PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St. Michael's Hospital, Toronto, Canada
AUTHORS	Armstrong, Bonnie; Spaniol, Julia; Persaud, Nav

VERSION 1 – REVIEW

REVIEWER	Dr. Michelle McDowell Max Planck Institute for Human Development, Harding Center for Risk Literacy, Berlin, Germany.
REVIEW RETURNED	25-Sep-2017

GENERAL COMMENTS	The authors contrast simulated experience-based probabilities with a numerical format they refer to as natural frequencies. Natural frequencies have been confused in the literature with normalised frequencies, a descriptive format that is computationally more complex and is similar to conditional probability formats. Natural frequencies, are not normalised, meaning that the computation is simpler. In fact, some authors have suggested that natural frequencies facilitate understanding because they are the end product of experience-based sampling in that they present, as a description, the joint frequencies that result from a natural sampling process. As there was no detail in the present paper to confirm that the correct format that is mentioned was used in the study, I looked up the supplementary material from a previous publication by the authors that was cited within the present manuscript. Here, the format that has been used is normalised frequencies, not natural frequencies. I refer the authors to the publication by Gigerenzer and Hoffrage (2007; Behavioral & Brain Sciences, 30 (3), Figure 1 on p.265) for a visual representation of different formats. If the authors compare Figure 1(1) with 1(3), the difference between the representation formats is shown. Importantly, natural frequencies are joint frequencies. In light of this, I have two suggestions for the authors on how to frame the study: 1. Change all mentions (and citations) of natural frequencies in the study to correctly identify them as normalised frequencies, and highlight that normalised frequencies are similar to conditional probability representations, but the numerical format is in frequencies. In the discussion somewhere, it would be important to mention that natural frequencies are an alternative numerical format that has been shown to enhance performance (see point below).

	2. I do think there is still merit in reporting the study, and the argument is rather that simulated (experience) formats based on natural sampling are better than normalised frequency descriptive formats. Normalised formats are still used to communicate test statistics to health professionals, despite a vast literature showing that people have difficulty solving problems presented in this way (e.g., calculating the PPV). Natural frequencies are a format that improve on normalised formats, based in part on the concept of natural sampling that underlies the experience-based presentation used here. It would therefore be relevant to compare experience-based presentations to natural frequency descriptive formats. The authors could mention this in their discussion. However, I do emphasise that the formats need to be referred to correctly throughout the manuscript. I have a few more minor comments: The abstract does not list the mean absolute error (only the 95% CI). Please include this in the abstract to facilitate understanding of the results. Further, the metric for evaluating accuracy is not clear in the abstract. It is not mentioned until the Method. As such, the reader may be confused about lower values indicating better accuracy (typically results from these studies have been reported as percentage correct, higher=better). Did the authors expect any differences across the three diagnostic test conditions (gold standard vs. low sensitivity vs. low specificity) or were these conditions only included for repeated tests for participants (e.g., to check the robustness of the effect across variations of the problem)? The Figure needs a caption and should explain what the axis means (value == MAE) and the red line.

REVIEWER	Penny Whiting Bristol Medical School, University of Bristol, England
REVIEW RETURNED	04-Oct-2017

GENERAL COMMENTS	General Although an interesting topic, I think manuscript currently lacks clarity and needs careful attention to make it suitable for publication. A key issue is that is not at all clear what information was provided to clinicians. Abstract: The design is not clear in the abstract - reading this alone I do not understand what you have done. I think the abstract needs to be completely re-written to convey what was done in the study and what was found. What is an “experience condition” and a numerical condition? What are the proportions in brackets? Are these the CI for the proportion of clinicians correctly judging PPV? Why have you not also included the actual proportion? These numbers still appear very low even for the “experience condition” – is that not also something to highlight? Why "posterior values" rather than the more commonly used "predictive values"?

