nature portfolio

Peer Review File

Counterfactual evaluation of elementary and secondary school policies in the COVID-19 pandemic

Corresponding Author: Ms Benedetta Canfora

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this work, Canfora et al. perform a counterfactual evaluation of the impact of school closures on COVID-19 transmission in the Netherlands. Overall this paper is extremely well written, of very high scientific quality, and relevant. In particular, I really appreciated the elasticity analysis: I suspect this approach will become more utilized after publication of this paper. Prior to publication though, I would request the authors consider the following:

Larger comments
-My main comment concerns the “q” values, which quantify the level of contacts in schools during the different periods compared to the baseline. It seems like your model suggests school contacts are extremely important for transmission – changing school opening/closure has massive impacts on cases/hospitalizations. Yet, this must be quite sensitive to “q”: if you assume that contacts in schools are very limited, then reopening or closing them would have more of a modest effect. Therefore, because of what I perceive to be a critical impact of this parameter on the results, I would request that (i) more detail be provided related to exactly how “q” is calculated (I don’t see it as a fitted parameter in Figs S5-S8), (ii) plots of “q” over time be shown, and (iii) some sensitivity analyses be performed to determine how sensitive the results are to the value of “q”.

Minor comments
-Do you consider vaccine waning? This isn’t explicitly stated in your “Transmission model” section I don’t believe.
-Where does the reported infection curve come from in Fig 2? Is that PCR test data?
-In Figure 2 caption, you say shaded regions AND black outlined segments show the counterfactual periods – this is a bit confusing (although I realize they are coincident).
-Figure S3: shaded regions aren’t really yellow, and it’s hard to see the vertical dashed lines for the lockdown measures
-Is this correct in your Table S2 caption: “calibrated: values taken from [1] and other relevant studies, or fixed: set via grid search”? The meaning of “fixed” and “calibrated” seems reversed?
-small typo on p16 of the supplement: “which we can readily programmed”
-Page 11: the phrasing “high children contacts” is awkward

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this study the authors perform a retrospective analysis using counterfactual scenarios to assess the impact of school closures and reopening strategies at the elementary and secondary school levels in the Netherlands during the first two years of the COVID-19 pandemic. Overall, I really like this study--it is well conducted and the manuscript is clearly written. I especially like the use of multiple datasources in the parameter estimation, and the elasticity analysis. School closures were one of the main non-pharmaceutical interventions implemented worldwide in the early stages of the pandemic. It is important to do such retrospective analyses to estimate their impact as they have been also linked to substantial learning losses and adverse effects on children's mental health. Such studies are necessary to identify optimal epidemic planning scenarios that

minimize both disease burden and societal disruption. However, there are certain aspects of the methods that are currently not very clear. I'm noting my comments below and hopefully this will help strengthen the manuscript.

Major comments:

1. My first comment is related to the splitting of the overall contact matrices into school and non-school components. As an aside, the surveys used (RIVM PiCo) seem to be a great resource to study longitudinal changes in contact patterns! After having gone through the methods in the main text and the supplement I have a rough idea but it will be good to provide more details about how the fraction of school contacts q was estimated for each period? It seems to me that there will be a range of values of q that will satisfy the imposed constraints so how were the final values chosen? I might be missing something, but the school and non-school split of the contact matrices is probably driving the results quite a bit so it needs to be very clear how they were calculated. Relatedly, it might be good to do a sensitivity analysis to see how this choice is influencing the final results.

2. There are some aspects of the fitting procedure that are not very clear currently. At what point in the fitting pipeline is the grid search done? Is the model fitting of the estimated parameters and the grid search of the fixed parameters done simultaneously? Or is the model fitting done first assuming some values for the fixed parameters and then a grid search is done to optimize the fixed parameter values? It is probably something like the latter since as per supplementary Figure S9, the model fit and the baseline counterfactual trajectories are different, and also this is what the code seems to indicate. But I'm a bit confused since the inference_fit code for say CP1 already seems to have the fixed parameter values hardcoded in it. Maybe it is more like model fit -> grid search -> model fit? Will be good to clarify in the text without having to dig through the code.

3. Can you comment on why the hospitalizations peak is very underestimated during the CP1 period? Wonder if this is indicating that the global optimum is not reached. Overall, it would be good to see the MCMC parameter trace plots!

Minor comments:

1. Right now the counterfactual scenarios look at each period in isolation and how just that period would have been affected if the school closure policies were changed during it (while keeping the past policies the same). Maybe it is worth stressing this explicitly since the cumulative effect of changing multiple policies hasn't been investigated here.

2. Figure 3: Since the light pink school status doesn't exist in this figure, I'd remove it from the legend. It is a bit confusing since that shade is pretty similar to the one used for children's cumulative elasticities.

3. Figure S2: This is probably due to rounding but for many of the contact matrix entries, the school and non-school contacts don't quite add up to the total contacts; for example, subplot A.

4. Figure S9: The first line of the caption has a typo, the subplot description is incorrect--the top row isn't just cases and the bottom isn't just hospitalizations.

(Remarks on code availability)
The README is very thorough and easy to follow. I ran the basic inference code (inference_fit_CP1.m, inference_baseline_CP1.m) for CP1 in Matlab and it seemed to work without any errors. The code in these files is also well commented. Just to note however that since the code is in Matlab it is not quite open-source.

Reviewer #3

(Remarks to the Author)
The manuscript „Counterfactual evaluation of elementary and secondary school policies in the COVID-19 pandemic" by Canfora et al. addresses the epidemiological effects of elementary and secondary school closures during the SARS-CoV-2 pandemic. This is a modelling study that uses time-varying contact matrices from surveys to study how case numbers and hospitalizations would have been different und counterfactual policies with regard to schools being closed or open. Whereas other modelling studies often relied either on assumptions regarding how the contact matrices would have changed or on crude and aggregate proxies such as mobility, the availability of this survey data allows the authors to study this issue more comprehensively and with fewer ad hoc assumptions. This is therefore an interesting and important study that adds to the literature. There are, however, some major issues that the authors would need to address in a potential revision.

Major comments:

(1) Notation and symbol clarity. There are occasions where multiple meanings are given for the same symbol (e.g., S is used for "susceptible" and "secondary"), indices that can be mistaken for exponents (both for the compartments and the betas, gammas, etas, etc.), and inconsistent time/period labelling (the paper would be easier to follow if time-dependence is explicitly given).

(2) Direct versus indirect effects of school closures. There is a rich literature on the effects of school closures showing that both direct and indirect effects are at play. In brief, if schools were closed, parents were more likely to also work from home to watch the kids. Can these effects be separated and estimated from the contact-matrices? And can this be used to also estimate the corresponding contributions to increases/decreases in case numbers and hospitalizations? I think such an

analysis could help readers better understand the mechanisms behind the results.

(3) CP implementation, counterfactual design, and hypothesis testing: It is unclear to me to which extent counterfactuals are evaluated in isolation or sequentially, what exactly is held fixed per CP, and what hypotheses tests this corresponds to. The manuscript indicates parameters were estimated separately per period, with posterior distributions from one period used to initialize the next, and that some parameters were fixed via grid search so the ‘baseline counterfactual’ matched real data These choices should be explicitly tied to the counterfactuals.

More concrete, I understand that at the moment CPs are evaluated sequentially. This means that the evaluation of the last CP also depends on all previous CPs. How much of an issue is this? How much do the evaluations depends on whether the CPs are evaluated in isolation (keeping the baseline dynamics unchanged in all CPs prior to the one that is evaluated) or in specific sequences, introducing path dependence of the results? Would it be more meaningful to consider naïve but consistent policy scenarios, such as “schools always open” or “schools always closed” or “schools close if hospitalizations exceed a threshold X”? What exactly are the hypothesis embedded in the current design of the CPs?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Canfora et. al. did a truly commendable job revising this paper – I think it’s the most thorough paper review I’ve ever seen. I focused on the sections related to q and the sensitivity analyses as that was where my initial comments were centered. I will let the other reviewers comment on the fitting procedures.

Overall, I have a much better understanding of the q value now. After going through the changes though, I am left with the following questions:

-In the supplement when you write “As there are several values that satisfy (S7), we selected the minimum q(j)…” (just before S8), you mean that there are multiple q values for the different age groups right? But each one solves the equality in Eq S7 (even though it’s presented as an inequality), is that right? So you take the minimum value across age groups, not the minimum value for a given age group that could apply to > instead of = in Eq S7?

-Something seems off in the survey dates in Table S1.. do surveys 5 and 6 overlap? And is the end date correct for survey 7?

-Supplementary lines 850-852: I thought ckk^1 is pre-pandemic? Not 1st lockdown?

-I think the grids in S3/S4 could be formatted to be a bit bigger which would help with readability

-In your sensitivity analyses, when you changed q, did you also adjust non-school contacts? Or are total contacts potentially higher/lower when you perturb q?

-S22/24 figures are very small – can you make the figure size closer to S23?

