THE ROYAL SOCIETY
PUBLISHING

PROCEEDINGS B

Sounds of sickness: can people identify infectious disease using sounds of coughs and sneezes?

Nicholas Michael Michalak, Oliver Sng, Iris M. Wang and Joshua Ackerman

Article citation details
Proc. R. Soc. B **287**: 20200944.
http://dx.doi.org/10.1098/rspb.2020.0944

Review timeline

Original submission:	22 July 2019
1st revised submission:	15 November 2019
2nd revised submission:	25 April 2020
3rd revised submission:	13 May 2020
Final acceptance:	14 May 2020

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2019-1719.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Reject – article is not of sufficient interest (we will consider a transfer to another journal)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Acceptable

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
The present manuscript concerns how well people can judge whether coughs/sneezes are made by infectious or non-infectuous individuals. The stimuli material, 40 audio-clips, of coughs (n=20) and sneezes (n=20), from youtube clips. The control stimuli came from non-infectious sounds in response to environmental irritants (such as allergies, maul stimulation of nasal passages etc). The authors then performed 4 separate studies on MTurk to these clips to see how accurately they could determine whether the clips were from infectious or non-infectious individuals. The results show that people were not able to accurately do this. However, they do find that the more discusting a sound is the more likely they are to judge the person as infectious. Certainty of the answer was also a predictor increasing the change for judging a clip as infectious. The article is written in excellent English. While the manuscript is contributing with new knowledge, there are a number of aspects that reduces my enthusiasm for the paper as it is now. See comments below.

Major concerns.
1) The study focus on how well one can discern coughs/sneezes from infectious vs non-infectious individuals. While this is interesting, a more appropriate questions to ask, at this stage in what we know in the field, are ”Do humans use sounds to decide whether someone is sick or not?” and ”What sounds do we use to tell if someone is sick?”. The usage of coughs and sneezes are of course probable sounds that help us judge if someone is sick or not. However, it does not seem likely that evolution has strongly advocated a selection for detecting which coughs that have a pathogenic origins or an alternative cause. Here, evolution has probably favored those thinking that coughs and sneezes may be signals that we should be careful to interact with this individual. While the article is interesting, it is more focused on the detail to whether humans have different kind of coughs and sneezes depending on whether they are driven by pathogens or of other origin. However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc. Additionally, the authors do tap into why some coughs and sneezes are judged differently, and the mechanisms for this would be an interesting venue as well to explore.

2) It is hard to judge the quality of the audio-clips. Where the infectious subjects really infectious at the time of collection and from what pathogen? And where the control clips really from people who were not carriers of any pathogens? A strength with previous studies is the usage of stimuli material where sickness had been experimentally controlled, and where subjects were properly healthy in the control condition.

3) The use of MTurk is common and tempting approach for rating projects. A major problem is that some are not so interested in doing their best rather than earning money as fast as possible, for example the existence of ″Super Turkers″ shows this point. Several approaches has been developed to combat such behavior, such as inclusion of control questions or removing subjects who do not vary much in their responses or respond very fast. The authors have not shown any approach dealing with this problem. Hence, the material is very likely to include a lot of error variance, and hence has much lower power than stated.

4) The authors measure accuracy, a combination of correctly identifying both infectious and non-infectious responses in relation to failures. Since they also discuss costs of their decisions (eg. Row 90 onwards), they should also include information, at least some, on specificity and sensitivity.

5) Fig 2 needs better explanations. The information is confusing. A reader does not understand what ″sound rating″ or ″sound origin″ means. It is much better to present it in terms of ″coughs/sneezes″ and ″infectious individuals″. Now, ′sound origin′ is explained as ″different colors and types″, which is unclear. Also clarify what the results are so people understand the purpose of the figure.

Minor comments
- The authors discuss the costs of detection and bias in detection, and trait differences in detection accuracy. There are a few recent analyses and commentaries on this in Proceedings B last year. See Kurves and Wolf 2018 and response by Axelsson et al 2018. These data are in opposite of the findings here (i.e. Kurves find very strong individual differences in strategy) and should be discussed (at least in part by the limitations of using MTurk).

- Row 158. Explain what ″sound stimuli″ is.

- Row 176. Unclear what the reference ″2″ refers to.

- Row 179. There is a reference for Figure 1 when reporting study 1. However, there is no data plotted for study 1, only from all the other data collections. This is confusing.

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> Yes
>
> **Is it clear?**
> Yes
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
I have very rarely recommended a paper for publication without any edits, but was very close to with this one. It's a comprehensive set of studies, transparently reported, with extremely clear codebooks and code. Excellent. My only substantial recommendation is that the authors bring a little more clarity to their description of the stimuli. I also thought it might be useful for the authors to discuss recent studies in the visual domain suggesting that previously reported links between susecptibility to infectious illnesses and facial cues are not ribust (Ziyi Cai et al. and Yong Foo et al's recent papers, for example).

Review form: Reviewer 3 (Leonid Tiokhin)

Recommendation
Reject – article is not of sufficient interest (we will consider a transfer to another journal)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Marginal

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
No

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
See attached. (See Appendix A)

Decision letter (RSPB-2019-1719.R0)

29-Aug-2019

Dear Mr Michalak:

I am writing to inform you that we have now obtained responses from referees on manuscript RSPB-2019-1719 entitled "Sounds of sickness: Can people identify infectious disease using auditory cues?" which you submitted to Proceedings B.

Unfortunately, on the advice of the Associate Editor and the referees, your manuscript has been rejected following full peer review. As you will see, the reviews are mixed, with all reviewers noting enthusiasm for your study, but two of the reviewers highlighting a number of concerns, particularly about the reliability of the stimuli and potential confounds in your data. Competition for space in Proceedings B is currently extremely severe, as many more manuscripts are submitted to us than we have space to print. We are therefore only able to publish those that are exceptional, convincing and present significant advances of broad interest, and must reject many good manuscripts. Given the nature of the concerns, it is unlikely that even with a substantive revision your manuscript would be accepted to Proceedings B. I know that this is frustrating and I am sorry not to be more positive on this occasion.

On a more positive note, however, based on the advice we have received, we would like to offer you the opportunity to transfer your manuscript file to another Royal Society journal, Royal Society Open Science, which we think may be an excellent fit for your study. Royal Society Open Science is a fast, open journal publishing high-quality research across all of science and mathematics. The journal operates objective peer review, optional open peer review, and will publish any article deemed to sufficiently advance the field by the reviewers and editors, leaving judgement of potential impact of the work to the reader. The journal publishes Registered Reports and encourages the submission of negative results. You can find out more about the scope of the journal and the benefits of publication here https://royalsocietypublishing.org/journal/rsos

If you wish to have your manuscript transferred to Royal Society Open Science please ensure that you revise your text to address all of the reviewers' comments relating to scientific soundness.

Please particularly ensure that your conclusions do not overstate the results of your study. Once submitted to Royal Society Open Science your manuscript will be assessed by an Associate Editor who will decide whether further reviewer advice is required. If no further advice is needed and all of your revisions are satisfactory your manuscript will be immediately accepted for publication.

If you agree to transfer your paper, and it is accepted for publication, you will be asked to pay the article processing charge, unless you request a waiver and this is approved by Royal Society Publishing. You can find out more about the charges at https://royalsocietypublishing.org/rsos/charges.

You can approve or reject this transfer using the links below:

Approve transfer - *** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/prsb?URL_MASK=4c6d2964793d4eb28183dc3acfe08b47
After approving the transfer you will need to log in to your Royal Society Open Science author centre (https://mc.manuscriptcentral.com/rsos) to complete your the submission. At this stage you will have chance to address any of the reviewers' or editor's concerns.

Reject transfer - *** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/prsb?URL_MASK=06d1df15c1904c96bca2399124f8a819

or by clicking 'approve' or 'reject' in your Author Center.

Once you have approved the transfer you will be prompted to complete the transfer of your article via the Royal Society Open Science submission system.

Please find below the comments received from referees concerning your manuscript, not including confidential reports to the Editor. If you approve transfer to Royal Society Open Science, these reviews will accompany your paper.

Thank you for your interest in Proceedings B.

Sincerely,
Dr. Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
Your manuscript has now been reviewed by three experts in the field. Although all three reviewers had positive things to say about your manuscript, on balance I think that the critical comments outweigh them. Moreover, I am not convinced that a revision would be able to deal with some of the comments that were raised -- particularly with respect to possible confounds. In the end, I must recommend rejection. You might consider submitting your manuscript to Royal Society Open Science, an open journal publishing high-quality original research across the entire range of science on the basis of objective peer-review.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)

The present manuscript concerns how well people can judge whether coughs/sneezes are made by infectious or non-infectuous individuals. The stimuli material, 40 audio-clips, of coughs (n=20) and sneezes (n=20), from youtube clips. The control stimuli came from non-infectious sounds in response to environmental irritants (such as allergies, maul stimulation of nasal passages etc). The authors then performed 4 separate studies on MTurk to these clips to see how accurately they could determine whether the clips were from infectious or non-infectious individuals. The results show that people were not able to accurately do this. However, they do find that the more discusting a sound is the more likely they are to judge the person as infectious. Certainty of the answer was also a predictor increasing the change for judging a clip as infectious. The article is written in excellent English. While the manuscript is contributing with new knowledge, there are a number of aspects that reduces my enthusiasm for the paper as it is now. See comments below.

Major concerns.

1) The study focus on how well one can discern coughs/sneezes from infectious vs non-infectious individuals. While this is interesting, a more appropriate questions to ask, at this stage in what we know in the field, are ”Do humans use sounds to decide whether someone is sick or not?” and ”What sounds do we use to tell if someone is sick?”. The usage of coughs and sneezes are of course probable sounds that help us judge if someone is sick or not. However, it does not seem likely that evolution has strongly advocated a selection for detecting which coughs that have a pathogenic origins or an alternative cause. Here, evolution has probably favored those thinking that coughs and sneezes may be signals that we should be careful to interact with this individual. While the article is interesting, it is more focused on the detail to whether humans have different kind of coughs and sneezes depending on whether they are driven by pathogens or of other origin. However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc. Additionally, the authors do tap into why some coughs and sneezes are judged differently, and the mechanisms for this would be an interesting venue as well to explore.

2) It is hard to judge the quality of the audio-clips. Where the infectious subjects really infectious at the time of collection and from what pathogen? And where the control clips really from people who were not carriers of any pathogens? A strength with previous studies is the usage of stimuli material where sickness had been experimentally controlled, and where subjects were properly healthy in the control condition.

3) The use of MTurk is common and tempting approach for rating projects. A major problem is that some are not so interested in doing their best rather than earning money as fast as possible, for example the existence of ”Super Turkers” shows this point. Several approaches has been developed to combat such behavior, such as inclusion of control questions or removing subjects who do not vary much in their responses or respond very fast. The authors have not shown any approach dealing with this problem. Hence, the material is very likely to include a lot of error variance, and hence has much lower power than stated.

4) The authors measure accuracy, a combination of correctly identifying both infectious and non-infectious responses in relation to failures. Since they also discuss costs of their decisions (eg. Row 90 onwards), they should also include information, at least some, on specificity and sensitivity.

5) Fig 2 needs better explanations. The information is confusing. A reader does not understand

what ″sound rating″ or ″sound origin″ means. It is much better to present it in terms of ″coughs/sneezes″ and ″infectious individuals″. Now, ′sound origin′ is explained as ″different colors and types″, which is unclear. Also clarify what the results are so people understand the purpose of the figure.

Minor comments
- The authors discuss the costs of detection and bias in detection, and trait differences in detection accuracy. There are a few recent analyses and commentaries on this in Proceedings B last year. See Kurves and Wolf 2018 and response by Axelsson et al 2018. These data are in opposite of the findings here (i.e. Kurves find very strong individual differences in strategy) and should be discussed (at least in part by the limitations of using MTurk).