	Background The background does not adequately introduce the topic and is difficult to follow. You say in the first paragraph what you are trying to look at and then repeat this in the last paragraph of the background, much of this information would be more appropriate in the methods. Avoid abbreviations such as OB-GYNs I don't think that our review (Whiting 2015) showed that probability of no disease given a negative test result is underestimated – you reference our review to back up this statement. We found "tendency to overestimation for both positive and negative test results." I do not know what you mean by details such as prevalence, test sensitivity and specificity being communicated via leaflets. Also, I am not sure that it is true that "clinicians interpret test results on the basis the basis of relevant statistics such as disease prevalence, test sensitivity and test specificity." You make this statement but do not support it with a reference. I do not find the term "experiential format" helpful. You are also not consistent in terminology – e.g. experience vs experiential. Methods I would like to see more details on the examples presented and exactly what information was provided. It is really not clear what you were comparing. In the background you state that "In the numerical format, participants were presented with explicit statistical summaries, framed in natural frequencies." In the methods you state "In the numerical format, participants read a summary about each diagnostic test, explicitly describing the disease prevalence (constant throughout the experiment), as well as test sensitivity and the false-positive rate." These are completely different ways of presenting information – which was it? Could you provide examples as web appendices? Looking at the results presented you have the "mean absolute error rate". This is not clearly described in the methods. Why have you chosen this rather than presenting something that would be much easier to understand such as the proportion of clinicians who correctly estimated PPV/NPV within say 5 or 10%? Or the difference in estimated and actual PPV/NPV? Or is that what you mean by the error rate? Table 1 This is rather lacking in detail. Is there not any other information that would be helpful here? Figures. It is unclear what these are showing. What does "value" mean? What are the red lines?

REVIEWER	Dr Clare Davenport, Senior Clinical Lecturer University of Birmingham UK
REVIEW RETURNED	07-Oct-2017

GENERAL COMMENTS	I consider this an interesting experiment and to my knowledge novel in its approach in using simulated patients as a form as fast paced experiential learning. Description of methods: The influence of language and presentation

	on understanding of test accuracy information is the subject of this research and the authors reference some of the existing work undertaken in this area. However the description of how information was presented to research participants is lacking in important detail. The language used to describe (essentially define) the PPV and the false positive rate in the numerical presentation as well as how participants were asked to express their estimates is an important potential confounder in this comparison. I am unclear why numerical information was presented in the form of the sensitivity and the false positive rate. This is another potential source of confounding: the numerical presentation describes test performance using test +ve results only whereas the experience presentation presents test positive and test negative outcomes.There is evidence that deriving the probability of disease after a positive test result differs to after a positive result. The values of prevalence and the variation in the sensitivity and specificity of the 3 tests may also affect estimates (for example evidence from the risk communication literature demonstrates that manipulation of small numbers increases difficulty). Independently, the prevalence of the target condition may have had an impact on the difference in accuracy of estimates of PPV and NPV in both groups (for example in the experience format, a low prevalence would have made recall of the denominator (disease positives) easier. What is meant by 'representative' patients? Some more information could be provided in the main body of text but I consider including the questionnaire as an appendix is necessary. Presentation of results: I am not sure table 1 adds value. The figure would benefit from better labelling (correct PPV and NPV values; numbers in the bars). Abstract, discussion, study limitations: I consider that conclusion are overstated. For example in the abstract tat statement 'experiential exposure significantly improves clinicians PPV and NPV judgements': is not accurate for NPV judgements. I consider use of the term 'judgement' misleading as this implies using information whereas this study is concerned with understanding information. The impact of understanding about test accuracy on diagnostic judgements is a different question. The importance and implications of the results for practice is not clear. There is already a wealth of evidence demonstrating that the use of PVs is intuitive as they convey the probability of disease in an individual rather than the probability of a test result in individuals with or without disease. Also that health professionals' knowledge of test performance is heavily influenced by experience. However because the derivation of PVs is dependent on prevalence the context in which this type of experiential learning takes place is key to its value and accuracy. The notion of fast paced, simulated experience may represent a valuable tool in terms of illustrating the concept of test accuracy but its role in improving knowledge about the performance of tests in use is difficult to imagine. References: Christensen-Szalanski JJJ, Bushyhead JB. Physicians' Misunderstanding of Normal Findings. Med Decis Making 1983; 3:169-175.

	Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. Journal of General Internal Medicine 1994 (b); 9:488-495. Poses RM, Cebul RD, Wigton RS. You can lead a horse to water - improving physicians' knowledge of probabilities may not affect their decisions. Medical Decision Making 1995; 15(1):65-75. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: Practicing physicians' use of quantitative measures of test accuracy. The American Journal of Medicine 1998; 104(4):374-380.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Dr. Michelle McDowell
Institution and Country: Max Planck Institute for Human Development, Harding Center for Risk Literacy, Berlin, Germany.
Please state any competing interests: None declared.

1. The authors contrast simulated experience-based probabilities with a numerical format they refer to as natural frequencies. Natural frequencies have been confused in the literature with normalised frequencies, a descriptive format that is computationally more complex and is similar to conditional probability formats. Natural frequencies, are not normalised, meaning that the computation is simpler. In fact, some authors have suggested that natural frequencies facilitate understanding because they are the end product of experience-based sampling in that they present, as a description, the joint frequencies that result from a natural sampling process. As there was no detail in the present paper to confirm that the correct format that is mentioned was used in the study, I looked up the supplementary material from a previous publication by the authors that was cited within the present manuscript. Here, the format that has been used is normalised frequencies, not natural frequencies. I refer the authors to the publication by Gigerenzer and Hoffrage (2007; Behavioral & Brain Sciences, 30 (3), Figure 1 on p.265) for a visual representation of different formats. If the authors compare Figure 1(1) with 1(3), the difference between the representation formats is shown. Importantly, natural frequencies are joint frequencies.

2. In light of this, I have two suggestions for the authors on how to frame the study: Change all mentions (and citations) of natural frequencies in the study to correctly identify them as normalised frequencies, and highlight that normalised frequencies are similar to conditional probability representations, but the numerical format is in frequencies. In the discussion somewhere, it would be important to mention that natural frequencies are an alternative numerical format that has been shown to enhance performance (see point below).

Response: We thank the reviewer for this important comment. We have corrected the terminology throughout the manuscript and we now mention previous research on natural frequencies in the introduction and discussion (p. 4,8,9).

3. I do think there is still merit in reporting the study, and the argument is rather that simulated (experience) formats based on natural sampling are better than normalised frequency descriptive formats. Normalised formats are still used to communicate test statistics to health professionals, despite a vast literature showing that people have difficulty solving problems presented in this way (e.g., calculating the PPV). Natural frequencies are a format that improve on normalised formats, based in part on the concept of natural sampling that underlies the experience-based presentation used here. It would therefore be relevant to compare experience-based presentations to natural

frequency descriptive formats. The authors could mention this in their discussion. However, I do emphasise that the formats need to be referred to correctly throughout the manuscript.

Response: In line with this suggestion, we now mention as a future direction the comparison of experience and natural frequency formats (p. 9).

I have a few more minor comments:

4. The abstract does not list the mean absolute error (only the 95% CI). Please include this in the abstract to facilitate understanding of the results.

Response: The abstract now includes the mean absolute errors.

5. Further, the metric for evaluating accuracy is not clear in the abstract. It is not mentioned until the Method. As such, the reader may be confused about lower values indicating better accuracy (typically results from these studies have been reported as percentage correct, higher=better).

Response: We have added a description at the beginning of the results section of the abstract to clarify the metric for evaluating accuracy.

6. Did the authors expect any differences across the three diagnostic test conditions (gold standard vs. low sensitivity vs. low specificity) or were these conditions only included for repeated tests for participants (e.g., to check the robustness of the effect across variations of the problem)?

Response: We had no a priori hypotheses for differences among the tests. Rather, we included 3 tests to check the robustness of the format effect. We now say this explicitly in the paper (p. 5,6,8).