- “As mentioned in Supplement Section 2.1, in most time intervals, the increase in q just redistributes survey contacts by increasing school and reducing non-school contacts. This implies that total contacts are the same and therefore in principle the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution”.
I don’t understand this. If you change q, then the non-school contacts change to preserve the same total contacts (Eq S7). When you run your baseline scenario then, would it not be possible for the baseline to be different now (even without changing school opening/closure) because contacts outside of school are different?

-“The baseline cumulative hospitalizations for q not perturbed and perturbed are shown in Table S7, and they are identical in CP2-CP3 for different perturbations of q, because the only time period where baseline total contacts differ from the original q is time interval F, and in that time period all perturbation lead to the same capped value q = 1, while this is not the case for CP4. As Table S7 shows, the CP2 and CP4 baseline results are very similar relative to q not perturbed. However, the CP3 baseline hospitalizations are much larger when q is perturbed, and as a result the baseline does not match the observed data as in the original analysis, making it an unrealistic scenario.”
I’m having trouble following this paragraph. Firstly nothing seems “identical” to me in Table S7, unless I’m missing something. Second the paragraph reads as saying things are identical in CP3 when q is perturbed first, but then later says baselines are much larger in CP3 when q is perturbed. Also, if q is increased in CP3, then shouldn’t non-school contacts go down? Even if you then adjust non-school contacts, overall this value should be lower shouldn’t it? So then why do you say contacts are higher resulting in the higher effective R and higher effective hospitalizations?

-I think you forgot to update the caption of figure S4 to reflect the visual changes you made

-“School capacity” is a slightly confusing term because it sounds like the total number of students the school can accommodate. What you’re varying is the extent of closure/opening. I would consider changing the terminology.

-Figures S25-28: what are those values? Are those numbers of infections (i.e. # of infections in perturbed scenario - # in baseline) or percentages?

-In your sensitivity analysis when you vary non-school contacts (Section 12 in the supplement), you’re not preserving total contacts right? So the total number of contacts could be changing? Related, I think this sentence is worth noting in the main too if you didn’t already “This suggests that sufficiently strong reductions in adult non-school contacts could, in principle, have allowed higher levels of school activity in some periods without leading to large increases in infections and hospitalizations.” This is an interesting finding given the large negative impacts of school closure, and is something to consider when coming up with how to prioritize public health decisions.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Thank you for addressing all of my comments so thoroughly! I have no further comments and recommend the manuscript for publication.

(Remarks on code availability)
I had reviewed the CP1 code in the first revision.

Reviewer #3

(Remarks to the Author)
I thank the authors for addressing all my comments in a comprehensive and satisfactory way. I have no further comments and congratulate the authors for their work.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors did a great job addressing my comments. I have a couple remaining things I don’t quite follow, but this shouldn’t affect the decision to proceed with publication.

-Supplement 870/871: Do you mean something like “In these OTHER time periods…”

-When you describe the other factor used in periods F,H,L, your fitting procedure is not 100% clear. What is a baseline counterfactual? My understanding is you choose different factor values, then fit your model in one CP to the actual data to get the values of all other parameters, simulate 300 times, and then find the error between the hospitalization data and your median. You pick the factor that results in the smallest error. But I’m confused what “baseline counterfactual” means – the baseline is not counterfactual, is it?

-Out of curiosity, why is your fit in Figure S14 different from your baseline simulation for each CP? Is that just the result of stochasticity and the fitting procedure?

-I’m stumped by this, which I also asked in my last round of questions, and you say it or something similar several times: “As mentioned in Supplement Section \ref{suppsubsect:q_calculation}, in most time intervals, the increase in q just redistributes
survey contacts by increasing school and reducing non-school contacts. This implies that total contacts are the same, and therefore, in principle, the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution.”

Why does only total contacts matter for baseline scenarios if you’re explicitly modeling school and non-school interactions, and these have different governing dynamics? To me if you change q, even keeping the opening/closure of schools the same (i.e. not introducing the counterfactual), the total number of hospitalizations may change. This leads to confusion for me to understand why varying q in your sensitivity analyses doesn’t change your “baseline” (even when you don’t have the issue of total contacts changing due to the adjustment factor). Why would this not be the case? For the sensitivity analyses is it because you recalibrate the model with the new qs?

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this work, Canfora et al. perform a counterfactual evaluation of the impact of school closures on COVID-19 transmission in the Netherlands. Overall this paper is extremely well written, of very high scientific quality, and relevant. In particular, I really appreciated the elasticity analysis: I suspect this approach will become more utilized after publication of this paper. Prior to publication though, I would request the authors consider the following:

Response: We would like to thank the Reviewer for their time, the valuable feedback and the opportunity to revise our manuscript. Below, you will find a point-by-point response, highlighted in blue. Unless stated otherwise, line numbers, figures, and citations refer to the revised manuscript without the track-changes.

Larger comments

My main comment concerns the "q" values, which quantify the level of contacts in schools during the different periods compared to the baseline. It seems like your model suggests school contacts are extremely important for transmission – changing school opening/closure has massive impacts on cases/hospitalizations. Yet, this must be quite sensitive to "q": if you assume that contacts in schools are very limited, then reopening or closing them would have more of a modest effect. Therefore, because of what I perceive to be a critical impact of this parameter on the results, I would request that

(i) more detail be provided related to exactly how "q" is calculated (I don't see it as a fitted parameter in Figs S5-S8),

Response: We thank the Reviewer for this thoughtful comment and for highlighting the central role of the parameter *q* in shaping the estimated impact of school closures and reopenings. We agree that a clear description of how *q* is constructed is essential for interpreting the results.

In the revised manuscript, we have substantially expanded the description of the calculation of *q*. Specifically, we have added a new Section 2.1 in the Supplementary Material (pages 7-14, lines 780-885), which details how the survey-specific *q* values were derived, the assumptions underlying their construction, and how elementary school, secondary school, and non-school contacts were subsequently obtained based on the *q* values and used in both the model fitting and the counterfactual analyses.

We note that *q* is not a fitted parameter. Rather, it is a scaling factor applied to pre-pandemic school contact matrices to reflect reduced contact intensity in schools during the pandemic, prior to fitting the model to the data. Allowing school contacts to return to their full pre-pandemic levels during periods when schools were open would, for several surveys, result in school contacts exceeding the total number of reported contacts, thereby implying meaningless (negative) non-school contacts. To avoid this inconsistency, we reduced school contacts to survey-specific fractions of their pre-pandemic values.

More precisely, for each survey between Surveys 3 and 7, we selected a value of *q* between zero and one, defined as the minimum across the 3×3 entries of the matrix, where the entries are calculated as:

q = (survey contacts – first lockdown contacts)/pre-pandemic school contacts.

In the above equation, the numerator is a proxy for the school contacts for a specific survey, under the assumption that school contacts were zero in the first lockdown because schools were closed. This approach ensures that (i) non-school contacts, computed for a specific survey as the difference between total contacts and reduced school contacts, remain positive, and (ii) reduced school contacts during the pandemic reflect the presence of mitigation measures in schools compared to the pre-pandemic period.

We believe that the additional explanation provided in Section 2.1 of the Supplementary Material (Lines 780-885) clarifies both the rationale for the construction of *q* and its role in defining school and non-school contact patterns throughout the analysis.

(ii) plots of “q” over time be shown,

Response: Thank you for the suggestion. The parameter *q* is calculated separately for each survey, and we have added plots of its values over time in Figure S2 of the Supplementary Materials.

(iii) and some sensitivity analyses be performed to determine how sensitive the results are to the value of “q”.

Response: We thank the Reviewer for this suggestion. We agree that assessing the sensitivity of the results to the value of *q* is important. We conducted a sensitivity analysis by perturbing the survey-specific *q* values, considering increases of 2%, 5%, 10%, and 20%, capped at q=1. This cap reflects the interpretation of *q* as a fraction of pre-pandemic school contacts, which cannot exceed 1 due to ongoing mitigation measures in schools during the first two years of the pandemic. The results are reported and discussed in detail in the new Section 10 of the Supplementary Material (pages 33-40, Lines 980-1019) and illustrated in Figures S22-S24 and Tables S8-S10.

These results show that increasing *q* does not change the qualitative conclusions of the study: the relative importance of different school-opening and school-closing strategies remain consistent across counterfactual periods and scenarios. The results are also quantitatively robust for CP2 and CP4, with minor changes in the baseline and counterfactuals as shown in Figures S22 and S24 and Tables S7, S8 and S10. However, the changes are larger for CP3, see Figure S23 and Table S9. Table S7 shows that these discrepancies are due to very high baseline hospitalizations that do not match the empirical data, making the perturbations of *q* in CP3 an unrealistic scenario to explore without further adjustments in contacts.

We clarify this in the main text as well, lines 178-182, where we now state:

"Our results were qualitatively robust to moderate variations in the allocation of total contacts between school and non-school settings across counterfactual periods. The relative importance of interventions targeting elementary versus secondary schools was preserved in all counterfactual periods, with minor differences in absolute number of hospitalizations for CP2 and CP4 and larger differences for CP3. The latter can be explained by a very high baseline hospitalization wave compared to the original data, making it an unrealistic scenario."