- Row 158. Explain what ″sound stimuli″ is.

- Row 176. Unclear what the reference ″2″ refers to.

- Row 179. There is a reference for Figure 1 when reporting study 1. However, there is no data plotted for study 1, only from all the other data collections. This is confusing.

Referee: 2

Comments to the Author(s)
I have very rarely recommended a paper for publication without any edits, but was very close to with this one. It's a comprehensive set of studies, transparently reported, with extremely clear codebooks and code. Excellent. My only substantial recommendation is that the authors bring a little more clarity to their description of the stimuli. I also thought it might be useful for the authors to discuss recent studies in the visual domain suggesting that previously reported links between susecptibility to infectious illnesses and facial cues are not ribust (Ziyi Cai et al. and Yong Foo et al's recent papers, for example).

Referee: 3

Comments to the Author(s)
See attached.

Author's Response to Decision Letter for (RSPB-2019-1719.R0)

See Appendix B.

RSPB-2019-2674.R0

Review form: Reviewer 1 (John Axelsson)

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
I have read thoroughly through the manuscript and can see that the authors have made a number of improvements. While the videos are of non-experimentally controlled origin, and McTurk has a number of weaknesses, the authors have done a thorough job in maximizing the designs and also reasonably deal and argue of their validity. I also agree that the outcomes are scientifically sound, i.e. that humans do not seem very good at judging whether sneezes coughs are from a contagious origin stimulated by something else, and that rated disgust and certainty are not related to whether the sounds were from a contagious individual or not. There are, however, some aspects of the ms that are not properly addressed.

1) The title makes the reader believe they have tested whether people can identify infectious disease using auditory cues. This is a very bold statement considering that the authors have only included specific short sounds from sick people. For the statements made about the ability to detect sick sounds, they also have to show proof of concept that they actually used the most relevant sickness sounds. Since this is not done in the ms, only including particular sounds presented for a few seconds, the title should be updated to fit the question studied. For example. 'Sounds of sickness: Can people identify infectious disease using auditory clips of sneezes and coughs?'

2) It is up to the authors to disagree with my previous statement "…. However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc…" i.e. that there has NOT been a large evolutionary

pressure for developing different coughs and sneezes depending on the causation. It would be suitable for the authors to include/discuss this possibility in the limitations, particularly since it is a likely reason for the 'null findings'.

3) Going through the OSF-online material (also appreciating the effort to find sick and not sick related sounds), I could not open the links in the "url1_archive", and only the original flies.

4) I am surprised that individuals did not differ in their rating behavior or accuracy (see method, rows 156-159), This is very different from the analyses made by Kurvers &amp; Wolf, 2018, and large parts of the literature where individuals often show large differences in rating behavior. It seems appropriate to include 'rater' as a factor in all models since there are 4 separate studies and it may differ from study to study, and also report the ICC data from both 'rater' and 'sound originator'.

Review form: Reviewer 2

Recommendation
Accept as is

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> Yes
>
> **Is it clear?**
> Yes
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
I think the authors have done a superb job addressing the issues raised during the review process. I was disappointed to see so many comments raised that I felt were very clearly addressed in the supp mats and original submission, so it is great that the authors were given the opportunity to clarify those points in a revision. I think this makes an important and timely contribution to the ongoing debate about whether or not humans display reliable cues of health.

Review form: Reviewer 3

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Acceptable

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
No

Should the paper be seen by a specialist statistical reviewer?
Yes

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> Yes
>
> **Is it clear?**
> Yes
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Review of first round of revisions for:
Sounds of sickness: Can people identify infectious disease using auditory cues?
Manuscript ID: RSPB-2019_1719

TO THE EDITOR/authors:
The authors have submitted a revised version of the manuscript. In this revision, they claim that most major points raised by the referees have been addressed. I think that the revised manuscript represents an improvement in clarity and detail over the original submission. That said, I still have concerns regarding the ability of this study to test whether people can identify infectious disease using auditory cues.
My most substantive remaining concern, which I do not feel is adequately addressed, concerns the ability of the study design to adequately answer the research question. The use of user-uploaded Youtube stimuli mean that there are plausible confounds between conditions, which make it difficult to interpret an effect in any direction in this study. I think that it is the responsibility of the original authors to do everything possible to control for such confounds, and I do not think that the current design adequately achieves this. Infected videos may have been uploaded by people who are different in any number of ways (e.g., age, sex, non-infectious disease comordibity), or had a number of confounds (e.g., video quality). On the other hand, a finding of no effect is also difficult to interpret, because we do not actually know people's infectious disease status. We also don't know what the difference in the severity of the affliction is between the different types of people – are we comparing people with mild allergies to severe infectious diseases or vice versa? The authors attempt to address this by limiting analyses to cases where they are more confident that the individual is sick (e.g., because the video is uploaded from a hospital) but this does not solve the problem that 1) the study still relies on participants to accurately report their diagnosis and 2) you don't know the disease status of people who generated the non-sick stimuli (e.g., they may have been sick at the time of stuffing hot peppers in their nose; people with bronchitis may have had chronic bronchitis, which, in contrast to acute bronchitis, is not infectious). I do not mean to say that the study is entirely uninformative – the probability that the hospital people are indeed sick with some infectious disease is higher than the pepper-powder up-nose people – but it is hard to know what these probabilities are.

Less major concerns

I don't have a good sense for how the stimuli used in this study adequately capture the properties of infectious and non-infectious diseases that people are most likely to encounter. Some the stimuli (e.g., whooping cough) are infectious diseases that people don't encounter very often, while the one arguably most-frequently encountered by people (e.g., cold) is labelled as "cold maybe", so we don't know exactly what these people have. Participants are also provided with such minimal information (e.g., 1 second decontextualized clips) that it is not clear to me how informative a null effect is regarding people's ability to do this in the real world. This is indeed a different question than the one addressed in this paper (i.e., whether people can tell infectious disease status from short-audio clips of a small subset of diseases) but it seems to me to be closer to the broader question that the paper is attempting to address.
Other minor remaining issues which should be addressed are 1) the lack of discussion of why the study found strong evidence that participants are worse than chance and 2) the lack of detail regarding what exactly was specified in the pre-reg versus what was conducted (e.g., the pre-reg said things like "We will use repeated measures analyses and contrasts to examine the effect of condition on mean disgust ratings." without specifying the specific functional form of the analyses – the authors should be more transparent about this), and 3) lack of consistent presentation of p-values and confidence intervals (e.g., lines 198-202; line 241).

However, I think that other aspects of this paper are impressive in many ways. The study is laudably open and transparent and provides detailed information regarding stimuli, data, analysis code, and the extensive sensitivity analyses. As such, I think that the main results are robust to alternative ways of analyzing the data. The large sample size provides quite certain effect-size estimates, in aggregate (conditional on the design); and even if the representativeness/generalizability of the stimuli is unclear, I think it is interesting, in principle, to explore whether people can tell infectious-disease status from audio clips. It is interesting to know that people's level of disgust is a strong predictor of whether they judge someone as infected, regardless of the person’s actual infectious disease status, and that certain information (e.g., base rates) may not have much of an effect on people’s judgements.

I don’t think that it would be possible for the authors to address my major concern, or the first part of my first less-major concern, given that the design is what it is. However, I think that the other concerns could be plausibly addressed in a revision, should one be invited by the editor.

I hope that my comments will help the authors to strengthen their manuscript.

For accountability and transparency, I would like to sign this review.
Leonid Tiokhin

Decision letter (RSPB-2019-2674.R0)

12-Dec-2019

Dear Dr. Michalak,

Your revised manuscript RSPB-2019-2674 entitled "Sounds of sickness: Can people identify infectious disease using auditory cues?" has been seen again by the reviewers. Based on their advice, your manuscript has, in its current form, been rejected for publication in Proceedings B. Although one reviewer is very pleased with your revision, two others still raise substantive concerns that we would like you give you the opportunity to address. In particular, two of the reviewers continue to raise concerns about the validity and reliability of the user uploaded stimuli. I realize that there are good reasons to have done it this way, and that it is challenging to exclude all possible confounds from null results, but encourage you to include more justification or evidence that these stimuli were by and large representative of what the users claimed. I also agree with the third reviewer that you should consider an alternate title. With this in mind we would be happy to consider a resubmission, provided the comments of the referees are fully addressed. However please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful.
Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

1) A ‘response to referees’ document including details of how you have responded to the comments, and the adjustments you have made.
2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
3) Line numbers in your main document.

To upload a resubmitted manuscript, log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,
Prof Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor Board Member
Comments to Author:
Your re-submitted manuscript has now been reviewed. As you will see, although all the reviewers have positive things to say, two of them still raise some substantive concerns. I am recommending rejection, but with an opportunity to deal with these concerns in a revised manuscript. I realize that some of these concerns are difficult to handle but I think you should be given an opportunity to mute some of the issues that have been raised.

Reviewer(s)' Comments to Author:

Referee: 3

Comments to the Author(s).
Review of first round of revisions for:
Sounds of sickness: Can people identify infectious disease using auditory cues?
Manuscript ID: RSPB-2019_1719

TO THE EDITOR/authors:
The authors have submitted a revised version of the manuscript. In this revision, they claim that most major points raised by the referees have been addressed. I think that the revised manuscript represents an improvement in clarity and detail over the original submission. That said, I still have concerns regarding the ability of this study to test whether people can identify infectious disease using auditory cues.
My most substantive remaining concern, which I do not feel is adequately addressed, concerns the ability of the study design to adequately answer the research question. The use of user-uploaded Youtube stimuli mean that there are plausible confounds between conditions, which make it difficult to interpret an effect in any direction in this study. I think that it is the responsibility of the original authors to do everything possible to control for such confounds, and I do not think that the current design adequately achieves this. Infected videos may have been uploaded by people who are different in any number of ways (e.g., age, sex, non-infectious disease comordibity), or had a number of confounds (e.g., video quality). On the other hand, a finding of no effect is also difficult to interpret, because we do not actually know people's infectious disease status. We also don't know what the difference in the severity of the affliction is between the different types of people – are we comparing people with mild allergies to severe infectious diseases or vice versa? The authors attempt to address this by limiting analyses to cases where they are more confident that the individual is sick (e.g., because the video is uploaded from a hospital) but this does not solve the problem that 1) the study still relies on participants to accurately report their diagnosis and 2) you don't know the disease status of people who generated the non-sick stimuli (e.g., they may have been sick at the time of stuffing hot peppers in their nose; people with bronchitis may have had chronic bronchitis, which, in contrast to acute bronchitis, is not infectious). I do not mean to say that the study is entirely uninformative – the probability that the hospital people are indeed sick with some infectious disease is higher than the pepper-powder up-nose people – but it is hard to know what these probabilities are.

Less major concerns

I don't have a good sense for how the stimuli used in this study adequately capture the properties of infectious and non-infectious diseases that people are most likely to encounter. Some the stimuli (e.g., whooping cough) are infectious diseases that people don't encounter very often, while the one arguably most-frequently encountered by people (e.g., cold) is labelled as "cold maybe", so we don't know exactly what these people have. Participants are also provided with such minimal information (e.g., 1 second decontextualized clips) that it is not clear to me how informative a null effect is regarding people's ability to do this in the real world. This is indeed a different question than the one addressed in this paper (i.e., whether people can tell infectious disease status from short-audio clips of a small subset of diseases) but it seems to me to be closer to the broader question that the paper is attempting to address.
Other minor remaining issues which should be addressed are 1) the lack of discussion of why the study found strong evidence that participants are worse than chance and 2) the lack of detail regarding what exactly was specified in the pre-reg versus what was conducted (e.g., the pre-reg said things like "We will use repeated measures analyses and contrasts to examine the effect of condition on mean disgust ratings." without specifying the specific functional form of the analyses – the authors should be more transparent about this), and 3) lack of consistent presentation of p-values and confidence intervals (e.g., lines 198-202; line 241).