7. The Figure needs a caption and should explain what the axis means (value == MAE) and the red line.

Response: Figure 2 now has a caption that includes an explanation of the y-axis and the red line. The y-axis does not show the MAE but rather the mean estimates provided by participants. The red lines indicate the true values of PPV and NPV. Figure 2 thus provides additional information over and above the MAE results reported in the results section. To clarify this, we have now also included a note in the results section (p. 7).

Reviewer: 2
Reviewer Name: Penny Whiting
Institution and Country: Bristol Medical School, University of Bristol, England
Please state any competing interests: None

General
1. Although an interesting topic, I think manuscript currently lacks clarity and needs careful attention to make it suitable for publication. A key issue is that is not at all clear what information was provided to clinicians.

Response: We have revised the method section extensively to convey this information more clearly. As you suggest below, we now also provide examples of the numerical format (Appendix) and the experience format (Figure 1).

Abstract:

2. a) The design is not clear in the abstract - reading this alone I do not understand what you have done. I think the abstract needs to be completely re-written to convey what was done in the study and what was found.
b) What is an “experience condition” and a numerical condition?
c) What are the proportions in brackets? Are these the CI for the proportion of clinicians correctly judging PPV? Why have you not also included the actual proportion? These numbers still appear very low even for the “experience condition” – is that not also something to highlight?
d) Why "posterior values" rather than the more commonly used "predictive values"?

Response: We have rewritten the abstract to convey these details more clearly. We now use the term predictive values as suggested.

Background
3. The background does not adequately introduce the topic and is difficult to follow. You say in the first paragraph what you are trying to look at and then repeat this in the last paragraph of the background, much of this information would be more appropriate in the methods.

Response: We have extensively revised the background section and moved some information to the method section, as suggested.

4. Avoid abbreviations such as OB-GYNs.

Response: We have removed the abbreviation.

5. I don’t think that our review (Whiting 2015) showed that probability of no disease given a negative test result is underestimated – you reference our review to back up this statement. We found “tendency to overestimation for both positive and negative test results.”

Response: We thank the reviewer for pointing out this error. We have corrected it, and we now refer to the overestimation of positive and negative predictive values (see abstract and p. 4).

6. I do not know what you mean by details such as prevalence, test sensitivity and specificity being communicated via leaflets. Also, I am not sure that it is true that “clinicians interpret test results on the basis the basis of relevant statistics such as disease prevalence, test sensitivity and test specificity.” You make this statement but do not support it with a reference.

Response: We have removed the mention of leaflets. The example of numerical summaries being similar to leaflets (or in other words, a pamphlet) has been used in prior literature (see Gigerenzer, Gaissmaier, Kurz-Milcke, et al., 2007).

7. I do not find the term “experiential format” helpful. You are also not consistent in terminology – e.g. experience vs experiential.

Response: The term "experience format" has been used in the literature (Hau, Pleskac, Kiefer, & Hertwig, 2008; Tyszka & Sawicki, 2001; Fraenkel, Peters, Tyra, & Oelberg, 2015; Armstrong & Spaniol, 2017) so we continue its use to remain consistent with prior studies. We have replaced "experiential format" with "experience format" throughout.

8. Methods
I would like to see more details on the examples presented and exactly what information was provided. It is really not clear what you were comparing. In the background you state that “In the numerical format, participants were presented with explicit statistical summaries, framed in natural

frequencies." In the methods you state "In the numerical format, participants read a summary about each diagnostic test, explicitly describing the disease prevalence (constant throughout the experiment), as well as test sensitivity and the false-positive rate." These are completely different ways of presenting information – which was it? Could you provide examples as web appendices?

Response: We have rewritten the method section to provide greater clarity on both formats. As mentioned, we have also provided a new figure (the new Figure 1) and an Appendix.

9. Looking at the results presented you have the "mean absolute error rate". This is not clearly described in the methods. Why have you chosen this rather than presenting something that would be much easier to understand such as the proportion of clinicians who correctly estimated PPV/NPV within say 5 or 10%? Or the difference in estimated and actual PPV/NPV? Or is that what you mean by the error rate?