Minor comments

-Do you consider vaccine waning? This isn't explicitly stated in your "Transmission model" section I don't believe.

Response: We thank the Reviewer for raising this point. Vaccine waning is accounted in our model by adjusting the parameters reducing susceptibility, $p_{I,k}(t)$, infectivity, $p_{TR,k}(t)$ and $p_{TU,k}(t)$, and probability of hospitalization, $p_{H,k}(t)$, after vaccination, which were used in the model detailed in Supplementary Material Section 1. To describe vaccine waning, these reduction parameters are allowed to vary over time, decaying over a six-month period following vaccination.

We have clarified this in the revised Transmission Model section:

(Lines 352-354, main): "Waning of vaccine-induced immunity is modeled by allowing the corresponding reduction parameters to vary over time, decaying over a six-month period following vaccination. Further details of modeling vaccine waning can be found in [64], Supplementary Section 3.1."

-Where does the reported infection curve come from in Fig 2? Is that PCR test data?

Response: Yes. We have clarified this in the revised caption of Figure 2 (Figure 1 in the updated manuscript) as follows: "[...] and the gray area shows daily PCR-confirmed SARS-CoV-2 infections based on national surveillance data from the National Institute for Public Health and the Environment (right axis), [...]". This is also mentioned in the Data subsection (page 13, lines 319-321): "Daily reported PCR-confirmed infections by age and province, as well as multiple hospitalization datasets, covering different periods and levels of aggregation, were obtained from the National Institute for Public Health and the Environment (RIVM)."

-In Figure 2 caption, you say shaded regions AND black outlined segments show the counterfactual periods – this is a bit confusing (although I realize they are coincident).

Response: We thank the Reviewer for pointing this out. We agree that, in Figure 2 (which in the revised version is Figure 1), the shaded regions and black-outlined boxes indicate the same counterfactual periods. Both elements were intentionally retained because the black outlines help keep the color blocks representing school operational status visually distinct, particularly in the results figures (Figures 2–5 in the revised version). To avoid any potential confusion, we have revised the caption as follows: "[...] Black-outlined segments mark the periods used for counterfactual analyses, which coincide with the shaded areas, as they represent the same periods. [...]".

-Figure S3: shaded regions aren't really yellow, and it's hard to see the vertical dashed lines for the lockdown measures

Response: We thank the Reviewer for this observation. We retained the shaded regions to maintain consistency with previous figures depicting the counterfactual periods. To improve visual clarity, we changed the color representing the South provinces to green, making the shaded periods more distinguishable. In addition, the vertical lines indicating lockdown measures have been changed to black solid lines to enhance their visibility (Figure S4 in the revised version).

-Is this correct in your Table S2 caption: "calibrated: values taken from [1] and other relevant studies, or fixed: set via grid search"? The meaning of "fixed" and "calibrated" seems reversed?

Response: We thank the Reviewer for pointing out that our terminology caused some confusion. The terms "fixed" and "calibrated" were indeed used incorrectly in the original version.

We have now revised the Supplementary Material to replace "fixed" with "grid search" for parameters estimated by searching over a predefined grid and selecting the value that minimizes the sum of absolute differences between the median (over 300 ensemble simulations) of simulated daily infections plus hospitalizations in the baseline scenario and the corresponding empirical data. Conversely, parameters previously labeled as "calibrated" are now correctly referred to as "fixed."

These changes have been implemented as follows:

- in the Supplementary Materials Table S2 (which in the revised version is Table S3) caption: "[...] Parameters are classified as *estimated*: inferred from data, *fixed*: values taken from [1] and other relevant studies, or *grid search*: set via grid search. [...]", and
- in Supplementary Materials, Section S3, where the revised terminology and procedure are now explicitly described as follows (Lines 898-905, supplement): "[...] The other parameters whose ensemble Kalman filter-based posteriors were not sufficiently tight were instead selected by grid search simultaneously with the non-school contact reduction/increase factors described in Section 2. Specifically, a grid was defined for each parameter, and for each combination of parameter values the model was refit by the ensemble Kalman filter to retrieve the posteriors of the other parameters. We selected the combination of values on the multidimensional grid that minimize the sum of the daily absolute distances between the median - over 300 ensembles - simulated infections plus hospitalizations in the baseline scenario and the corresponding empirical data to ensure the best match to the data."

-small typo on p16 of the supplement: "which we can readily programmed"

Response: The sentence has been revised to (Lines 956, supplement): "which we can readily program to get the time-varying reproduction number of the system."

-Page 11: the phrasing "high children contacts" is awkward

Response: The sentence has been revised to (Lines 151-154, main): "[...] These resulting high *R(t)* values are consistent with a large number of contacts among children and a rapidly decreasing proportion of susceptible individuals in this group (Supplementary Figures S5 and S21). [...]"

Reviewer #2 (Remarks to the Author):

In this study the authors perform a retrospective analysis using counterfactual scenarios to assess the impact of school closures and reopening strategies at the elementary and secondary school levels in the Netherlands during the first two years of the COVID-19 pandemic. Overall, I really like this study--it is well conducted and the manuscript is clearly written. I especially like the use of multiple datasources in the parameter estimation, and the elasticity analysis. School closures were one of the main non-pharmaceutical interventions implemented worldwide in the early stages of the pandemic. It is important to do such retrospective analyses to estimate their impact as they have been also linked to substantial learning losses and adverse effects on children's mental health. Such studies are necessary to identify optimal epidemic planning scenarios that minimize both disease burden and societal disruption. However, there are certain aspects of the methods that are currently not very clear. I'm noting my comments below and hopefully this will help strengthen the manuscript.

Response: We would like to thank the Reviewer for their time, the valuable feedback and the opportunity to revise our manuscript. Below, you will find a point-by-point response, highlighted in blue. Unless stated otherwise, line numbers, figures, and citations refer to the revised manuscript without the track-changes.

Major comments

1. My first comment is related to the splitting of the overall contact matrices into school and non-school components. As an aside, the surveys used (RIVM PiCo) seem to be a great resource to study longitudinal changes in contact patterns! After having gone through the methods in the main text and the supplement I have a rough idea but it will be good to provide more details about how the fraction of school contacts q was estimated for each period? It seems to me that there will be a range of values of q that will satisfy the imposed constraints so how were the final values chosen? I might be missing something, but the school and non-school split of the contact matrices is probably driving the results quite a bit so it needs to be very clear how they were calculated. Relatedly, it might be good to do a sensitivity analysis to see how this choice is influencing the final results.

Response: We thank the Reviewer for this thoughtful comment and for highlighting the role of the school-contact fraction *q* in shaping the estimated impact of school closures and reopenings. We agree that a clear description of how q is calculated is essential for interpreting the results. In the revised manuscript, we have substantially expanded the description of q in the Supplementary Material (Section 2.1, pages 7-14, lines 780-885). This section details how survey-specific q values were derived, the assumptions underlying their construction, and how elementary school, secondary school, and non-school contacts were subsequently calculated for use in both model fitting and the counterfactual analyses.

More precisely, for each survey between Surveys 3 and 7, we selected a value of *q* between zero and one, defined as the minimum across the 3×3 entries of the matrix, where the entries are calculated as:

q = (survey contacts – first lockdown contacts)/pre-pandemic school contacts.

In this formula, the numerator is a proxy for the school contacts for a specific survey, under the assumption that school contacts were zero in the first lockdown because schools were closed. This approach ensures that (i) non-school contacts, computed for a specific survey as the difference between total contacts and reduced school contacts, remain positive, and (ii) reduced school contacts during the pandemic reflect the presence of mitigation measures in schools compared to the pre-pandemic period.

Following the Reviewer's suggestion, we also conducted a sensitivity analysis by perturbing the survey-specific q values, considering increases of 2%, 5%, 10%, and 20%, capped at q=1. This cap reflects the interpretation of q as a fraction of pre-pandemic school contacts, which cannot exceed 1 due to ongoing mitigation measures in schools during the first two years of the pandemic. The results are reported and discussed in detail in the new Section 10 of the Supplementary Material (pages 33-40, Lines 980-1019) and illustrated in Figures S22-S24 and Tables S8-S10.

These results show that increasing q does not change the qualitative conclusions of the study: the relative importance of different school-opening and school-closing strategies remain consistent across counterfactual periods and scenarios. The results are also quantitatively robust for CP2 and CP4, with minor changes in the baseline and counterfactuals as shown in Figures S22 and S24 and Tables S7, S8 and S10. However, the changes are larger for CP3, see Figure S23 and Table S9. Table S7 shows that these discrepancies are due to very high baseline hospitalizations that do not match the original data, making the perturbations of q in CP3 an unrealistic scenario to explore without further adjustments in contacts.

We clarify this in the main text as well, lines 178-182, where we now state:

"Our results were qualitatively robust to moderate variations in the allocation of total contacts between school and non-school settings across counterfactual periods. The relative importance of interventions targeting elementary versus secondary schools was preserved in all counterfactual periods, with minor differences in absolute number of hospitalizations for CP2 and CP4 and larger differences for CP3. The latter can be explained by a very high baseline hospitalization wave compared to the original data, making it an unrealistic scenario."