However, I think that other aspects of this paper are impressive in many ways. The study is laudably open and transparent and provides detailed information regarding stimuli, data, analysis code, and the extensive sensitivity analyses. As such, I think that the main results are robust to alternative ways of analyzing the data. The large sample size provides quite certain effect-size estimates, in aggregate (conditional on the design); and even if the representativeness/generalizability of the stimuli is unclear, I think it is interesting, in principle, to explore whether people can tell infectious-disease status from audio clips. It is interesting to know that people's level of disgust is a strong predictor of whether they judge someone as infected, regardless of the person's actual infectious disease status, and that certain information (e.g., base rates) may not have much of an effect on people's judgements.

I don't think that it would be possible for the authors to address my major concern, or the first part of my first less-major concern, given that the design is what it is. However, I think that the other concerns could be plausibly addressed in a revision, should one be invited by the editor.

I hope that my comments will help the authors to strengthen their manuscript.

For accountability and transparency, I would like to sign this review.
Leonid Tiokhin

Referee: 2

Comments to the Author(s).
I think the authors have done a superb job addressing the issues raised during the review process. I was disappointed to see so many comments raised that I felt were very clearly addressed in the supp mats and original submission, so it is great that the authors were given the opportunity to clarify those points in a revision. I think this makes an important and timely contribution to the ongoing debate about whether or not humans display reliable cues of health.

Referee: 1

Comments to the Author(s).
I have read thoroughly through the manuscript and can see that the authors have made a number of improvements. While the videos are of non-experimentally controlled origin, and McTurk has a number of weaknesses, the authors have done a thorough job in maximizing the designs and also reasonably deal and argue of their validity. I also agree that the outcomes are scientifically sound, i.e. that humans do not seem very good at judging whether sneezes coughs are from a

contagious origin stimulated by something else, and that rated disgust and certainty are not related to whether the sounds were from a contagious individual or not. There are, however, some aspects of the ms that are not properly addressed.

1) The title makes the reader believe they have tested whether people can identify infectious disease using auditory cues. This is a very bold statement considering that the authors have only included specific short sounds from sick people. For the statements made about the ability to detect sick sounds, they also have to show proof of concept that they actually used the most relevant sickness sounds. Since this is not done in the ms, only including particular sounds presented for a few seconds, the title should be updated to fit the question studied. For example. 'Sounds of sickness: Can people identify infectious disease using auditory clips of sneezes and coughs?'

2) It is up to the authors to disagree with my previous statement ".... However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc…" i.e. that there has NOT been a large evolutionary pressure for developing different coughs and sneezes depending on the causation. It would be suitable for the authors to include/discuss this possibility in the limitations, particularly since it is a likely reason for the 'null findings'.

3) Going through the OSF-online material (also appreciating the effort to find sick and not sick related sounds), I could not open the links in the "url1_archive", and only the original flies.

4) I am surprised that individuals did not differ in their rating behavior or accuracy (see method, rows 156-159), This is very different from the analyses made by Kurvers & Wolf, 2018, and large parts of the literature where individuals often show large differences in rating behavior. It seems appropriate to include 'rater' as a factor in all models since there are 4 separate studies and it may differ from study to study, and also report the ICC data from both 'rater' and 'sound originator'.

Author's Response to Decision Letter for (RSPB-2019-2674.R0)

See Appendix C.

RSPB-2020-0944.R0

Review form: Reviewer 1 (John Axelsson)

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors have done a good job with updating the ms. I have a few minor comments

1) General discussion, 3rd paragraph: Aouthors state "Recent work indicates that this sound variation is available (e.g., to statistical learning algorithms; Porter et al., 2019), but perhaps human hearing mechanisms cannot use it reliably, even with clinical training (e.g., Smith et al., 2006). ," An alternative is likely, and I suggest to directly mention that the sound variation in coughs and sneezes are rather limited.

2) References: Please check that references are correct. I know the study by Olsson et al, did not have A. Soop as last author (it is M Lekander).

3) The authors have uploaded the clips for others to use also on the webarchive, but I still have problems opening most of the youtube clips provided on this page
https://nickmichalak.shinyapps.io/sounds_of_sickness_sensitivity/
Please check the links are working.

Review form: Reviewer 3 (Leonid Tiokhin)

Recommendation
Accept as is

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Acceptable

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
NA. I appreciate your effort to seriously engage with the reviewer comments and to make all materials, data, and code openly available so that the scientific community can easily judge the quality of your paper.

Decision letter (RSPB-2020-0944.R0)

12-May-2020

Dear Mr Michalak

I am pleased to inform you that your manuscript RSPB-2020-0944 entitled "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" has been accepted for publication in Proceedings B.

The referee(s) have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the referee(s)' comments and revise your manuscript. Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days. If you do not think you will be able to meet this date please let us know.

To revise your manuscript, log into https://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referee(s) and upload a file "Response to Referees". You can use this to document any changes you make to the original manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Before uploading your revised files please make sure that you have:

1) A text file of the manuscript (doc, txt, rtf or tex), including the references, tables (including captions) and figure captions. Please remove any tracked changes from the text before submission. PDF files are not an accepted format for the "Main Document".

2) A separate electronic file of each figure (tiff, EPS or print-quality PDF preferred). The format should be produced directly from original creation package, or original software format. PowerPoint files are not accepted.

3) Electronic supplementary material: this should be contained in a separate file and where possible, all ESM should be combined into a single file. All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

4) A media summary: a short non-technical summary (up to 100 words) of the key findings/importance of your manuscript.

5) Data accessibility section and data citation
It is a condition of publication that data supporting your paper are made available either in the electronic supplementary material or through an appropriate repository.

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should be fully cited. To ensure archived data are available to readers, authors should include a 'data accessibility' section immediately after the acknowledgements section. This should list the database and accession number for all data from the article that has been made publicly available, for instance:

- DNA sequences: Genbank accessions F234391-F234402
- Phylogenetic data: TreeBASE accession number S9123
- Final DNA sequence assembly uploaded as online supplemental material
- Climate data and MaxEnt input files: Dryad doi:10.5521/dryad.12311

NB. From April 1 2013, peer reviewed articles based on research funded wholly or partly by RCUK must include, if applicable, a statement on how the underlying research materials – such as data, samples or models – can be accessed. This statement should be included in the data accessibility section.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link http://datadryad.org/submit?journalID=RSPB&manu=(Document not available) which will take you to your unique entry in the Dryad repository. If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link. Please see https://royalsociety.org/journals/ethics-policies/data-sharing-mining/ for more details.

6) For more information on our Licence to Publish, Open Access, Cover images and Media summaries, please visit https://royalsociety.org/journals/authors/author-guidelines/.

Once again, thank you for submitting your manuscript to Proceedings B and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Sincerely,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member
Comments to Author:
We have now heard from our reviewers about your revised manuscript. As you see, they are both positive. Reviewer 1 has some minor comments that you should be able to deal with easily before we move forward. I am recommending acceptance. Congratulations on an interesting paper.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s).
The authors have done a good job with updating the ms. I have a few minor comments

1) General discussion, 3rd paragraph: Aouthors state ”Recent work indicates that this sound variation is available (e.g., to statistical learning algorithms; Porter et al., 2019), but perhaps human hearing mechanisms cannot use it reliably, even with clinical training (e.g., Smith et al., 2006). ,” An alternative is likely, and I suggest to directly mention that the sound variation in coughs and sneezes are rather limited.

2) References: Please check that references are correct. I know the study by Olsson et al, did not have A. Soop as last author (it is M Lekander).

3) The authors have uploaded the clips for others to use also on the webarchive, but I still have problems opening most of the youtube clips provided on this page https://nickmichalak.shinyapps.io/sounds_of_sickness_sensitivity/
Please check the links are working.

Referee: 3

Comments to the Author(s).
NA. I appreciate your effort to seriously engage with the reviewer comments and to make all materials, data, and code openly available so that the scientific community can easily judge the quality of your paper.

Author's Response to Decision Letter for (RSPB-2020-0944.R0)

See Appendix D.

Decision letter (RSPB-2020-0944.R1)

14-May-2020

Dear Mr Michalak

I am pleased to inform you that your manuscript entitled "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Your article has been estimated as being 7 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Paper charges
An e-mail request for payment of any related charges will be sent out shortly. The preferred payment method is by credit card; however, other payment options are available.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

You are allowed to post any version of your manuscript on a personal website, repository or preprint server. However, the work remains under media embargo and you should not discuss it with the press until the date of publication. Please visit https://royalsociety.org/journals/ethics-policies/media-embargo for more information.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Appendix A

Review of:

Sounds of sickness: Can people identify infectious disease using auditory cues?

Manuscript ID: RSPB-2019_1719

TO THE EDITOR/authors:

The authors conducted several MTURK studies to explore whether participants could identify whether coughs and sneezes came from infected vs non-infected individuals. The studies provided no evidence for this effect. They provided strong evidence that pariticpants' subjective assessment of how disgusting these sounds were correlated with their judgements of whether the sound was infectious or not.

I have both major and minor concerns with the submission in its current form (see below). In addition to a typical subjective assessment of the manuscript, I have evaluated the manuscript for some common limitations, in part based on:

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them?. *Journal of the Royal Society of Medicine, 101*(10), 507-514.

I hope that my comments will help the authors to strengthen their manuscript.

For accountability and transparency, I would like to sign this review.

Leo Tiokhin

COMMON LIMITATIONS

Is the study properly justified (i.e., does it appropriately describe previous research and its marginal contribution on top of prior work)?

Yes, but see OTHER CONCERNS section below for comments on intro.

Are the participants properly randomized between test groups? Any issues with response rates?

All studies - not clear to me how participants were randomized. Should be more explicit.

Issues with experimental design?

Pg 139 – why were infectious diseases defined but non-infectious ones were not? This means that you have primed participants to think of infectious diseases as x, but allow participants to think of non-infectious diseases in whatever folk-model way that they want. Idk if this is a huge problem, but it certainly introduces noise.

Study 1 – not clear to me whether participants were blinded to purpose of experiment.

What measures did you take to ensure that the stimuli in the 2 groups (infectious vs non-infectious) didn't differ along other parameters that were irrelevant to your study? For instance, what if the infectious sounds came from people who were more severely afflicted than the non-infectious sounds? If you didn't take any such measures, then there are potentially confounds galore in this study. For example, on line 180-182, you note that people more often thought that non-infectious sounds were infectious than infectious sounds were infectious. That’s a bit weird, if it’s a reliable finding. And one possibility is that the non-infectious sounds you used were just different in other ways than infectious sounds, which confounds the design. Anyways, I may be wrong, but as is, the paper doesn’t provide me with any information that indicates that you made sure to control for other potential confounds.

Sample size – is the sample size appropriately justified? Is there a power analysis and is it adequate?

Not really. How did you do the power calculation? What test did you use? Why did you chose an effect size of cohen’s h of 0.23 or 0.19 or whatever it is for each individual study? Was this apriori or posthoc? That is, why would you expect there to be an xx% difference from 50%? Does prior work on sound and disgust give you any indication of how small the expected effects are?

Transparency – is the study pre-registered? Are the data, materials, and code openly available?

The study is laudably transparent, providing open data, materials, and code. The submission and OSF description also implies that the study was pre-registered. However, I was not able to find the original pre-registration, or the csv / .xlsx files for it. The authors should provide these (or make them easier to find) if a resubmission is invited.