Response: The mean absolute error refers to the difference in estimated and actual PPV and NPV. We have now clarified this in the abstract and in the method section (ps. 2, and 7). Please also see our response to Reviewer 1, Comment 7.

10. Table 1
This is rather lacking in detail. Is there not any other information that would be helpful here?

We have removed the table that describes participant information, and have instead added it in-text at the beginning of the results section (p. 7).

11. Figures.
It is unclear what these are showing. What does "value" mean? What are the red lines?

Response: Figure 2 now has a caption that includes an explanation of the y-axis and the red line. The y-axis does not show the MAE but rather the mean estimates provided by participants. The red lines indicate the true values of PPV and NPV. Figure 2 thus provides additional information over and above the MAE results reported in the results section. To clarify this, we have now also included a note in the results section (p. 7).

Reviewer: 3
Reviewer Name: Dr Clare Davenport, Senior Clinical Lecturer
Institution and Country: University of Birmingham, UK
Please state any competing interests: None declared

Please leave your comments for the authors below
I consider this an interesting experiment and to my knowledge novel in its approach in using simulated patients as a form as fast paced experiential learning.

Description of methods:
1. The influence of language and presentation on understanding of test accuracy information is the subject of this research and the authors reference some of the existing work undertaken in this area. However the description of how information was presented to research participants is lacking in important detail. The language used to describe (essentially define) the PPV and the false positive rate in the numerical presentation as well as how participants were asked to express their estimates is an important potential confounder in this comparison. I am unclear why numerical information was presented in the form of the sensitivity and the false positive rate.

Response: As noted in our responses to Reviewers 1 and 2, we have extensively revised the method section to provide clearer descriptions of the numerical and experience formats. We have also provided a new figure (the new Figure 1) and an Appendix.

2.This is another potential source of confounding: the numerical presentation describes test performance using test +ve results only whereas the experience presentation presents test positive and test negative outcomes. There is evidence that deriving the probability of disease after a positive test result differs to after a positive result. The values of prevalence and the variation in the sensitivity and specificity of the 3 tests may also affect estimates (for example evidence from the risk communication literature demonstrates that manipulation of small numbers increases difficulty). Independently, the prevalence of the target condition may have had an impact on the difference in accuracy of estimates of PPV and NPV in both groups (for example in the experience format, a low prevalence would have made recall of the denominator (disease positives) easier. What is meant by 'representative' patients? Some more information could be provided in the main body of text but I consider including the questionnaire as an appendix is necessary.

Response: As noted in our responses to Reviewers 1 and 2, we have extensively revised the method section to provide clearer descriptions of the numerical and experience formats. We have also provided a new figure (the new Figure 1) and an Appendix. We hope this clarifies the above comment.

3. Presentation of results:
I am not sure table 1 adds value.

Response: We have removed the table that describes participant information, and have instead added it in-text at the beginning of the results section (p. 7).

The figure would benefit from better labeling (correct PPV and NPV values; numbers in the bars).

Figure 2 now has a caption that includes an explanation of the y-axis and the red line.
The y-axis does not show the MAE but rather the mean estimates provided by participants. The red lines indicate the true values of PPV and NPV. Figure 2 thus provides additional information over and above the MAE results reported in the results section. To clarify this, we have now also included a note in the results section (p. 7).

4. Abstract, discussion, study limitations:
I consider that conclusion are overstated. For example in the abstract tat statement 'experiential exposure significantly improves clinicians PPV and NPV judgements': is not accurate for NPV judgements.

Response: NPV estimates were indeed more accurate in the experience format compared to the numerical format. We now say this more clearly in the abstract and in the results section (ps. 2 and 7).

5. I consider use of the term 'judgement' misleading as this implies using information whereas this study is concerned with understanding information. The impact of understanding about test accuracy on diagnostic judgements is a different question.