2. There are some aspects of the fitting procedure that are not very clear currently. At what point in the fitting pipeline is the grid search done? Is the model fitting of the estimated parameters and the grid search of the fixed parameters done simultaneously? Or is the model fitting done first assuming some values for the fixed parameters and then a grid search is done to optimize the fixed parameter values? It is probably something like the latter since as per supplementary Figure S9, the model fit and the baseline counterfactual trajectories are different, and also this is what the code seems to indicate. But I'm a bit confused since the inference_fit code for say CP1 already seems to have the fixed parameter values hardcoded in it. Maybe it is more like model fit -> grid search -> model fit? Will be good to clarify in the text without having to dig through the code.

Response: We thank the Reviewer for the careful reading and for highlighting the lack of clarity in the description of our fitting procedure.

The fitting pipeline is indeed sequential. For each counterfactual period, we first performed an initial model fit using the standard initialization parameters and the corresponding baseline counterfactual trajectories generated from the posterior parameter values. If this initial baseline did not adequately reproduce the observed data, we identified parameters with posterior distributions that were too wide and performed a multi-dimensional grid search over these parameters to identify values that minimized the discrepancy between the baseline counterfactual trajectories and the data. The model was refitted at each point on the grid. Once suitable values were identified, the model was refitted with these updated parameter values, and the resulting posterior distributions were then used to generate the baseline counterfactual trajectories and the other intervention scenarios presented in the Results.

The difference between the model fit and the baseline counterfactual trajectories (Supplementary Figure S14, formerly Figure S9) arises from the temporal treatment of parameter posteriors: the model fit allows parameters to vary over time, while the baseline scenario is generated using parameter posteriors fixed at the end of the sample period, which explains the observed differences.

We have clarified this "model fit → grid search → model fit" workflow explicitly in the revised Methods section as follows (Lines 400-404, main):

"[…] For each counterfactual period, we first performed initial model fitting for some parameters, where each model fit was implemented for a single value on a multidimensional grid over the remaining parameters. We then selected the value on the grid that minimized the distance between the baseline counterfactual and the observed data, and refitted the model with these parameter values to obtain the posterior distributions used in the baseline simulations (see Supplementary Material, Section 3). […]"

Additionally, in the Supplementary Materials (Section 3, page 15), we have added the following clarification (Lines 894-896 and 902-905, Supplement):

"Values of parameters that were not estimated but fixed are given in Table S3. In particular, μ, D, and Z were fixed based on the average posteriors over time in [1] (Supplementary Materials, Figure A2).
[…]
We selected the combination of values on the multidimensional grid that minimize the sum of the daily absolute distances between the median - over 300 ensembles - simulated infections plus hospitalizations in the baseline scenario and the corresponding empirical data to ensure the best match to the data."

3. Can you comment on why the hospitalizations peak is very underestimated during the CP1 period? Wonder if this is indicating that the global optimum is not reached. Overall, it would be good to see the MCMC parameter trace plots!

Response: Thank you for pointing this out. We want to clarify that we are not using Bayesian estimation with MCMC. The method we use is an approximate Bayesian method called ensemble adjustment Kalman filter.

The reason we need to use filtering to update the (unobserved) state variables (S, I^U, R) is because we had to treat them as stochastic rather than deterministic due to many compartments with relatively small or depleting populations. An exact method for obtaining these updates for the unobserved state variables in nonlinear models is the particle filter. However, this breaks down in high dimensional models like ours, with many state variables, as shown in Bengtsson, Bickel and Li (2008), *Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems*. *Institute of Mathematical Statistics Collections, Vol. 2, pages 316-334.* Instead, we used the ensemble Kalman filter, which starts with many (300 in our case) initial draws or ensemble members for each state variable, and calculates their mean across ensemble members. For these means, 300 Poisson draws are taken for each model component that is an integer, where the Poisson distribution was chosen to reflect that when the values of state variables like infections and hospitalizations are high, so is the variability around them. For each of these draws, we employ the nonlinear transition function characterizing the model equations. We then integrate the model over a day for each draw, after which a closed form update is performed for each of these draws and the operation is repeated for the next day. This closed form update uses each observed series one-by-one, and is derived by assuming the model is linear and everything is normally distributed, in which case the "posterior" distribution of the state variables is normal. Since in nonlinear models the posterior is not normally distributed even if the initial states would be, it is an approximate filtering method.

Typically, filtering is used to calculate the likelihood when the unobserved states are stochastic. Then, a Bayesian estimation or a likelihood maximization is employed on top to estimate the parameters. However, for computational efficiency, we treat the parameters as state variables with constant transition function, and update them in the same way as the state variables, based on the same closed form expression, therefore also assuming they are also normally distributed. Since, as the (time-series and cross-sectional) sample size grows, the parameter estimates should converge to the normal distribution under reasonable statistical assumptions, this approximation should yield the right parameter estimates after sufficiently many updates, although a formal proof of this conjecture is not available to our knowledge, despite the widespread use of the ensemble Kalman filter across disciplines. Since we are using daily data, and, for each day, 36 series of infection data (12 regions x 3 age categories), and 14 series of hospitalization data (12 regions for adults, national for children and adolescents), we have sufficient observations in each CP1-CP4 period for convergence.

One way to assess convergence in our setting is to look at whether the parameter posterior updates stabilize after a while. The plots in Supplementary Material, Section 3, Supplementary Figures S6-S13 show that this is indeed the case.

So why do we underfit the hospitalizations in CP1? One possible explanation is that the hospitalization data for adults shows so much variability at the peak compared to the day before and after, that an approximate filter cannot capture it. This is despite the fact that we tried to address this by taking Poisson draws and also assuming that the observed series have variance equal to their mean. We tried to make this latter variance even larger (4 times the mean), without any effect on the model fit.

Another likely explanation is that our age stratification of adults is not sufficiently fine-grained to capture such large peaks; these peaks would be smaller if we had used 10-year age strata by region, and thus more amenable to detection by the filtering technique we employ.

We have now added this limitation in the updated manuscript as follows:

(Lines 289-291, main) "[...] The lack of finer age stratification among adults, combined with using an approximate rather than an exact Bayesian inference method, may also explain why the model fit did not fully capture the large hospitalization peak during the first wave."

Minor comments

1. Right now the counterfactual scenarios look at each period in isolation and how just that period would have been affected if the school closure policies were changed during it (while keeping the past policies the same). Maybe it is worth stressing this explicitly since the cumulative effect of changing multiple policies hasn't been investigated here.

Response: Thank you for this comment. We agree that this aspect of the counterfactual design should be stated more explicitly. In our analysis, each counterfactual period is evaluated in isolation, conditional on the observed epidemic trajectory and school policy history up to the beginning of that period. Counterfactual scenarios modify school operation only during the period under consideration, while all other past policies and epidemic dynamics are kept unchanged.

Our aim is therefore to assess the impact of time-localized and historically plausible changes in school policies, rather than the cumulative effects of altering multiple policies over longer time horizons. We have revised the manuscript to explicitly clarify this point.

(Lines 67-70, main) "[...] Each counterfactual period was evaluated in isolation, conditional on the observed epidemic trajectory up to that point in time. The counterfactual scenarios differed from the baseline scenario only in terms of school-related contacts, while all other epidemiological parameters, non-school interventions, and past policies were kept unchanged."

(Lines 218-222, main) "[...] Rather than evaluating globally uniform policies, such as schools being always open or always closed [38], we focused on period-specific policy decisions made under the epidemiological and social conditions prevailing at each phase of the pandemic. This approach reflects how school policies were considered and implemented in practice, enabling assessment of their impact in realistic, historically grounded contexts."

2. Figure 3: Since the light pink school status doesn't exist in this figure, I'd remove it from the legend. It is a bit confusing since that shade is pretty similar to the one used for children's cumulative elasticities.

Response: Thank you for pointing this out. The "School holiday" entry in the legend has now been removed from the figures that do not show this specific period, such as CP 1 and CP 4. (Figures 2 and 5 in the revised version)

3. Figure S2: This is probably due to rounding but for many of the contact matrix entries, the school and non-school contacts don't quite add up to the total contacts; for example, subplot A.

Response: Thank you for this observation. The discrepancies are indeed due to rounding. We have verified all values, and the school and non-school contact matrices correctly sum to the total contact matrix in the original data. Figure S3 has now school and non-school contacts that add up to the total contacts.

4. Figure S9: The first line of the caption has a typo, the subplot description is incorrect--the top row isn't just cases and the bottom isn't just hospitalizations.

Response: Thank you for highlighting this. We have corrected the caption of Figure S14 accordingly, updating the figure caption to accurately describe the top and bottom rows.

Reviewer #2 (Remarks on code availability)

The README is very thorough and easy to follow. I ran the basic inference code (inference_fit_CP1.m, inference_baseline_CP1.m) for CP1 in Matlab and it seemed to work without any errors. The code in these files is also well commented. Just to note however that since the code is in Matlab it is not quite open-source.