Researcher degrees of freedom – are there signs that researcher degrees of freedom exist that may affect minor or major conclusions in the manuscript?

All studies – not clear to me whether exclusion criteria were decided on apriori, or posthoc. Do the results qualitatively change if you don’t exclude participants? How robust are the results to a multiverse analysis where you conduct all reasonable specifications of statistical models?

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

I’m not saying that you need to do a huge extra amount of analyses, but I do think it is important to conduct more sensitivity checks.

Measurement (see Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. https://doi.org/10.31234/osf.io/hs7wm)

Have the measures been shown to be valid (i.e., measure the construct of interest) and reliable (e.g., test-retest reliability; questionnaire items are highly correlated with one another). Have all modifications to the scales been appropriately described? How have any measurement summary stats been calculated? Why has this measure been used as opposed to another?

Line 151 – 153. Please report these in the main manuscript. I think you should provide readers with more information, such as the reliability of these scales, why you chose them, how they have been used in past work, whether you made any modifications, etc.

Issues with statistical analyses

I think that it's useful to describe the exact statistical model that you used for each analysis, in the main text. You should also report any sensitivity checks you did and whether the results are robust to these.

Line 159 – 161. I have a hard time understanding the statistical model you used from this. When you say that the participant factor didn't account for meaningful variance, are you saying that you *didn't* end up using a random intercept for participants? Or something else? I'm not an expert on multi-level models, but I don't know of when it is ever justified to not include random intercepts for participant when there are multiple observations per participant. The statistical models used here should definitely be evaluated by someone with more expertise than me.

Line 176 – 178. The fact that the confidence interval upper band translates to 53% accuracy does NOT mean that you can reject values above 53%. Confidence intervals are a frequentist long-run concept. In the long run, 95% of confidence intervals will contain the true mean. If the true mean is 50%, one of your confidence intervals in one study could be 30-51% and another one could be 49-70%. So, if you were to use a single confidence interval to make an inference about the "true" level of accuracy, you will be very mislead based on a single study. See for example https://daniellakens.blogspot.com/2016/03/the-difference-between-confidence.html

Please also address the above point throughout the rest of the manuscript, as the same mistake is made in the inference about the other studies. For instance, there are many cases where you find that some thing "did not depend on" some other thing, but the confidence intervals are huge, so it's not clear to me how informative this is, without the appropriate statistical test.

You argue that people didn't accurately distinguish infectious vs non-infectious sounds but provide no statistical test that is able to assess evidence for the null hypothesis. For example, the non-significant p values just indicate given the null hypothesis the data is not unlikely. And the confidence intervals are consistent with a range of effect sizes. To quantify evidence for a null effect, you could consider using Bayes factors or equivalence tests:

Harms, C., & Lakens, D. (2018). Making'null effects' informative: statistical techniques and inferential frameworks. *Journal of clinical and translational research*, *3*(Suppl 2), 382.

Are the conclusions justified, given the nature of the study and data?

In every study, participants are worse than chance at correctly identifying both types of sounds. Can you discuss why you think this is? It's a bit of a weird finding, no?

I think that the dscription of the conclusions could be more accurate. You used a subset of sounds (from youtube) without controlling or testing for potential differences between them. The participants were mturk. You used coughs and sneazes. And you found no evidence, although you didn't conduct analyses that would indicate evidence for the null effect (I bet if you do these, you will find strong evidence for the hypothesis that people are worse than chance, not for null effect, but also certainly not for the fact that people can accurately identify things). In my opinion, the conclusions would be more justified if they discussed these limitations of the study.

OTHER CONCERNS

Pg 47 – citation for "existing work suggests"?

Pg 51 – “surprisingly” – not clear to me why this is susprising. Do we have a strong prior about this in the opposite direction?’

Pg 62 – 64 – also consider citing Curtis V. A. 2014. Infection-avoidance behaviour in humans and other animals. Trends in Immunology
35:457–464.

You refer to “people” throughout the intro, but this makes it unclear where the sample populations come from. For example, Curtis 2004 is an online cross-cultural sample while some of the other studies in the intro are undergrads. Worth clarifying, for sake of transparency.

A general note for the intro – do you know how well replicated these studies are? For example, the LPS and health rating study? If there are concerns about the replicability of any of this work, it’s worth noting, so that we don’t get an overconfident picture of current knowledge.

Line 83 – belching – how is this at all relevant to adaptations for pathogen detection? Also, the p-value for the effect of auditory cues on that study provides weak evidence against the null (see below quote). I’m not saying that this means the effect isn’t “real”, but again, it really would be a more comprehensive intro if you provided more information about the evidentiary value of previous work, instead of just putting it in the typical narrative frame of “past work has shown x y z”. I am guilty of doing this myself, but I think that we as a field can do better.

- “The olfactory disgust stimuli showed the highest significance difference compared to the control ($p < 0.001$) while visual disgust and auditory disgust showed smaller responses, but they still had significantly higher probability to increase hand washing attempts than the control (for all cases $p = 0.025$).

Line 91-93 – argument that mistaking infectious person as non-infectious is less costly than the latter error – what evidence do we have for this? For instance, the expected costs of this mistake will depend on the base rate with which you encounted sick vs not sick individuals. If you’re constantly encountering healthy individuals but not interacting with them, this “small” cost will be large in aggregate. If you’re going to use error management stuff, it’s worth taking a look (and citing) this, for a more nuanced theoretical perspective:

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and human behavior*, *31*(5), 309-319.

Line 107 – Research Overview – I think that, if you want to mention the pilot study, you need to provide more information about what was done exactly, whether it was pre-registered, what motivated the study, etc. I appreciate the transparency of noting that you ran this study, but I would appreciate more information about how you changed your design as a result of it, what you learned, why you did the study in the first place, etc.

Line 130 – do you mean “different” types of sounds?

Footnote #1 – I don’t understand this. Can you please elaborate on exactly what you did? And also report all analyses in which you used both a subset of stimuli and all the stimuli?

Line 136 – who were these coughs from? That is, were they generated from individuals all over the world, from the U.S., etc? It may be more difficult for people to detect infectious disease state for individuals from different cultural backgrounds or something. I don't know, but you should provide this information. Also, have you made this stimulus set openly available on the OSF?

Line 190 – is a rating of 6.7 out of 9 really "highly certain"? Also address this in the rest of the manuscript (e.g., line 256)

Line 197 and 199– report the exact p value. Same for everywhere else in the manuscript (e.g., lines 302 – 306).

Figures – I had a hard time telling which was "figure 1". If it is the figure on page 23, why doesn't it show study 1?

Pg 218 – please elaborate. I don't quite follow this base rate idea. Is it that you're saying people usually encounter healthy people, and so they may have a prior that sounds indicate non-infection? If so, then you should find lots of false negatives, right? As opposed to the false-positive narrative that was presented in the intro?

Pg 321 – 323 – What theories are these exactly? To me, this study came across as completely exploratory "Hey, people haven't looked at sound yet, why don't we do that?". And that's totally fine for me – we need more honest high-quality exploratory research. But then here in the discussion it's written up as if it was a confirmatory thing testing a theory. Really doesn't seem that way to me. I don't see how the sound thing is a deductive prediction from a theory. I think it should be written up as exploratory.

Appendix B

2019-11-15

Dear Dr. Brosnan,

We previously submitted a manuscript titled "Sounds of sickness: Can people identify infectious disease using auditory cues?" (RSPB-2019-1719) for review at Proceedings of the Royal Society B: Biological Sciences. Our manuscript initially received a Reject decision though you decided to accept a revision following an appeal. We appreciate your handling of our manuscript along with the AE and for noting the positive comments made by referees. In our revision comments, we refer to many of the same points raised in our appeal.

While evaluating the referees' comments in detail, we noted that the majority of their concerns were already addressed in either our manuscript or in our supplementary materials (we used our supplement when space was limited). This is the case for almost all points raised by Referee 1 and most major points raised by Referee 3 (Referee 2's few suggestions are easy to accommodate). To summarize, the major potential concerns discussed by referees involved (1) our selection of stimuli and (2) specific features of our statistical analyses. In the originally submitted supplement, we reported extensive robustness analyses showing that our key finding—no evidence that people could accurately (above 50%) distinguish infectious coughs and sneezes from non-infectious ones—did not hinge upon any particular set of stimuli. Our key finding remained robust when analyses were limited to sounds recorded from the most convincingly infected individuals (i.e., those reporting diagnoses from medical professionals). We discussed these analyses in the main text and more extensively in our supplement. Further, the supplement included detailed descriptions of the stimuli along with the actual sound files to assist with referee evaluation. Together, the stimulus material and analyses originally included in our manuscript and in our supplement address the majority of the referee feedback and should obviate most concerns about potential confounds.

Below, we have responded to each comment and noted when we also revised our manuscript. In many cases, our responses to referee concerns reference our preregistrations, study materials, deidentified data, analysis code, and supplementary write-ups in our external OSF repository. Referees were able to access this information via an anonymized link to our OSF Project (see below), which was provided with the original submission materials. Such information was limited to the supplement because of space limitations inherent to Proceedings B manuscripts. Because reviewers mentioned trouble accessing our preregistrations specifically, we have updated the link. Supplementary materials can be found at:

- Submitted with original manuscript: https://osf.io/4c7vr/?view_only=de1a2eb674374f3bba5498f207c47882
- Updated preregistrations and R code: https://osf.io/4c7vr/?view_only=dd859850b1314242a7962411d7fe0da7

In addition, we have updated our figures to improve ease of interpretation and to provide additional relevant information. Specifically, Figure 1 now depicts accuracy scores as both means and average scores for each stimulus (depicted with large points). Also, Figures 1 and 2 now label the sound origins of stimuli as "Non-Infectious Person" and "Infectious Person."

Thank you and the referees for your time and effort with this manuscript. If there are any concerns we missed, please let us know and we would be happy to respond to them.

Regards,
Nicholas Michalak, Oliver Sng, Iris Wang, Joshua Ackerman

Note. In the following text, we quote in *italics* points raised by each referee and provide our response to those points in regular font.

Referee: 1

Major concerns.
1) The study focus on how well one can discern coughs/sneezes from infectious vs non-infectious individuals. While this is interesting, a more appropriate questions to ask, at this stage in what we know in the field, are "Do humans use sounds to decide whether someone is sick or not?" and "What sounds do we use to tell if someone is sick?". The usage of coughs and sneezes are of course probable sounds that help us judge if someone is sick or not. However, it does not seem likely that evolution has strongly advocated a selection for detecting which coughs that have a pathogenic origins or an alternative cause. Here, evolution has probably favored those thinking that coughs and sneezes may be signals that we should be careful to interact with this individual. While the article is interesting, it is more focused on the detail to whether humans have different kind of coughs and sneezes depending on whether they are driven by pathogens or of other origin. However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc. Additionally, the authors do tap into why some coughs and sneezes are judged differently, and the mechanisms for this would be an interesting venue as well to explore.

We thank the referee for these comments, as we think they cut to the heart of an alternative, reasonable prediction that could be made about infection identification through sound. The referee suggests selection could have favored mechanisms that motivated humans to be cautious around others who cough and sneeze more generally and would not have selected for the ability to distinguish the infectious origins of such sounds. In contrast, we proposed that selection could have favored human sensitivity to the particular coughs and sneezes that diagnose infection. These are (non-mutually exclusive) hypotheses that require empirical tests to disentangle—people could be extra cautious toward others who cough and sneeze and as well as discriminating toward particular coughs and sneezes. We believe there are reasons to think that, in principle, humans could have evolved abilities to identify the infectious/non-infectious origin of sounds like coughs and sneezes. For instance, we know from everyday experience that people do not always react to others' coughs and sneezes as though these sounds indicate infection. Additionally, medical doctors can be trained to use auditory information to more accurately diagnose sickness—a similar process is true even with non-human animals (e.g., Ferrari, Silva, Guarino, Aerts, & Berckmans, 2008)—and so at least some relevant informational content must be present in the sounds (which selection could have targeted over the course of evolution).

