptions[1], and [8] *never reports the sensitivity of NPI effectiveness estimates in the tests they perform*. **[R2]** correctly points out that these models make "assumptions that we now know are violated", exactly why our novel mathematical results (§5 *Effectiveness in Context*) are important steps forward. Our results show that when commonly made assumptions are violated, estimates must be interpreted as averages, taken over contexts of the dataset, and expert judgement is required to adjust them to local, unique circumstances.

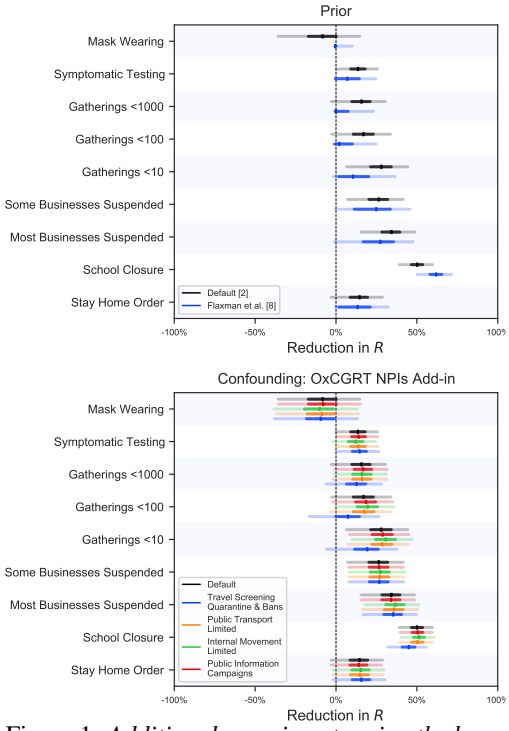

Figure 1: *Additional experiments using the baseline model. Top:* ***[R2]*** *prior from [8]. Bottom:* ***[R3]*** ***[R4]*** *additional confounding tests. The NPIs from the OxCGRT dataset (as labeled) are now observed.*

[R2] Implementation. Our implementation of the model of [8] is correct; we only model latent infections as a discrete renewal process while deaths are modelled as in [8] and [2] i.e., produced via discrete convolutions. **We believe this misunderstanding is due to typographical errors in the supplement, Eqs. (128), (129)**. The correct equation, modelling only deaths, is $N^{(D)}_{t,c} = R_{t,c} \sum_{\tau=1}^{t} N^{(D)}_{t-\tau,c} \cdot \pi_{SI}[\tau]$ where $\pi_{SI}[\tau]$ is the discretised serial interval distribution, $N^{(D)}_{t,c}$ is the daily number of infections that result in fatalities. $R_{t,c}$ is the instantaneous reproduction number at time t in country c. We seed this with a latent variable $N^{(D)}_{0,c}$ that incorporates the country-specific infection fatality rate, IFR_c. Other than truncation and naming, this is identical to $c_{t,m} = R_{t,m} \sum_{\tau=0}^{t-1} c_{\tau,m} g_{t-t}$ [8] where the convolution has been rewritten indexing over the other variable. Since c represents the *total* number of infections, we have $c_{t,c} = N^{(D)}_{t,c}/\text{IFR}_c$. We compute the expected number of deaths as $\bar{D}_{t,c} = \sum_{\tau=1}^{63} N^{(D)}_{t-\tau,c} \pi_D[\tau]$, where $\pi_D[\tau]$ is the discretised infection-to-death delay. We implemented all models ourselves to minimise discrepancies between models and make fair comparisons.

[R3] [R4] Confounding. Thank you for pointing out the no-confounder assumption. We agree that this assumption is critical, and will update the tone of the conclusion to reflect the assumptions we tested. For clarity, the NPI leave-out test assesses how much the effect of unobserved interventions are attributed to observed NPIs **[R2]** **thereby testing this assumption**. We apologise for not clarifying the purpose of this test. In light of your feedback, we have run additional experiments finding low sensitivity when previously unobserved NPIs from the OxCGRT NPI dataset [*Thomas Hale et al. Oxford COVID-19 Government Response Tracker. (2020)*] are observed (Fig. 1, bottom). These tests are imperfect but considered best practice [*Rosenbaum et al., Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome, 1983*]. We highlight that our results show that confounding is the key limitation of such NPI effectiveness models. For example, if we had found that effectiveness estimates fluctuate widely under different epidemiological parameters, we would not have been able to make strong conclusions regardless of whether we observe all relevant factors.

[R2] Effectiveness Prior. Thank you for your comment. We take our effectiveness prior from [2], and it reflects the belief that NPIs have moderate effects. Low posterior correlation $r < 0.4$ between NPI effectiveness estimates and low sensitivity suggests that collinearity is manageable. Furthermore, we have added run a test using the suggested prior from [8] (Fig. 1, top).

[1]The most recent version of [2] reproduces our structural sensitivity analysis from the preprint corresponding to this submission.