Response: Thank you so much for running the code and for mentioning that it is well commented. The reason we chose Matlab is due to computational efficiency. Matlab offers user-friendly commands to facilitate vectorizing the code to speed it up. We wrote the code such that it stores five-dimensional arrays and does operations with three-dimensional arrays, avoiding iterations and row-by-row operations when possible to increase speed. As a result, the code for estimating the parameters runs pretty fast.

Reviewer #3 (Remarks to the Author):

The manuscript „Counterfactual evaluation of elementary and secondary school policies in the COVID-19 pandemic" by Canfora et al. addresses the epidemiological effects of elementary and secondary school closures during the SARS-CoV-2 pandemic. This is a modelling study that uses time-varying contact matrices from surveys to study how case numbers and hospitalizations would have been different und counterfactual policies with regard to schools being closed or open. Whereas other modelling studies often relied either on assumptions regarding how the contact matrices would have changed or on crude and aggregate proxies such as mobility, the availability of this survey data allows the authors to study this issue more comprehensively and with fewer ad hoc assumptions. This is therefore an interesting and important study that adds to the literature. There are, however, some major issues that the authors would need to address in a potential revision.

Response: We would like to thank the Reviewer for their time, the valuable feedback and the opportunity to revise our manuscript. Below, you will find a point-by-point response, highlighted in blue. Unless stated otherwise, line numbers, figures, and citations refer to the revised manuscript without the track-changes.

Major comments

(1) Notation and symbol clarity. There are occasions where multiple meanings are given for the same symbol (e.g., S is used for "susceptible" and "secondary"), indices that can be mistaken for exponents (both for the compartments and the betas, gammas, etas, etc.), and inconsistent time/period labelling (the paper would be easier to follow if time-dependence is explicitly given).

Response: We thank the Reviewer for this observation. We have implemented the following changes to improve clarity regarding notation:

1. We have revised the numerical superscripts into subscripts to avoid confusion with mathematical exponents.
2. Time-dependence has been explicitly shown (with *(t)* as in, e.g., *S(t)*) for compartments and parameters. This improved notation is visible in the Transmission Model subsection of the Methods and in the Supplementary Materials Sections 1 and 5.
3. We acknowledge the potential notational conflict but find this particular "abuse of notation" unavoidable for the following reasons: (a) S (Susceptible) and E (Exposed) compartments are standard in infectious disease modeling and (b) S (Secondary) and E (Elementary) schools are central to the scenarios we investigate. To clearly differentiate between these uses, we have adopted a consistent typographic standard:
    - We refer to the school terms with normal typeface (S, E).
    - We refer to the compartments with italics and subscripts (*S_a* and *E_a* for *a=1, 2*). In addition, these terms appear exclusively in the Transmission Model subsection of the Methods in the main paper.
4. We introduced new labels for the survey contact periods (Supplementary Materials, Table S2). While "S" and "E" were avoided, we used "I" and "H", which are also used for the compartments. Similarly, we use a normal typeface for these period labels, whereas the compartments are italicized.
5. The use of the terms "period" and "time" has been modified to avoid confusion as well, e.g.,

- We state "period" when we pertain to changes within and among CPs. For example, we revise the sentence in the caption of Figure 1 in the revised version: "During CP3 schools were closed for only part of the *period* [...]"
- We state "time" when we pertain to changes through time that are common among CPs. For example, "*time*-varying reproduction number" or "median hospital bed occupancy over *time*".
- In addition, parameters "Z" and "D" are now average *durations* instead of average *times*.

6. We omitted the subscript in the notation for the effective reproduction number (from $R_e(t)$ to $R(t)$ since "*e*" is used as the notation for the cumulative elasticities.

(2) Direct versus indirect effects of school closures. There is a rich literature on the effects of school closures showing that both direct and indirect effects are at play. In brief, if schools were closed, parents were more likely to also work from home to watch the kids. Can these effects be separated and estimated from the contact-matrices? And can this be used to also estimate the corresponding contributions to increases/decreases in case numbers and hospitalizations? I think such an analysis could help readers better understand the mechanisms behind the results.

Response: We thank the Reviewer for raising the important point of distinguishing direct and indirect effects of school closures. We fully agree that both mechanisms contribute to transmission changes. However, with the available contact-survey data, a formal decomposition of these effects is not identifiable: the surveys provide total contacts but do not contain behavioral information (e.g., parental work-from-home arrangements, household supervision patterns) that would allow direct and indirect pathways to be separately estimated. To address your concern, we performed two complementary sensitivity analyses that independently probe the two mechanisms you mention. First, to explore direct effects of school contacts, we varied the operating capacity of elementary and secondary schools from 0% to 100% (Supplementary Section 11, Figure S25-S28), while keeping non-school contacts fixed to their baseline values. This isolates the impact of increasing or reducing school-related contacts. Second, to explore indirect effects related to changes in behavior outside school, we independently perturbed non-school contacts of adults and youth (children + adolescents) by −10%, −5%, 0%, +5%, and +10% (Supplementary Section 12, Figures S29–S36), while keeping school contacts fixed to the configuration of each scenario. These perturbations are designed to mimic plausible behavioral responses to school opening/closure, such as parents working from home or changes in extra-school social and recreational activities.

Across all control periods, both sensitivity analyses confirm that the ranking of importance of elementary and secondary school opening and closure in generating infections and hospitalizations are similar, while the magnitude of these effects varies across the explored ranges (Supplementary Sections 11 and 12). For example, in CP1, reopening both elementary and secondary schools is consistently associated with the largest increase in transmission, while reopening only one school level leads to a smaller increase across the full 0–100% capacity grid (Supplementary Section 11, Figure S25).

The school capacity sensitivity analysis further reveals that partial school openings or closures may have been feasible in most counterfactual periods, with the exception of CP1. This is due, for example, to minor changes in hospitalizations when elementary and secondary schools are open at a low capacity, say 25%, compared to the case when they are closed (Supplementary Section 11, Figure S25-S28).
The non-school contact sensitivity analysis further shows that changes in adult non-school contacts have a larger effect on the magnitude of outcomes than comparable changes in youth contacts, and that higher adult non-school contacts tend to amplify the impact of school

reopening (Supplementary Section 12). We revised the manuscript to clarify how indirect behavioral effects can be interpreted and to explicitly state the limitations of our approach in quantifying these effects.

(Lines 215-218, main) “The effects identified in our analyses primarily reflect direct changes in school-related contacts, which are explicitly modified in the counterfactual analyses. While school closures may also induce indirect behavioral responses outside the school setting, such as changes in work-related or other social contacts of adults, these effects cannot be uniquely identified with the available data and are only examined through sensitivity analyses.”

(Lines 440-447, main) “Secondly, the counterfactual scenarios we performed assume either full school openings or complete closures, which represent extreme cases that do not account for other mitigating strategies in schools. To address this, we performed a sensitivity analysis by varying the capacities of elementary and secondary schools across 0%, 25%, 50%, 75%, and 100%. The resulting grid of capacities includes intermediate scenarios ranging between the baseline and the specific counterfactual scenarios. Finally, for all counterfactual periods, we conducted a sensitivity analysis by perturbing the non-school contacts of adults and youth (children and adolescents) by -10%, -5%, 5%, and 10% to examine indirect behavioral impacts of school-based measures.”

(Lines 183-194, main) “To move beyond our considered scenarios of full school closure or complete reopening, we conducted additional sensitivity analyses in which the capacities of elementary and secondary schools were varied between 0% and 100%. The resulting grid of school capacities includes intermediate scenarios between the baseline and the defined counterfactuals. Across all periods, the ranking of elementary and secondary schools in terms of their impact on infections and hospitalizations remained unchanged, while the magnitude of these impacts varied. Notably, the results suggest that partial school openings or closures may have been feasible in most counterfactual periods, with the exception of CP1.

Finally, we examined the sensitivity of our findings to the assumption that school openings and closures may indirectly affect contact patterns in non-school settings. For each counterfactual period, we perturbed non-school contacts among adults and among children and adolescents by ±10%. Changes in adult non-school contacts had a larger influence on both infections and hospitalizations than comparable changes in youth contacts. Moreover, increases in adult non-school contacts tended to amplify the impact of school reopenings.”

(3) CP implementation, counterfactual design, and hypothesis testing: It is unclear to me to which extent counterfactuals are evaluated in isolation or sequentially, what exactly is held fixed per CP, and what hypotheses tests this corresponds to. The manuscript indicates parameters were estimated separately per period, with posterior distributions from one period used to initialize the next, and that some parameters were fixed via grid search so the ‘baseline counterfactual’ matched real data These choices should be explicitly tied to the counterfactuals.

More concrete, I understand that at the moment CPs are evaluated sequentially. This means that the evaluation of the last CP also depends on all previous CPs. How much of an issue is this? How much do the evaluations depends on whether the CPs are evaluated in isolation (keeping the baseline dynamics unchanged in all CPs prior to the one that is evaluated) or in specific sequences, introducing path dependence of the results? Would it be more meaningful to consider naïve but consistent policy scenarios, such as “schools always open” or “schools always closed” or “schools close if hospitalizations exceed a threshold X”? What exactly are the hypothesis embedded in the current design of the CPs?