Ferrari, S., Silva, M., Guarino, M., Aerts, J. M., & Berckmans, D. (2008). Cough sound analysis to identify respiratory infection in pigs. *Computers and Electronics in Agriculture*, *64*(2), 318-325.

2) It is hard to judge the quality of the audio-clips. Where the infectious subjects really infectious at the time of collection and from what pathogen? And where the control clips really from people who were not carriers of any pathogens? A strength with previous studies is the usage of stimuli material where sickness had been experimentally controlled, and where subjects were properly healthy in the control condition.

We fully agree with the referee that it is important to ensure that targets were appropriate to their labeled conditions. For infected targets, we relied on videos of people who self-reported sickness (as described and linked in our supplemental materials). In fact, many targets reported receiving official medical diagnoses on camera. In contrast, we are certain that targets in the non-infection condition produced sounds due to pathogen-irrelevant reasons. All such targets actively engaged in activities to produce coughs and sneezes in the videos (e.g., plucking a nose hair, inhaling a powder). We do agree with the referee that it can be difficult to judge these factors from the audio clips alone (this is the same difficulty participants had), and so we encourage all referees to view the videos or read the summaries provided in the supplement. These materials should allow for better evaluation of the sounds. We also conducted additional analyses limiting the infected target group to only those individuals explicitly receiving medical diagnoses (originally reported in Footnote 1 within the manuscript and in our supplement, but now reported in the analysis outline section, Lines 170-174). In other words, we conducted a focused test on the subset of stimuli for which we could most confidently say were from infected sources. There was no change to the key findings.

3) The use of MTurk is common and tempting approach for rating projects. A major problem is that some are not so interested in doing their best rather than earning money as fast as possible, for example the existence of "Super Turkers" shows this point. Several approaches has been developed to combat such behavior, such as inclusion of control questions or removing subjects who do not vary much in their responses or respond very fast. The authors have not shown any approach dealing with this problem. Hence, the material is very likely to include a lot of error variance, and hence has much lower power than stated.

In line with the referee's concern, we are sensitive to the issue of participant error variance. We offer several responses to this issue. First, we recruited only participants who met Amazon MTurk's specification of worker approval rate of 95% or greater, helping to ensure that participants took the study seriously. This is now specified more clearly in the manuscript, alongside a citation to the *Behavioral Research Methods* paper (Litman, Robinson, & Abberbock, 2017) that investigates MTurk data more thoroughly (Line 125). Second, as stated in the manuscript, we excluded people for reasons that would affect their responding (e.g., technical problems, training in infection identification). Third, we would push back a bit on the idea that data from MTurk participants is inherently problematic. Direct comparison of MTurk samples with non-MTurk samples often suggests they do not differ with respect to many outcomes, and online participants are not more prone to "satisfice" than are other types of participants (Casler, Bickel, & Hackett, 2013; Snowberg & Yariv, 2018). Further, use of control or check questions has been found to generate unique problems (e.g., increasing deliberative thinking), and may itself actually create confounds in experiments (Hauser & Schwarz, 2015, 2016). Finally, it is not clear why error variance due to use of MTurk data would explain why we found many significant effects apart from those for accuracy: Disgust significantly predicted rating sounds as infectious, and our widely-used individual difference scales replicated psychometric properties (e.g., reliability) commonly reported in the literature. So, though Referee 1 is correct that satisficing in survey responses introduces noise and reduces power to detect hypothesized effects, these issues do not seem to apply substantially to our data.

4) The authors measure accuracy, a combination of correctly identifying both infectious and non-infectious responses in relation to failures. Since they also discuss costs of their decisions (eg. Row 90 onwards), they should also include information, at least some, on specificity and sensitivity.

We thank the referee for this suggestion, and have included the false positive rate, specificity (1 – false positive rate), and true positive rate/sensitivity for each study (Lines 183-184; Lines 238-240, and Lines 287-289).

5) Fig 2 needs better explanations. The information is confusing. A reader does not understand what ”sound rating” or ”sound origin” means. It is much better to present it in terms of ”coughs/sneezes” and ”infectious individuals”. Now, ’sound origin’ is explained as ”different colors and types”, which is unclear. Also clarify what the results are so people understand the purpose of the figure.

We understand the confusion in interpreting this figure. To help with this, we have relabeled the sound origin factor levels “Infectious Person” and “Non-Infectious Person” (before, they were labeled “Infectious” and “Non-Infectious”). These represent people who coughed or sneezed in each clip. We also titled the x-axes with the unique ratings made by participants in either condition (i.e., clarity or disgust). We also briefly describe the results in the figure caption (Lines 300-307).

Minor comments
- The authors discuss the costs of detection and bias in detection, and trait differences in detection accuracy. There are a few recent analyses and commentaries on this in Proceedings B last year. See Kurves and Wolf 2018 and response by Axelsson et al 2018. These data are in opposite of the findings here (i.e. Kurves find very strong individual differences in strategy) and should be discussed (at least in part by the limitations of using MTurk).

We agree that discussion of this recent work is important for our investigation (we included points from the Axelsson paper in our original manuscript but also now refer to the Kurvers & Wolf paper as well, Lines 69-70).

- Row 158. Explain what ”sound stimuli” is.

This refers to the sounds clips used in the studies. In our original manuscript, we described sound stimuli in our materials and procedure section (Lines 132-140): “Sound stimuli featuring coughs and sneezes were extracted from online, U.S.-based videos (e.g., YouTube) (see materials supplement). We included different types of sounds to improve ecological validity, though we had no a priori predictions about sound type differences. Individuals who generated the infectious sounds self-reported with certainty experiencing sickness with an infectious disease (e.g., cold, flu). Individuals who generated the non-infectious sounds responded to an environmental irritant (e.g., allergies, consumption of powdery spices). We trimmed videos to 1-2 second audio clips featuring only the target sound. The full stimulus set comprised 20 coughs and 20 sneezes, with half of each sound type being infectious or non-infectious in origin.”

- Row 176. Unclear what the reference ”2” refers to.

The second footnote (now footnote 1) refers to the method used to compute the confidence interval. We use this method throughout to compute confidence intervals.

- Row 179. There is a reference for Figure 1 when reporting study 1. However, there is no data plotted for study 1, only from all the other data collections. This is confusing.

Thank you to the referee for pointing this out. To fix this issue, we have replaced “Study 2,” “Study 3,” and “Study 4” in Figure 1 with “Study 1,”, “Study 2,” and “Study 3” (in our original figure, the study numbering was off because the pilot study was incorrectly labeled as Study 1)**.**

Referee: 2

Comments to the Author(s)
I have very rarely recommended a paper for publication without any edits, but was very close to with this one. It's a comprehensive set of studies, transparently reported, with extremely clear codebooks and code. Excellent. My only substantial recommendation is that the authors bring a little more clarity to their description of the stimuli. I also thought it might be useful for the authors to discuss recent studies in the visual domain suggesting that previously reported links between susecptibility to infectious illnesses and facial cues are not ribust (Ziyi Cai et al. and Yong Foo et al's recent papers, for example).

We thank the referee for their very positive feedback. In following the suggested changes, we now include these references, which help to make the case that much more research is necessary to support firm conclusions in this literature (Line 71). Regarding stimuli descriptions, we have now clarified the origins of the sound stimuli and reference a spreadsheet in our supplement with details on each sound clip.

Referee: 3

Are the participants properly randomized between test groups? Any issues with response rates?
All studies - not clear to me how participants were randomized. Should be more explicit.
Issues with experimental design?

We used Qualtrics's randomization procedure to randomly present stimuli within all studies and to randomly assign conditions in Study 3. Our Qualtrics .qsf files for all study surveys are available in our supplemental repository, so our randomization procedure can easily be reproduced.

Pg 139 – why were infectious diseases defined but non-infectious ones were not? This means that you have primed participants to think of infectious diseases as x, but allow participants to think of non-infectious diseases in whatever folk-model way that they want. Idk if this is a huge problem, but it certainly introduces noise.

By defining infectious illnesses, non-infectious sources include anything outside that definition – this difference is precisely what we wanted to evaluate. In other words, whatever conceptions participants had about non-infectious sounds are not noise, but conceptions that fall within our intended definition of non-infectious sounds.

Study 1 – not clear to me whether participants were blinded to purpose of experiment.

The purpose of the experiment was to test whether people can distinguish between non-infectious and infectious sounds. This was made explicit to the participants (else they could not have completed the study appropriately), and so we were not able to blind participants.

What measures did you take to ensure that the stimuli in the 2 groups (infectious vs noninfectious) didn't differ along other parameters that were irrelevant to your study? For instance, what if the infectious sounds came from people who were more severely afflicted than the non-infectious sounds? If you didn't take any such measures, then there are potentially confounds galore in this study. For example, on line 180-182, you note that people more often thought that non-infectious sounds were infectious than infectious sounds were infectious. That's a bit weird, if it's a reliable finding. And one possibility is that the noninfectious sounds you used were just different in other ways than infectious sounds, which confounds the design. Anyways, I may be wrong, but as is, the paper doesn't provide me with any information that indicates that you made sure to control for other potential confounds.

The stimuli came from two sources – videos of people actively manipulating their nasal & oral cavities and videos of people reporting infection – and involved a range of predisposing factors in both conditions. As stated in the manuscript/supplement, to create equivalent audio clips, we truncated files to include 1-2 seconds of sound (ensuring the same average frequency & duration of sounds across conditions) and normalized sound files to ensure equal volume between clips. To examine differences in sounds, we assessed (and statistically adjusted for) perceived clarity and disgustingness of each clip in the studies. These aspects were reported in the original manuscript and all stimuli descriptions and sources are presented in the supplement. Beyond these parameters, other factors were not examined. However, many of these are inherent to the difference between conditions. For instance, an infected person is necessarily different in “affliction” compared to a non-infected person. Given that our focus was on infection vs. non-infection identification, we would argue that, absent referee predictions about specific sound features, differences between sounds are not in fact confounds but instead auditory features that could be examined to help characterize how infectious and non-infectious sounds are alike/unalike.

Sample size – is the sample size appropriately justified? Is there a power analysis and is it adequate?
Not really. How did you do the power calculation? What test did you use? Why did you chose an effect size of cohen’s h of 0.23 or 0.19 or whatever it is for each individual study? Was this apriori or posthoc? That is, why would you expect there to be an xx% difference from 50%? Does prior work on sound and disgust give you any indication of how small the expected effects are?

Power analysis calculations (including our assumptions) and details about specific tests were available in our supplement. Regarding effect sizes, we computed the effect size that we were able to detect with the sample size collected (i.e., sensitivity tests). As with any investigation that cannot rely on existing data to make arguments about appropriate effect sizes for power calculations (because prior research on this specific topic has not been conducted), we must make some decisions about what degree of effect size we would consider “meaningful” to examine. We were 80% powered to detect whether accuracy was ~ 10-11% different than 50% accuracy. It may in fact be that perceivers are able to detect differences between sound origins at lower levels (e.g., improvements in accuracy from 1-9%), but we were not able to detect with 80% power such small effects with any individual study sample we collected.