Response: We thank the Reviewer for this comment, which raises important points about the interpretation of the counterfactual periods (CPs) and their implementation. In our analysis, each counterfactual period is evaluated in isolation, conditional on the observed epidemic trajectory and school policy history up to the beginning of that period. Counterfactual scenarios modify school operation only during the period under consideration, while all past policies and epidemic dynamics are kept unchanged.

Our aim was therefore to assess the impact of time-localized and historically plausible changes in school policies, rather than the cumulative effects of altering multiple policies over longer time horizons. We have revised the manuscript to explicitly clarify this point.

(Lines 67-70, main) "[...] Each counterfactual period was evaluated in isolation, conditional on the observed epidemic trajectory up to that point in time. The counterfactual scenarios differed from the baseline scenario only in terms of school-related contacts, while all other epidemiological parameters, non-school interventions, and past policies were kept unchanged."

(Lines 218-222, main) "[...] Rather than evaluating globally uniform policies, such as schools being always open or always closed [38], we focused on period-specific policy decisions made under the epidemiological and social conditions prevailing at each phase of the pandemic. This approach reflects how school policies were considered and implemented in practice, enabling assessment of their impact in realistic, historically grounded contexts."

Alternative designs based on globally consistent or long-term policies, such as schools always open or always closed, correspond to different research questions and have been explored elsewhere in the literature, for example:

"*Estimating the impact of school closures on the COVID-19 dynamics in 74 countries: A modelling analysis*" Ragonnet et al. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1004512).

While informative, such approaches aim to quantify the cumulative impact of sustained interventions over the entire epidemic and are less suited to disentangling the effects of specific policy decisions. Our counterfactual design instead prioritizes retrospective policy evaluation under realistic conditions, reflecting how school-opening and school-closing decisions were actually considered and implemented during the pandemic.

We also want to note that the grid search for some parameters that were not properly identified via filtering is done so that the *baseline* scenarios - where all school policies are as they occurred in reality - match the observed data. Therefore, these parameter choices are not tied to counterfactual scenarios. We now emphasize this in the main text, in conjunction with explaining the estimation pipeline more clearly:

(Lines 398-405, main) "[...] The estimated parameters vary across the periods considered, reflecting the separate model fitting for each period, where the posterior distributions from one period were used to initialize the next period.

For each counterfactual period, we first performed initial model fitting for some parameters, where each model fit was implemented for a single value on a multidimensional grid over the remaining parameters. We then selected the value on the grid that minimized the distance between the baseline counterfactual and the observed data, and refitted the model with these parameter values to obtain the posterior distributions used in the baseline simulations (see Supplementary Material, Section 3). This step was essential for internal model validity, given the variations in data and epidemic context over time."

(Lines 67-70, main) “Each counterfactual period was evaluated in isolation, conditional on the observed epidemic trajectory up to that point in time. The counterfactual scenarios differed from the baseline scenario only in terms of school-related contacts, while all other epidemiological parameters, non-school interventions, and past policies were kept unchanged.”

(Lines 218-222, main) “Rather than evaluating globally uniform policies, such as schools being always open or always closed [38], we focused on period-specific policy decisions made under the epidemiological and social conditions prevailing at each phase of the pandemic. This approach reflects how school policies were considered and implemented in practice, enabling assessment of their impact in realistic, historically grounded contexts.”

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Canfora et. al. did a truly commendable job revising this paper – I think it's the most thorough paper review I've ever seen. I focused on the sections related to q and the sensitivity analyses as that was where my initial comments were centered. I will let the other reviewers comment on the fitting procedures.

Overall, I have a much better understanding of the q value now. After going through the changes though, I am left with the following questions:

Response: We sincerely thank the reviewer for the positive and encouraging feedback on the revised manuscript. We are pleased that the revisions helped clarify the role of the q parameter and the associated sensitivity analyses. Below, we address the reviewer's remaining questions in detail.

-In the supplement when you write "As there are several values that satisfy (S7), we selected the minimum q(j)…" (just before S8), you mean that there are multiple q values for the different age groups right? But each one solves the equality in Eq S7 (even though it's presented as an inequality), is that right? So you take the minimum value across age groups, not the minimum value for a given age group that could apply to > instead of = in Eq S7?

Response: We thank the reviewer for this helpful comment. Yes, this explanation is correct. Equation S7 yields a value of q^(j^*) for each age group pair (3 age groups yield 9 pairs). Since multiple values satisfy the condition in Eq. S7, we selected the minimum value across the different age group pairs. This choice ensures that the resulting parameter does not lead to unrealistic reductions in non-school contacts while remaining consistent with the constraint described in the manuscript. To clarify this point, we have revised the text in the Supplementary Materials accordingly.

(Lines 832-833, Supplement): "As there are several values that satisfy \eqref{eq:q_value}, we selected the minimum q^(j^*) across age group pairs to prevent unrealistic reductions in non-school contacts."

-Something seems off in the survey dates in Table S1.. do surveys 5 and 6 overlap? And is the end date correct for survey 7?

Response: We thank the reviewer for pointing this out. There was a typographical error in the maximum dates reported for Survey 5 and Survey 7 in Table S1. These dates have now been corrected in the revised Supplementary Materials.

-Supplementary lines 850-852: I thought ckk^1 is pre-pandemic? Not 1st lockdown?

Response: We thank the reviewer for pointing this out. The reviewer is correct: ckk^1 refers to the pre-pandemic contact matrix, whereas ckk^2 corresponds to the first lockdown period. The reference in the Supplementary Materials was due to a typographical error and has now been corrected in the revised version.

-I think the grids in S3/S4 could be formatted to be a bit bigger which would help with readability

Response: We thank the reviewer for commenting on this. The figure has been revised to use larger font sizes for the text. The colorbars are also set to be common among the subplots.

-In your sensitivity analyses, when you changed q, did you also adjust non-school contacts? Or are total contacts potentially higher/lower when you perturb q?

Response: We thank the reviewer for this question. When q is perturbed, non-school contacts are adjusted accordingly so that the total number of contacts remains equal to the survey totals, as described in the first equality in Eq. S7. In other words, increasing q increases school contacts while proportionally reducing non-school contacts, resulting in a redistribution of contacts rather than a change in the total number of contacts.

However, there are a few exceptions. As explained in Supplement Section 2.1, in specific time intervals (F at the end of CP2/beginning of CP3 and L in CP4), to better reproduce the observed data, an additional reduction/increase in non-school contacts is applied after preliminary non-school contacts are calculated as the difference between total contacts and school contacts multiplied by q. Preliminary non-school contacts are still calculated as the total survey contacts minus the perturbed school contacts. However, the same reduction/increase as in the original baseline scenarios is then applied to non-school contacts; therefore, the final total contacts may differ slightly from the original survey totals, which explains the small differences observed in the perturbed baseline scenarios compared to the original baseline scenarios.

-S22/24 figures are very small – can you make the figure size closer to S23?

Response: We thank the reviewer for this observation. Figures S22 and S24 have now been expanded to two pages to clearly show the impact of each increase in q(t) on the model outcomes.

- "As mentioned in Supplement Section 2.1, in most time intervals, the increase in q just redistributes survey contacts by increasing school and reducing non-school contacts. This implies that total contacts are the same and therefore in principle the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution".
I don't understand this. If you change q, then the non-school contacts change to preserve the same total contacts (Eq S7). When you run your baseline scenario then, would it not be possible for the baseline to be different now (even without changing school opening/closure) because contacts outside of school are different?

Response: We thank the reviewer for raising this point. In principle, changing q does not modify the total number of contacts, but only redistributes them between school and non-school settings, as imposed by the equality in Eq. S7. Under this condition, the baseline scenarios remain unchanged because only total contacts matter for the baseline; however, for counterfactuals relative to that baseline, changes in school contacts matter.

As explained in Supplement Section 2.1 and above, though, this condition does not hold strictly in a few specific time intervals, such as F, H, and L. For example, in intervals F and L, non-school contacts are further reduced after preliminary non-school contacts are obtained by subtracting the perturbed school contacts from the survey totals. The reduction factors we applied to non-school contacts are the same in the original analysis and in the sensitivity analyses. As a result, the baseline total number of contacts in the sensitivity analysis may differ from the baseline total contacts in the original analysis, depending on the value of q applied.

Importantly, in CP2 and CP3, the only affected interval is F, where all perturbations lead to the same capped value q=1; therefore, the baseline results remain identical across perturbations. This is not the case for CP4, where small differences in the baseline emerge.

For clarity, we present below the relevant lines from Section 2.1 of the Supplement, where this point is explained in detail.

(Lines 856-864, Supplement): "In these time intervals (F, H, and L), we adjusted the non-school contacts upward or downward – to capture a further relaxation or a restriction, respectively – with a factor above or below one, in the latter case imposing the first lockdown (non-school) contacts as the lower bound. [...]"