Transparency – is the study pre-registered? Are the data, materials, and code openly available?
The study is laudably transparent, providing open data, materials, and code. The submission and OSF description also implies that the study was pre-registered. However, I was not able to find the original pre-registration, or the csv / .xlsx files for it. The authors should provide these (or make them easier to find) if a resubmission is invited.

In the original manuscript, we provided links to our OSF project page, which included supplemental material and pre-registration materials (link generated on 11-24-2018). However, it appears that our pre-registrations were not accessible via the read-only/anonymous OSF link (although all other supplemental information was). We have corrected this to make sure all material is available in the case of further review (updated link generated on 11-15-2019). Note that we have not updated any materials. We have only updated what can be viewed.

Researcher degrees of freedom – are there signs that researcher degrees of freedom exist that may affect minor or major conclusions in the manuscript?
All studies – not clear to me whether exclusion criteria were decided on apriori, or posthoc.

We preregistered exclusion criteria for Studies 2 and 3, and we used those same criteria in Study 1.

Do the results qualitatively change if you don't exclude participants? How robust are the results to a multiverse analysis where you conduct all reasonable specifications of statistical models?
Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712.
I'm not saying that you need to do a huge extra amount of analyses, but I do think it is important to conduct more sensitivity checks.

We reported many different specifications in our supplemental materials. The results do not substantively change if participants are not excluded. And the results are robust to all reasonable specifications of the model (presented in supplement).

Measurement (see Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. https://doi.org/10.31234/osf.io/hs7wm)
Have the measures been shown to be valid (i.e., measure the construct of interest) and reliable (e.g., test-retest reliability; questionnaire items are highly correlated with one another). Have all modifications to the scales been appropriately described? How have any measurement summary stats been calculated? Why has this measure been used as opposed to another?
Line 151 – 153. Please report these in the main manuscript. I think you should provide readers with more information, such as the reliability of these scales, why you chose them, how they have been used in past work, whether you made any modifications, etc.

We reported extensive psychometric properties of all our measurements in our supplement, and we cited development/validation papers for multi-item scales. Unfortunately, because of space limitations used at Proceedings B, we are not able to fit all of the requested information in the manuscript itself.

Issues with statistical analyses
I think that it's useful to describe the exact statistical model that you used for each analysis, in the main text. You should also report any sensitivity checks you did and whether the results are robust to these.

In the "Analysis outline" section (Lines 152-174), we detailed the models used across studies. We combined this information into one section in order to save space in the manuscript. Additionally, as mentioned in an earlier response, we conducted many different specifications in our supplemental materials (e.g., different random effects specifications, fixed effects for coughs vs. sneezes).

Line 159 – 161. I have a hard time understanding the statistical model you used from this. When you say that the participant factor didn't account for meaningful variance, are you saying that you didn't end up using a random intercept for participants? Or something else? I'm not an expert on multi-level models, but I don't know of when it is ever justified to not *include random intercepts for participant when there are multiple observations per participant. The statistical models used here should definitely be evaluated by someone with more expertise than me.*

In the manuscript (Lines 156-158), we specified that: “However, the participant factor never accounted for meaningful variance (i.e., participants judgments had high consensus), so we excluded it in all reported models.” The model estimated 0 variance for the subject factor (i.e., random subject factor which is the same thing as random intercepts). There are technical reasons why the model estimated exactly 0 (see p. 10 Bates, 2010), but, importantly, it made no substantive difference whether or not we included this factor. Therefore, it was excluded. However, we reported models with and without this factor in our supplement, thereby addressing the referee’s concern.

Line 176 – 178. The fact that the confidence interval upper band translates to 53% accuracy does NOT mean that you can reject values above 53%. Confidence intervals are a frequentist long-run concept. In the long run, 95% of confidence intervals will contain the true mean. If the true mean is 50%, one of your confidence intervals in one study could be 30-51% and another one could be 49-70%. So, if you were to use a single confidence interval to make an inference about the "true" level of accuracy, you will be very mislead based on a single study. See for example https://daniellakens.blogspot.com/2016/03/the-difference-betweenconfidence.html

We agree with the referee’s description of confidence intervals, but we disagree that our interpretation was incorrect. If one follows this procedure to infinity—sampling proportions, constructing 95% confidence intervals around them, rejecting all proportions outside of them—then one will incorrectly reject the population proportion 5% of the time. This is an uncontroversial use of confidence intervals, which include all null values one cannot reject in a significance test. Our claim that we can reject values above 53% in this example includes this built-in, procedural error. We would have made a similar claim had 50% (our focal null hypothesis) fallen below our 95% confidence interval (i.e., observed accuracy is significantly greater than 50% *and* every value below our lower bound).

Please also address the above point throughout the rest of the manuscript, as the same mistake is made in the inference about the other studies. For instance, there are many cases where you find that some thing "did not depend on" some other thing, but the confidence intervals are huge, so it's not clear to me how informative this is, without the appropriate statistical test. You argue that people didn't accurately distinguish infectious vs non-infectious sounds but provide no statistical test that is able to assess evidence for the null hypothesis. For example, the non-significant p values just indicate given the null hypothesis the data is not unlikely. And the confidence intervals are consistent with a range of effect sizes. To quantify evidence for a null effect, you could consider using Bayes factors or equivalence tests:
Harms, C., & Lakens, D. (2018). Making'null effects' informative: statistical techniques and inferential frameworks. Journal of clinical and translational research, 3(Suppl 2), 382.

We are sensitive to this issue of interpreting findings precisely, so we took pains to carefully report what we found. Contrary to the referee’s statement indicating that we argued that “people did not accurately distinguish infectious vs. non-infectious sounds," instead our language stated that "no sufficient evidence" was found for differences in identification, which follows logically from the null hypothesis tests we did conduct. As for quantifying evidence for the null hypothesis, we did so using two approaches: confidence or Bayesian credible intervals and Savage-Dickey ratios (conceptually similar to Bayes Factors). For each study in our main text, we reported the accuracy values greater than 50% that fell outside of confidence intervals (following similar logic to equivalence tests). In our supplement, we reported meta-analytic 90% confidence intervals (i.e., equivalence tests) and we reported Savage-Dickey ratios (like Bayes factors) that quantified a ratio of how likely 50% accuracy was under the null distribution (with mean equal 50%) vs. an alternative distribution (with mean less than 50%). We provide more details in our supplement, but, importantly, we did report appropriate tests for assessing evidence for the null hypothesis in the original documents.

Are the conclusions justified, given the nature of the study and data?
In every study, participants are worse than chance at correctly identifying both types of sounds. Can you discuss why you think this is? It's a bit of a weird finding, no?
I think that the dscription of the conclusions could be more accurate. You used a subset of sounds (from youtube) without controlling or testing for potential differences between them. The participants were mturk. You used coughs and sneazes. And you found no evidence, although you didn't conduct analyses that would indicate evidence for the null effect (I bet if you do these, you will find strong evidence for the hypothesis that people are worse than chance, not for null effect, but also certainly not for the fact that people can accurately identify things). In my opinion, the conclusions would be more justified if they discussed these limitations of the study.

We agree with the reviewer it seems odd that, across studies, identification accuracy consistently appears to be less than chance. We did not find that this difference from chance was significant in our planned tests, but, in aggregate, we do find evidence that participants were worse than 50% accurate (we had summarized this meta-analysis in our supplement). However, we hesitate to draw strong conclusions from post-hoc internal meta-analyses and so do not explore this aggregate finding further.

Regarding the stimuli themselves, we discussed this issue in a response to Referee 1 within this letter. To review, we did control for certain differences between sounds, and other differences are not clearly relevant to address (e.g., because they may be inherent to the basic difference between sounds of infectious vs. non-infectious origins). However, we do now include more limitations regarding the stimuli in our discussion (Lines 354-357): "… our set may have nonetheless been limited in type, quality, and breadth of eliciting conditions."

OTHER CONCERNS
Pg 47 – citation for "existing work suggests"?

This claim was used to introduce the topic in the first paragraph of the paper. We return to it later in the introduction (Lines 64-79), where we provide references.

Pg 51 – "surprisingly" – not clear to me why this is susprising. Do we have a strong prior *about this in the opposite direction?'*

We use surprising to mean that the results are unexpected given the hypotheses we laid out.

Pg 62 – 64 – also consider citing Curtis V. A. 2014. Infection-avoidance behaviour in humans and other animals. Trends in Immunology 35:457–464.

We agree about the relevance of this reference and have added it (Line 60).

You refer to "people" throughout the intro, but this makes it unclear where the sample populations come from. For example, Curtis 2004 is an online cross-cultural sample while some of the other studies in the intro are undergrads. Worth clarifying, for sake of transparency.

We have included a summary about the sample characteristics of studies included in our review (Lines 61-64): "Humans also use a variety of sensory cues to detect pathogen presence, although research with humans is relatively recent and commonly restricted to Western undergraduates. Moreover, human evidence is limited by relatively few studies within any single modality."

A general note for the intro – do you know how well replicated these studies are? For example, the LPS and health rating study? If there are concerns about the replicability of any of this work, it's worth noting, so that we don't get an overconfident picture of current knowledge.

To our knowledge, no published studies have either replicated or failed to replicate the LPS and health rating finding or many of the other reported findings. We do now explicitly point to the need for further evidence and replication in these literatures (Lines 61-64 and 360-363).

Line 83 – belching – how is this at all relevant to adaptations for pathogen detection? Also, the p-value for the effect of auditory cues on that study provides weak evidence against the null (see below quote). I'm not saying that this means the effect isn't "real", but again, it really would be a more comprehensive intro if you provided more information about the evidentiary value of previous work, instead of just putting it in the typical narrative frame of "past work has shown x y z". I am guilty of doing this myself, but I think that we as a field can do better.
- "The olfactory disgust stimuli showed the highest significance difference compared to the control ($p < 0.001$) while visual disgust and auditory disgust showed smaller responses, but they still had significantly higher probability to increase hand washing attempts than the control (for all cases $p = 0.025$).

Belching involves some similar mechanical processes to coughing and could be perceived to spew germs into the air. Speaking to the broader issue of whether an introduction should compile existing literature or whether it should also attempt to interpret the validity of that work, we believe such a detailed review is beyond the scope of the current paper, particularly given space limitations. However, given the referee's point, we have tried to make a general conclusion about the evidentiary value of this literature. In our revised manuscript, we have noted that the literature on sensory modalities used to detect infection is rather shallow, containing initial evidence for a variety of effects rather than multiple investigations (replications) of a single sensory modality in the context of infection detection (Lines 61-64).

Line 91-93 – argument that mistaking infectious person as non-infectious is less costly than the latter error – what evidence do we have for this? For instance, the expected costs of this mistake will depend on the base rate with which you encounted sick vs not sick individuals. If you're constantly encountering healthy individuals but not interacting with them, this "small" cost will be large in aggregate. If you're going to use error management stuff, it's worth taking a look (and citing) this, for a more nuanced theoretical perspective: McKay, R., & Efferson, C. (2010). The subtleties of error management. Evolution and human behavior, 31(5), 309-319.

We cannot be sure, but we think R3 meant to ask whether there is evidence that it is *more* costly to mistake an infectious person as non-infectious than to mistake a non-infectious person as infectious (Lines 88-96). First, we agree that expected costs will depend on the base rate at which individuals encounter infectious vs. non-infection individuals just as they will depend on the potential damage incurred by specific forms of infection. We attempted to address base rate influences in our studies by providing this information to participants in Study 2. Of course, this only addresses how perceivers explicitly process such information in an immediate situation. We do know that the base rate of infected people is not zero outside the laboratory. Infectious disease is common in modern human populations and was likely even more so in the ancestral past.

We also agree that costs of a single error may be smaller than the cost of many smaller errors in aggregate. However, this is a moot point if a single false negative error results in a fatal infection. Consider that infectious diseases are thought to be one of the strongest selective forces on human evolution. Even recent World Health Organization statistics point to almost a quarter of all deaths being due to infectious disease (this number has thankfully fallen in the past decade).

Last, to clarify, we are not arguing that judging people as infectious will always result in "encountering healthy individuals but not interacting with them" (which could indeed be quite costly). In some percentage of situations, this avoidance will occur though. Existing error management research indicates that perceptions of strangers (which all sound stimuli used in the studies came from) are commonly biased toward inferences of threat (i.e., minimizing misses rather than false alarms), and it is not unreasonable to think therefore that people wish to avoid most interactions with potentially sick strangers. But other possibilities exist as well. For instance, judging another person as sick could lead people to engage in certain prophylactic behaviors (e.g., not sharing food, not kissing) while still maintaining reasonable levels of interaction.

Regarding McKay and Efferson (2010), we agree that biased judgments may stem from a cognitive bias (e.g., biased perceptions, biased beliefs) or from a cognitive evaluation (e.g., reasoned beliefs incorporating base rates and costs). This discussion is more nuanced and in-depth than we have space to elaborate on in the manuscript, but we have referenced this idea by noting that the expressed bias could be functional regardless of its particular origins (Lines 95-96).

Line 107 – Research Overview – I think that, if you want to mention the pilot study, you need to provide more information about what was done exactly, whether it was pre-registered, what motivated the study, etc. I appreciate the transparency of noting that you ran this study, but I would appreciate more information about how you changed your design as a result of it, what you learned, why you did the study in the first place, etc.

Details for the pilot study are included in our supplementary repository (but not the manuscript because of space limitations). The methods for this study are identical to those in Study 1, although fewer stimuli were used.

Line 130 – do you mean "different" types of sounds?

Thank you for this point. We indeed meant different types of sounds and have edited this sentence.

Footnote #1 – I don't understand this. Can you please elaborate on exactly what you did? And also report all analyses in which you used both a subset of stimuli and all the stimuli?

To address the confusion about this footnote, we have elevated it to the main manuscript text (Lines 170-174). In essence, to ensure a conservative test of our stimuli, we conducted additional analyses limiting the infected target group to only those individuals explicitly receiving medical diagnoses (i.e., the subset of stimuli for which we could most confidently say were from infected sources). There was no change to the key findings. (We report detail on these analyses in detail in our supplemental repository.

Line 136 – who were these coughs from? That is, were they generated from individuals all over the world, from the U.S., etc? It may be more difficult for people to detect infectious disease state for individuals from different cultural backgrounds or something. I don't know, but you should provide this information. Also, have you made this stimulus set openly available on the OSF?

Sound clips were taken from U.S.-based videos and participants were limited to U.S.-based individuals. We have edited the manuscript to specify the U.S. origin of the stimuli, which we agree is a helpful point to specify (Line 132-133). All of our materials, as well as a spreadsheet describing each stimulus video, are available in our supplement. These are accessible to referees and will be freely available to readers if the manuscript is accepted for publication.

Line 190 – is a rating of 6.7 out of 9 really "highly certain"? Also address this in the rest of the manuscript (e.g., line 256)

Though average ratings of certainty are statistically above the midpoint in every study (Lines 192-193; Lines 251-252; Lines 309-310), we agree that "highly" is a subjective interpretation.

We are have replaced “highly” certain with “reasonably” certain to indicate this subjective interpretation.

Line 197 and 199– report the exact p value. Same for everywhere else in the manuscript (e.g., lines 302 – 306).

We have added z-statistics and p-values for the conditional effects.

Figures – I had a hard time telling which was “figure 1”. If it is the figure on page 23, why doesn’t it show study 1?

As mentioned in an earlier response, the studies in this figure were poorly labeled. We have relabeled the figure to correctly refer to the pilot study and Studies 1-3.

Pg 218 – please elaborate. I don’t quite follow this base rate idea. Is it that you’re saying people usually encounter healthy people, and so they may have a prior that sounds indicate non-infection? If so, then you should find lots of false negatives, right? As opposed to the false-positive narrative that was presented in the intro?

This is a useful point to address. We believe that, in the natural world, reasonable arguments could be made about many types of base rate distributions. Presumably, these may often reflect the idea that most relevant sounds do not indicate infection. However, in the context of a study, it is certainly possible that participants will not use their common base rate presumptions. Therefore, to explicitly test whether base rate information accounts for the failure to find identification differences in sound origin, we specified these rates in Study 2 (Lines 218-220). According to the results of Study 2, base rate information does not account for our findings.

Pg 321 – 323 – What theories are these exactly? To me, this study came across as completely exploratory “Hey, people haven’t looked at sound yet, why don’t we do that?”. And that’s totally fine for me – we need more honest high-quality exploratory research. But then here in the discussion it’s written up as if it was a confirmatory thing testing a theory. Really doesn’t seem that way to me. I don’t see how the sound thing is a deductive prediction from a theory. I think it should be written up as exploratory.

We based our predictions off of an understanding of error management theory (itself a subset of signal detection theory). We do agree with the reviewer, however, that framing the investigation as more exploratory is appropriate given the lack of prior research on this sensory modality and the ambiguity with which a reader might see error management leading to specific predictions in this context. To better address this, we now state that “Though the nascent state of the literature makes this investigation largely exploratory in nature, the current work presents a test of these alternative hypotheses and advances future theorizing about pathogen threat processing.” (Lines 360-363)

References

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://webcom.upmf-grenoble.fr/LIP/Perso/DMuller/M2R/R_et_Mixed/documents/Bates-book.pdf

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.

Hauser, D. J., & Schwarz, N. (2015). It'sa trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, *5*(2), 2158244015584617.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. Com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Snowberg, E., & Yariv, L. (2018). *Testing the Waters: Behavior Across Participant Pools*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3210415

Appendix C

2020-04-25

Dear Dr. Brosnan,

We previously submitted a manuscript titled "Sounds of sickness: Can people identify infectious disease using auditory cues?" (RSPB-2019_1719) for review at *Proceedings of the Royal Society B: Biological Sciences*. Our manuscript received a reject decision, though you decided to accept a revision presuming we could fully address the referees' concerns. Our critical goals in revising our work were to include more justification or evidence that our stimuli were by and large representative of what the individuals featured in those stimuli claimed and to consider an alternate title that limits the generalizability of our claim. We have revised our manuscript to meet these goals. Summarizing our revisions, (1) we had research assistants code demographics (i.e., age, sex, race/ethnicity, SES) of the target individuals in the stimuli, and we tested whether our findings held when controlling for these potentially confounding influences. They did, thereby increasing confidence in our stimulus interpretations. We describe this process in more detail below. Next, (2) we changed our paper title so it better represents the conclusions our findings can support. Now we title our paper "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" Further, we have modified our claims/interpretations throughout the paper to address this same concern.

We appreciate your handling of our manuscript along with the AE and for noting the positive comments made by referees. Below, we respond to each referee point-by-point (referee comments are italicized to help distinguish our responses). As the concerns of Referee 3 were particularly relevant to the critical revisions, we respond first to Referee 3 below, followed by Referees 1 and 2.

Reviewer(s)' Comments to Author:

Referee: 3

Comments to the Author(s).
Review of first round of revisions for:
Sounds of sickness: Can people identify infectious disease using auditory cues?
Manuscript ID: RSPB-2019_1719

TO THE EDITOR/authors:
The authors have submitted a revised version of the manuscript. In this revision, they claim that most major points raised by the referees have been addressed. I think that the revised manuscript represents an improvement in clarity and detail over the original submission. That said, I still

have concerns regarding the ability of this study to test whether people can identify infectious disease using auditory cues.

My most substantive remaining concern, which I do not feel is adequately addressed, concerns the ability of the study design to adequately answer the research question. The use of user-uploaded Youtube stimuli mean that there are plausible confounds between conditions, which make it difficult to interpret an effect in any direction in this study. I think that it is the responsibility of the original authors to do everything possible to control for such confounds, and I do not think that the current design adequately achieves this. Infected videos may have been uploaded by people who are different in any number of ways (e.g., age, sex, non-infectious disease comordibity), or had a number of confounds (e.g., video quality).

Our understanding of the referee's concern is that our approach does not go far enough to control for potential confounds, such as the demographics of people featured in the stimuli, which may be suppressing evidence of true accuracy in participant responses. To address this issue, we focused on leveraging the value of the existing stimuli by coding and controlling for conceptually plausible confounds. We took this approach because we believe our assignment of stimuli to infectious and non-infectious categories was largely accurate based on the clear information available in the videos (this was especially the case for the subset of stimuli we used in our sensitivity analyses). To measure potentially confounding variables, we had research assistants who were blind to stimulus category watch and code standard demographics (age, gender, race/ethnicity, and SES) of the people who made the sounds we used as stimuli. Then we included this information in our analyses to test the referee's confound hypothesis. We have made all our materials and coding data available at our OSF Project (i.e., online supplement), along with the procedure, results, and interpretations of the additional analyses.

The key goal of this approach was to test whether demographic variables that might differ by stimulus condition acted like suppressors in the prior analyses (i.e., accuracy would improve when we account for these variables). A demographic variable that acts as a suppressor should (1) correlate with our sound origin variable (i.e., infectious or not), and (2) increase the association between our sound origin variable and participant judgments when it is controlled for (which is equivalent to increasing overall accuracy).

We found no sufficient evidence for demographic variable suppression effects in any study. Only gender significantly correlated with our sound origin variable (criterion 1 above). However, controlling for gender did not significantly increase the association between our sound origin variable and participant judgments (criterion 2 above). Thus, when we statistically control for standard demographic variables, we still find no sufficient evidence that people can accurately identify infectious diseases from sound clips of coughs and

sneezes. This suggests that a number of plausible confounds might not play a meaningful role in our studies. Further, certain other suggested confounding variables (e.g., video quality) cannot play a role because sound clips were extracted from the videos prior to data collection. Participants never saw the videos, and they also rated the clarity of the sound clips during the studies (we found no significant differences across categories, with both category means estimated at 8.4 (out of 9 possible) scale points).

On the other hand, a finding of no effect is also difficult to interpret, because we do not actually know people's infectious disease status. We also don't know what the difference in the severity of the affliction is between the different types of people – are we comparing people with mild allergies to severe infectious diseases or vice versa? The authors attempt to address this by limiting analyses to cases where they are more confident that the individual is sick (e.g., because the video is uploaded from a hospital) but this does not solve the problem that 1) the study still relies on participants to accurately report their diagnosis and 2) you don't know the disease status of people who generated the non-sick stimuli (e.g., they may have been sick at the time of stuffing hot peppers in their nose; people with bronchitis may have had chronic bronchitis, which, in contrast to acute bronchitis, is not infectious). I do not mean to say that the study is entirely uninformative – the probability that the hospital people are indeed sick with some infectious disease is higher than the pepper-powder up-nose people – but it is hard to know what these probabilities are.

Our understanding of the referee's concern is that we do not know with 100% certainty the infection status (infectious or not) of the targets. This is true. However, we suggest two reasons why this may not represent much of a problem in our studies. First, as mentioned by the reviewer, we conducted analyses reported in the prior manuscript submission on a limited subset of stimuli comprising individuals who reported medical diagnoses of infection status (e.g., while in the hospital). We see no convincing reason that individuals admitted to the hospital would be likely to misreport their status as infected. The reviewer also suggests that individuals in our non-infected category may have been suffering from sickness at the time their videos were recorded. Though possible, we believe it is more plausible that, given what we know about the phenomenological effects of infection (e.g., sickness behavior), sick individuals would be less likely than non-sick individuals to take part in the activities shown in the videos, such as putting substances up their noses. Thus, we believe our category assignments reflect the best interpretation of the stimuli.