(Lines 865-871, Supplement): "Note that with these adjustment factors in non-school contacts, the values of $q(t)$ influence the final total contacts in some time intervals, an insight relevant also to the sensitivity analysis in Section 10. [...] In these time intervals, the value of $q(t)$ changes both the final non-school contacts and the final total contacts relative to their survey equivalents."

We also report below the relevant lines from Supplement Section 10.

(Lines 997-1004, Supplement): "As mentioned in Supplement Section \ref{suppsubsect:q_calculation}, in most time intervals, the increase in q just redistributes survey contacts by increasing school and reducing non-school contacts.
This implies that total contacts are the same, and therefore, in principle, the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution. However, as shown in Table \ref{tab:surveys_periods_2y}, Figure \ref{fig:contactss}, and explained in Section \ref{suppsubsect:q_calculation}, in time intervals F at the end of CP2 and beginning of CP3, and L in CP4, non-school contacts were reduced.\footnote{The non-school contacts were increased in time interval H, but there is no q in the baseline of that time interval, as schools were closed; therefore, total baseline contacts are the same in H as well.}
Since this reduction is calculated after preliminary non-school contacts are obtained by subtracting the perturbed school contacts from the total survey contacts, the total contacts in the perturbed baseline scenario can differ from those in the original baseline scenario, resulting in slightly different baselines in CP2–CP4."

-"The baseline cumulative hospitalizations for q not perturbed and perturbed are shown in Table S7, and they are identical in CP2-CP3 for different perturbations of q, because the only time period where baseline total contacts differ from the original q is time interval F, and in that time period all perturbation lead to the same capped value q = 1, while this is not the case for CP4. As Table S7 shows, the CP2 and CP4 baseline results are very similar relative to q not perturbed. However, the CP3 baseline hospitalizations are much larger when q is perturbed, and as a result the baseline does not match the observed data as in the original analysis, making it an unrealistic scenario."
I'm having trouble following this paragraph. Firstly nothing seems "identical" to me in Table S7, unless I'm missing something. Second the paragraph reads as saying things are identical in CP3 when q is perturbed first, but then later says baselines are much larger in CP3 when q is perturbed. Also, if q is increased in CP3, then shouldn't non-school contacts go down? Even if you then adjust non-school contacts, overall this value should be lower shouldn't it? So then why do you say contacts are higher resulting in the higher effective R and higher effective hospitalizations?

Response: We thank the reviewer for raising this point and for the opportunity to clarify this paragraph.

First, when we state that the baseline hospitalizations are *identical in CP2–CP3*, we mean that within each of these periods, the baseline results are identical across the different perturbations of q (i.e., +2%, +5%, +10%, +20%). This occurs because the only time interval where baseline total contacts differ from the original q is interval F, and in that interval, all perturbations lead to the same capped value q=1. As a result, all perturbations produce the same baseline outcome within CP2 and within CP3.

For CP3, baseline hospitalizations are much larger when q is perturbed compared to when they are not, because perturbing q changes total contacts as well, leading to different parameter estimates. As a result, the perturbed baseline scenario no longer matches the observed data as in the original analysis, making it an unrealistic scenario.

If q is increased in CP3, non-school contacts decrease only in the first interval of that period (interval F). However, the school contacts increase, on balance, making the total contacts higher. In the following intervals (G and H), non-school contacts are not calculated as total contacts minus school contacts, as explained in Supplementary Section 2.1. Instead, they are taken from the original analysis: from Survey 2 for G, and from Survey 2 with non-school contacts increased for H. In H, schools are closed, so the q perturbation does not apply. Considering the whole CP3 period, this results in higher total contacts compared to the unperturbed case. The higher total contacts lead to an earlier increase of the median R_t above one and a longer period with $R_t>1$, resulting in a larger hospitalization wave.

For clarity, we report below the relevant lines from Supplement Section 2.1 where this point is explained in detail.

(Lines 852–864, Supplement): "In the other time intervals just before implementation of the second and third lockdown, G and M in Table \ref{tab:surveys_periods_2y}, we set the non-school contacts equal to the first lockdown (non-school) contacts $\widetilde{c}_{k k^*}^{(2)}$. In the rest of the time intervals, we noticed that the baseline counterfactuals did not closely reproduce the actual data, due to contact surveys covering long periods and not reflecting all the gradual restrictions or relaxations in those periods (Survey 4, Survey 7). In these time intervals (F, H, and L), we adjusted the non-school contacts upward or downward – to capture a further relaxation or a restriction, respectively – with a factor above or below one, in the latter case imposing the first lockdown (non-school) contacts as the lower bound. These factors were selected in each of the periods F, H and L via a grid search as follows: (i) we fitted the model for a range of these factors, and all the parameter posteriors were retrieved; (ii) we simulated 300 ensembles from a baseline counterfactual for each CP with these parameter posteriors and the same factor; (iii) we calculated the sum of the absolute distance between daily infection and hospitalization data and their corresponding median across 300 ensembles of the baseline counterfactual simulated infections and hospitalizations; (iv) we picked the factor that minimized the distance in (iii)."

We also revised the text in Section 10 to make these points clearer.

(Lines 989–994, Supplement): "These modified values pertaining to a particular CP were then used to refit the model, re-estimate the parameters fitted via the ensemble adjustment Kalman filter, and recompute all outputs exactly as in the main analysis, except that the grid searches for the other parameters and for the adjustment factors in non-school contacts were not repeated. For the latter, the grid-searched values obtained in the original analysis were preserved, so that the non-school contact adjustments remained the same in both the original baseline scenario and the perturbed baseline scenarios, ensuring comparability between them."

(Lines 997–1004, Supplement):"As mentioned in Supplement Section \ref{suppsubsect:q_calculation}, in most time intervals, the increase in q just redistributes survey contacts by increasing school and reducing non-school contacts.
This implies that total contacts are the same, and therefore, in principle, the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution. However, as shown in Table \ref{tab:surveys_periods_2y}, Figure \ref{fig:contactss}, and explained in Section \ref{suppsubsect:q_calculation}, in time intervals F at the end of CP2 and beginning of CP3, and L in CP4, non-school contacts were reduced.\footnote{The non-school contacts were increased in time interval H, but there is no q in the baseline of that time interval, as schools were closed; therefore, total baseline contacts are the same in H as well.}
Since this reduction is calculated after preliminary non-school contacts are obtained by subtracting the perturbed school contacts from the total survey contacts, the total contacts in the perturbed baseline scenario can differ from those in the original baseline scenario, resulting in slightly different baselines in CP2–CP4."

(Lines 1005–1014, Supplement): "The baseline cumulative hospitalizations for q not perturbed and perturbed are shown in Table \ref{tab:SA1_baseline}. The baseline values are identical within CP2 and within CP3 across the different perturbations of q, because the only time period where baseline total contacts differ from the original q is time interval F, and in that time period, all perturbations lead to the same capped value $q=1$, while this is not the case for CP4. As Table \ref{tab:SA1_baseline} shows, the CP2 and CP4 baseline results are very similar relative to q not perturbed. However, the CP3 baseline hospitalizations are much larger when q is perturbed compared to when they are not, and as a result, the baseline does not match the observed data as in the original analysis, making it an unrealistic scenario. Figure \ref{fig:SA1_cp3} Panel C shows why this is the case: since total contacts are higher at the beginning of the period relative to the case of no increase in q, the median $\mathcal R_t$ increases earlier above one in the period and stays above one for longer, resulting in a higher and longer hospitalization wave despite the small increase in q. "

-I think you forgot to update the caption of figure S4 to reflect the visual changes you made

Response: Thank you for this observation. We now revised it to point out the elements of Figure S4. In addition, the y-label is revised from "school capacity" to "school opening index", in line with the term used in Boldea et al. (2024).

-"School capacity" is a slightly confusing term because it sounds like the total number of students the school can accommodate. What you're varying is the extent of closure/opening. I would consider changing the terminology.

Response: Thank you for this comment. In response to this, we decided to use the term "school opening index" which describes the proportion of school contacts remaining. With that, we have revised the following:

- "Sensitivity analysis 2: different **school capacity**"
  - changed to "Sensitivity analysis 2: different **levels of school opening index**" (Line 1023, Supplement)
- "we performed an additional sensitivity analysis ... at different **levels of capacity**"

    - changed to “we performed an additional sensitivity analysis ... at different **levels of school opening index**” (Lines 1028-1030, Supplement)
    - added the sentence: “These school opening indices describe the proportion of school contacts remaining.” (Lines 1030-1031, Supplement)
- “Across all counterfactual periods, changing **school capacity**... ”
    - changed to “Across all counterfactual periods, changing **the school opening index**... ” (Line 1035, Supplement)
- “The relative ordering ... across the range of **capacities** considered...”

changed to “The relative ordering ... across the range of **school opening indices** considered...” (Lines 1038-1039, Supplement)

-Figures S25-28: what are those values? Are those numbers of infections (i.e. # of infections in perturbed scenario - # in baseline) or percentages?