Of course, reasonable minds may disagree. This is why our second approach was to compare subsets of sounds from such high-likelihood infected targets to *every* possible, equally-sized subset of non-infectious sounds. Simply put, the resulting poor accuracy when trying to distinguish between sounds from low vs. high-likelihood infected targets did not support the hypothesis that people can detect infection from coughs and sneezes sounds.

Certainly, these studies cannot rule out every possible alternative explanation. For example, as the reviewer proposes, perhaps the severity of affliction plays some role in infection accuracy. Or perhaps people could detect infection from sentences spoken by non-infected vs. infected targets rather than coughs and sneezes. We discuss such possibilities in our General Discussion (lines 337-352). We also limit our claims to sounds of coughs and sneezes (i.e., not sounds in general) throughout the manuscript.

Less major concerns

I don't have a good sense for how the stimuli used in this study adequately capture the properties of infectious and non-infectious diseases that people are most likely to encounter. Some the stimuli (e.g., whooping cough) are infectious diseases that people don't encounter very often, while the one arguably most-frequently encountered by people (e.g., cold) is labelled as "cold maybe", so we don't know exactly what these people have. Participants are also provided with such minimal information (e.g., 1 second decontextualized clips) that it is not clear to me how informative a null effect is regarding people's ability to do this in the real world. This is indeed a different question than the one addressed in this paper (i.e., whether people can tell infectious disease status from short-audio clips of a small subset of diseases) but it seems to me to be closer to the broader question that the paper is attempting to address.

We agree with the referee's point about generalizability. We have very much limited our conclusions in the revised manuscript. We have even limited the generalizability of claims in our title by now referring only to the types of sounds used in the studies: "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" We also agree that in typical settings, people do not encounter decontextualized sensory information such as disembodied sound clips. That said, isolating this information is a standard way to examine the unique influence of a sensory modality. For example, in tests of visual processing, participants would not also be given auditory, scent, or tactile information.

Other minor remaining issues which should be addressed are 1) the lack of discussion of why the study found strong evidence that participants are worse than chance

We can understand why the referee would want an explanation for the finding that participants were worse than chance in identifying the infectious origins of the sounds. We want to know why too, but we have not found support for a shortlist of hypotheses. Namely, we did not find support for differences (between stimulus categories) due to demographics nor differences due to the type of error (i.e., participants made approximately equal numbers of false positive and false negative errors). Even evidence for

differences in disgust is weak given significant but still large *p*-values ($0.01 \leq p < 0.05$) and wide confidence intervals (e.g., between 0.003 and 1.5 scale points). We write in footnote 2 (p.12): "Here and across studies, overall accuracy falls below 50%. Adjusting for sound type (cough or sneeze), target demographics, and sound disgustingness did not substantively change this pattern."

and 2) the lack of detail regarding what exactly was specified in the pre-reg versus what was conducted (e.g., the pre-reg said things like "We will use repeated measures analyses and contrasts to examine the effect of condition on mean disgust ratings." without specifying the specific functional form of the analyses – the authors should be more transparent about this)

We understand the referee's concern that the manuscript does not exactly specify what is different between our reported analyses and our preregistered analyses. We also understand our preregistered analyses could have been improved by specifying the functional form of our regression. To accommodate such concerns while hewing closely to our word limit, we now note in the manuscript that our reported analyses are different from our preregistered analyses (line 155), and readers can see detailed discrepancies in a spreadsheet in our supplement.

, and 3) lack of consistent presentation of p-values and confidence intervals (e.g., lines 198-202; line 241).

We thank the referee for their statistical reporting scrutiny. We provide *p*-values through lines 198-202 except when the values are smaller than $p < .001$, per APA style (Edition 7). Additionally, we have added the non-significant *p*-value next to our 95% confidence interval that includes 0 on line 242.

However, I think that other aspects of this paper are impressive in many ways. The study is laudably open and transparent and provides detailed information regarding stimuli, data, analysis code, and the extensive sensitivity analyses. As such, I think that the main results are robust to alternative ways of analyzing the data. The large sample size provides quite certain effect-size estimates, in aggregate (conditional on the design); and even if the representativeness/generalizability of the stimuli is unclear, I think it is interesting, in principle, to explore whether people can tell infectious-disease status from audio clips. It is interesting to know that people's level of disgust is a strong predictor of whether they judge someone as infected, regardless of the person's actual infectious disease status, and that certain information (e.g., base rates) may not have much of an effect on people's judgements.

We thank the referee for their interest and positive comments. The feedback has greatly helped in strengthening the manuscript.

Referee: 2

Comments to the Author(s).
I think the authors have done a superb job addressing the issues raised during the review process. I was disappointed to see so many comments raised that I felt were very clearly addressed in the supp mats and original submission, so it is great that the authors were given the opportunity to clarify those points in a revision. I think this makes an important and timely contribution to the ongoing debate about whether or not humans display reliable cues of health.

We thank the referee for their very positive feedback.

Referee: 1

Comments to the Author(s).
I have read thoroughly through the manuscript and can see that the authors have made a number of improvements. While the videos are of non-experimentally controlled origin, and McTurk has a number of weaknesses, the authors have done a thorough job in maximizing the designs and also reasonably deal and argue of their validity. I also agree that the outcomes are scientifically sound, i.e. that humans do not seem very good at judging whether sneezes coughs are from a contagious origin stimulated by something else, and that rated disgust and certainty are not related to whether the sounds were from a contagious individual or not. There are, however, some aspects of the ms that are not properly addressed.

1) The title makes the reader believe they have tested whether people can identify infectious disease using auditory cues. This is a very bold statement considering that the authors have only included specific short sounds from sick people. For the statements made about the ability to detect sick sounds, they also have to show proof of concept that they actually used the most relevant sickness sounds. Since this is not done in the ms, only including particular sounds presented for a few seconds, the title should be updated to fit the question studied. For example. 'Sounds of sickness: Can people identify infectious disease using auditory clips of sneezes and coughs?'

We understand the referee's concern and have changed our title to nearly the exact title the referee suggested: "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" We agree that this change does a better job of not speaking beyond the data.

2) It is up to the authors to disagree with my previous statement ".... However, it seems very likely that coughs and sneezes are strong physiological reflexes that has been rather consistent *through evolution, and can be triggered by a number of aspects that the authors tap into, e.g. pathogenic infections as well as allergens, etc…" i.e. that there has NOT been a large evolutionary pressure for developing different coughs and sneezes depending on the causation. It would be suitable for the authors to include/discuss this possibility in the limitations, particularly since it is a likely reason for the 'null findings'.*

We understand the referee's concern. It is possible that evolution has not shaped coughs and sneezes to reveal human-detectable infectiousness information. In contrast, we propose evolution could have favored perceptual mechanisms for detecting infections via coughs and sneezes, given the advantages of having versus not have such information. Moreover, algorithms can be trained to identify human illnesses from sounds (Porter et al., 2019, line 341), and doctors are also trained in medical school to identify certain illnesses through sound. These facts suggest that infection information is present in such sounds, allowing the possibility that perceivers could make accurate identifications, in principle. However, we have revised our manuscript to make these alternate possibilities clear (lines 337-352).

3) Going through the OSF-online material (also appreciating the effort to find sick and not sick related sounds), I could not open the links in the "url1_archive", and only the original flies.

We thank the referee for going through our OSF-online material. We had tried to archive the links to the videos in case the links broke over time (i.e., "link rot"), but some of these links do not seem to be working. We have downloaded all available videos and have made them available via our OSF Project. We have also removed broken links from our stimuli information spreadsheet.

4) I am surprised that individuals did not differ in their rating behavior or accuracy (see method, rows 156-159), This is very different from the analyses made by Kurvers & Wolf, 2018, and large parts of the literature where individuals often show large differences in rating behavior. It seems appropriate to include 'rater' as a factor in all models since there are 4 separate studies and it may differ from study to study, and also report the ICC data from both 'rater' and 'sound originator'.

We agree with the referee that, in principle and consistent with previous empirical reports, differences between raters should account for some differences in accuracy (i.e., a significant rater factor, either fixed or random). However, we found no sufficient evidence that the rater factor accounted for significant variability in accuracy. The model estimated 0 variance for the rater factor (i.e., random rater factor, which is the same thing as random intercepts). There are technical reasons why the model estimated exactly 0 (see p. 10 Bates, 2010), but, importantly, it made no substantive difference whether or not we included this factor. Therefore, it was excluded. However, we reported models with and without this

factor in our supplement, and, importantly, this point was made in the prior versions of our manuscript (lines 152-178), thereby addressing the referee’s concern.

Appendix D

2020-05-##

Dear Dr. Brosnan,

We previously submitted a manuscript titled "Sounds of sickness: Can people identify infectious disease using sounds of coughs and sneezes?" (RSPB-2020-0944) for review at *Proceedings of the Royal Society B: Biological Sciences*. We are excited that you and the AE have recommended our paper for acceptance—thank you very much!

Also, we appreciate your handling of our manuscript along with the AE. Below, we briefly respond to each of the referee's comments. We italicized referee comments to distinguish them from our responses.

Associate Editor
Board Member
Comments to Author:
We have now heard from our reviewers about your revised manuscript. As you see, they are both positive. Reviewer 1 has some minor comments that you should be able to deal with easily before we move forward. I am recommending acceptance. Congratulations on an interesting paper.

This is exciting news—thank you!

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s).
The authors have done a good job with updating the ms. I have a few minor comments

1) General discussion, 3rd paragraph: Aouthors state "Recent work indicates that this sound variation is available (e.g., to statistical learning algorithms; Porter et al., 2019), but perhaps human hearing mechanisms cannot use it reliably, even with clinical training (e.g., Smith et al., 2006). ," An alternative is likely, and I suggest to directly mention that the sound variation in coughs and sneezes are rather limited.

We agree with the referee and have revised that sentence to include the possibility that such sound variation is very limited. The new sentence (lines 341-343) now reads: "Such sound variation could be very limited, but recent work suggests that this variation is available (e.g., to statistical learning algorithms; Porter et al., 2019). However, human

hearing mechanisms may not be able to use it reliably, even with clinical training (e.g., Smith et al., 2006).”

2) References: Please check that references are correct. I know the study by Olsson et al, did not have A. Soop as last author (it is M Lekander).

This is a great catch! We revised the reference (lines 433-437):

Olsson, M. J., Lundström, J. N., Kimball, B. A., Gordon, A. R., Karshikoff, B., Hosseini, N., Sorjonen, K., Olgart Höglund, C., Solares, C., Soop, A., Axelsson, J., & Lekander, M. (2014). *The Scent of Disease: Human Body Odor Contains an Early Chemosensory Cue of Sickness. Psychological Science, 25*(3), 817–823. https://doi.org/10.1177/0956797613515681

3) The authors have uploaded the clips for others to use also on the webarchive, but I still have problems opening most of the youtube clips provided on this page
https://nickmichalak.shinyapps.io/sounds_of_sickness_sensitivity/
Please check the links are working.

We thank the referee for going through our OSF-online material. We had tried to archive the links to the videos in case the links broke over time (i.e., “link rot”), but some of these links do not seem to be working. We have downloaded all available videos and have made them available via our OSF Project. We left the broken links in the spreadsheet, but we added a column that notes whether the link is broken.

Referee: 3

Comments to the Author(s).
NA. I appreciate your effort to seriously engage with the reviewer comments and to make all materials, data, and code openly available so that the scientific community can easily judge the quality of your paper.

We appreciate the referee’s positive feedback regarding our response to reviews. Their critical feedback really strengthened our paper. We’re also glad they appreciate our open materials, data, and code.