Response: We thank the reviewer for this comment. The values reported in Figures S25–S28 represent percentage changes relative to the baseline scenario, as indicated in the figure captions. To make this even clearer, we have now explicitly specified “relative change (%)” in the figure labels in the revised Supplementary Materials (Figures S25-S28).

-In your sensitivity analysis when you vary non-school contacts (Section 12 in the supplement), you’re not preserving total contacts right? So the total number of contacts could be changing? Related, I think this sentence is worth noting in the main too if you didn’t already “This suggests that sufficiently strong reductions in adult non-school contacts could, in principle, have allowed higher levels of school activity in some periods without leading to large increases in infections and hospitalizations.” This is an interesting finding given the large negative impacts of school closure, and is something to consider when coming up with how to prioritize public health decisions.

Response: We thank the reviewer for this observation. Yes, this is correct. In the sensitivity analysis described in Supplement Section 12, non-school contacts are varied directly, and therefore, the total number of contacts is not preserved. As a result, the total contacts can change in those scenarios. We include a sentence in the Supplementary Material pointing this out.

(Lines 1053-1054, Supplement): “... It should be noted that these simulated changes to non-school contacts do not preserve the total contacts used in the main analyses. ...”

The suggested sentence is also included in the main text:

(Lines 194-196, Main text): “Our findings also suggest that sufficiently strong reductions in adult non-school contacts could, in principle, have allowed higher levels of school activity in some periods without leading to large increases in infections and hospitalizations.”

Reviewer #2 (Remarks to the Author):

Thank you for addressing all of my comments so thoroughly! I have no further comments and recommend the manuscript for publication.

Reviewer #2 (Remarks on code availability):

I had reviewed the CP1 code in the first revision.

Response: We are grateful to the reviewer for the positive assessment of our manuscript and for the helpful comments provided during the review process. We are pleased that the revisions have satisfactorily addressed all the issues raised.

Reviewer #3 (Remarks to the Author):

I thank the authors for addressing all my comments in a comprehensive and satisfactory way. I have no further comments and congratulate the authors for their work.

Response: We sincerely thank the reviewer for the positive feedback and for the time and effort dedicated to evaluating our manuscript. We greatly appreciate the reviewer's constructive comments throughout the review process, which have helped improve the clarity and quality of our work. We are pleased that the revisions have addressed the concerns satisfactorily.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors did a great job addressing my comments. I have a couple remaining things I don't quite follow, but this shouldn't affect the decision to proceed with publication.

Response: We thank the reviewer for the positive assessment of our revision and for the helpful remaining comments. We have clarified the relevant points in the manuscript to improve overall clarity, as detailed below.

-Supplement 870/871: Do you mean something like "In these OTHER time periods…"

Response: We thank the reviewer for this helpful suggestion. We agree that the phrasing was ambiguous, as the reference of "these time intervals" was not sufficiently clear. In the revised version, we explicitly distinguish the intervals in which total contacts are preserved from those in which non-school contacts are further adjusted, to avoid confusion.

(Lines 879-881, Supplement):"In the specific time intervals where non-school contacts are further adjusted, the value of $q(t)$ changes both the final non-school contacts and the final total contacts relative to their survey equivalents."

-When you describe the other factor used in periods F,H,L, your fitting procedure is not 100% clear. What is a baseline counterfactual? My understanding is you choose different factor values, then fit your model in one CP to the actual data to get the values of all other parameters, simulate 300 times, and then find the error between the hospitalization data and your median. You pick the factor that results in the smallest error. But I'm confused what "baseline counterfactual" means – the baseline is not counterfactual, is it?

Response: We thank the reviewer for this comment and for the accurate summary of our procedure.

In our framework, model parameters are first estimated by fitting the model to the observed data within each counterfactual period. Counterfactual analyses are then performed by sampling from the posterior distributions of these parameters, without further re-estimation.

We agree that the term "baseline counterfactual" may be confusing. By this, we refer to the simulation that reproduces the observed (historical) conditions, using the estimated parameters, and serves as a reference for comparison with alternative scenarios. It is therefore not counterfactual in the strict sense, but rather a reference simulation within the counterfactual framework.

This baseline is used to assess how well the model reproduces the observed data and to provide a consistent point of comparison for the counterfactual scenarios. As shown in Figure S14, the close agreement between the fit and the baseline scenario supports the reliability of the parameter estimates and the subsequent scenario analyses.

We have revised the manuscript to clarify this terminology. In the main text, we now refer readers to Supplementary Materials, Section 4, where this point is explained in detail. In addition, we have replaced the term "baseline counterfactual" with "baseline scenario" to improve clarity.

(Lines 431-433, Main): “After fitting the model for each of the four periods to observed infections and hospitalizations, we used it to generate the corresponding baseline scenarios (see Supplementary Figure \ref{fig:CPs_fitbaselinevsdata} and Supplementary Material, Section \ref{suppsect:baseline_fit} for details).”

(Lines 917-922, Supplement): “The model fit corresponds to the trajectory obtained during the data assimilation procedure (ensemble adjustment Kalman filter), where model states and parameters are sequentially updated to match the observed data.
The baseline scenario is generated after the fitting process by running the model using parameter values drawn from the posterior distributions at the end of the fitted period, while accounting for stochastic variability across multiple realizations.
Thus, the fit reflects the outcome of the data assimilation procedure, whereas the baseline scenario captures the propagation of uncertainty in the forward simulations.“

-Out of curiosity, why is your fit in Figure S14 different from your baseline simulation for each CP? Is that just the result of stochasticity and the fitting procedure?

Response: We thank the reviewer for this question. The difference between the model fit and the baseline simulation in Figure S14 arises because they represent two different stages of the analysis.

- The model fit corresponds to the trajectory generated during the data assimilation procedure (ensemble adjustment Kalman filter), where model states and parameters are sequentially updated to match the observed data.
- The baseline simulation is generated after the fitting process. It corresponds to the solution of the model using the final estimated parameters. It also accounts for stochastic variability propagated across multiple realizations.

Figure S14 shows the median across ensembles in both cases. The fit reflects the outcome of the data assimilation process, whereas the baseline simulation reflects posterior uncertainty through forward simulations. The small discrepancies between the two are therefore expected.

To clarify this point, we have revised the text in the Supplementary Materials accordingly.

(Lines 917-922, Supplement): “The model fit corresponds to the trajectory obtained during the data assimilation procedure (ensemble adjustment Kalman filter), where model states and parameters are sequentially updated to match the observed data.
The baseline scenario is generated after the fitting process by running the model using parameter values drawn from the posterior distributions at the end of the fitted period, while accounting for stochastic variability across multiple realizations.
Thus, the fit reflects the outcome of the data assimilation procedure, whereas the baseline scenario captures the propagation of uncertainty in the forward simulations.“

-I’m stumped by this, which I also asked in my last round of questions, and you say it or something similar several times: “As mentioned in Supplement Section \ref{suppsubsect:q_calculation}, in most time intervals, the increase in q just redistributes survey contacts by increasing school and reducing non-school contacts. This implies that total contacts are the same, and therefore, in principle, the baseline scenarios for CP2-CP4 should be the same, and only the counterfactual scenarios should change due to this redistribution.” Why does only total contacts matter for baseline scenarios if you’re explicitly modeling school and non-school interactions, and these have different governing dynamics? To me if you change q, even keeping the opening/closure of schools the same (i.e. not introducing the counterfactual), the total number of hospitalizations may change. This leads to confusion for me to understand why varying q in your sensitivity analyses doesn't change your "baseline" (even when you don't have the issue of total contacts changing due to the adjustment factor). Why would this not be the case? For the sensitivity analyses is it because you recalibrate the model with the new qs?

Response: We thank the reviewer for this important point.

We agree that, in principle, redistributing contacts between school and non-school settings could affect transmission dynamics, even if the total number of contacts remains unchanged, due to their different epidemiological roles.

However, in our framework, for each value of q we recalibrate the model parameters to observed data within each counterfactual period. This recalibration step allows to just redistribute the contacts. In the model used for fitting and the baseline scenario – Supplement, Section 1 – only the total number of contacts matter, and not how they are distributed across school and non-school contacts, because the transmission outside school is not separately modelled from transmission within schools. Therefore, in cases where the non-school contacts are not adjusted with an additional factor, so in cases where the total contacts are equal to survey contacts, the baseline scenario don't change. However, in the counterfactuals, school closers/openings do change the total number of contacts, leading to changes in the epidemic trajectory.

We have clarified this point in the manuscript to better explain the role of recalibration in the baseline scenarios.

For the sensitivity analysis, the same thing applies: if non-school contacts were not adjusted with an additional factor besides q, and total contacts were still equal to survey contacts, the baseline scenarios would be the same, however the counterfactuals change when changing q.

(Lines 1010-1013, Supplement): "This recalibration step ensures consistency with the observed data despite the redistribution of contacts. In the model, transmission depends on the total number of contacts, as transmission outside and within schools is not modeled separately. As a result, differences in the baseline scenarios are limited."