Response: We have replaced all mention of “judgments” with the term “estimate”.

6. The importance and implications of the results for practice is not clear. There is already a wealth of evidence demonstrating that the use of PVs is intuitive as they convey the probability of disease in an individual rather than the probability of a test result in individuals with or without disease. Also that

health professionals' knowledge of test performance is heavily influenced by experience. However because the derivation of PVs is dependent on prevalence the context in which this type of experiential learning takes place is key to its value and accuracy. The notion of fast paced, simulated experience may represent a valuable tool in terms of illustrating the concept of test accuracy but its role in improving knowledge about the performance of tests in use is difficult to imagine.

Response: We have acknowledged that the current results do not reflect subsequent choice nor other clinical outcomes, and suggest this an additional avenue for future research (p. 9).

7. References:
Christensen-Szalanski JJJ, Bushyhead JB. Physicians' Misunderstanding of Normal Findings. Med Decis Making 1983; 3:169-175.
Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. Journal of General Internal Medicine 1994 (b); 9:488-495.
Poses RM, Cebul RD, Wigton RS. You can lead a horse to water - improving physicians' knowledge of probabilities may not affect their decisions. Medical Decision Making 1995; 15(1):65-75.
Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: Practicing physicians' use of quantitative measures of test accuracy. The American Journal of Medicine 1998; 104(4):374-380.

Response: We thank the reviewer for providing these references. The manuscript now includes a reference to one of these papers, Lyman & Balducchi (1994).

VERSION 2 – REVIEW

REVIEWER	Dr. Michelle McDowell Max Planck Institute for Human Development, Berlin, Germany.
REVIEW RETURNED	29-Nov-2017

GENERAL COMMENTS	The authors have addressed my main concern from my initial review regarding the use of terminology. I have a few more comments that I believe can further strengthen the manuscript. For unfamiliar readers, the authors should Include in the abstract an example of PPV or NPV or both. In the first paragraph of the Introduction, the authors state "Overestimation of the PPV, for example, increases the risk of overdiagnosis…". The references refer to risks of overtreatment. Is there any support for the statement that overestimation of the PPV increases the risk of overdiagnosis? The first sentence on page 5 states that "an experience format increased respondents' understanding of patients' knowledge of the risks and benefits…" To clarify, does it increase respondents' understanding of the risks and benefits, increase patients' knowledge of the risks and benefits, or is the interpretation as it is written? The last paragraph of page 4 and the first of page 5 could be combined, especially if reference 23 and 24 are part of the "series of

	studies". Perhaps the statement "there is significant evidence" is a bit too strong terminology given that only a few studies have been reviewed (first sentence, last paragraph of introduction). One argument that is not so clear from the Introduction is why the authors anticipate the experience advantage would (or would not) extend to clinicians? The study uses hypothetical scenarios, so it cannot be that clinicians could have pre-existing ideas of PPV or NPV for tests that could influence responses irrespective of format. Have other studies shown a difference between format effects for experts versus non-experts? To facilitate understanding of the experimental conditions, Figure 1 could also include the text of one of the problems as presented in the numerical condition rather than having the text in an Appendix. I think this would be important for understanding the type of problem presented in the numerical condition. Could the authors include a definition of sensitivity and specificity in the Introduction. Could the authors provide a little more description of the counterbalancing. Each condition (numeric vs. experience x test type) were randomly presented? So participants did not first do the three numeric followed by the three experience tasks (or in a counterbalanced order) but could do any of the 6 problems in any order? I still think that there is not enough description of the outcome metric used (MAE). The other reviewers also question the use of this metric over others (e.g., percentage correct estimate) and I am not sure this is adequately addressed. Figure 2 now presents the percentage correct PPV and NPV values whereas the results present the MAE. Could the authors discuss both, and justify more clearly why they employed MAE and emphasise that lower values are better for this metric. One final question related to the results. What kinds of errors did clinicians make in each format? This could be an interesting additional analysis (e.g., that experience reduced hit rate errors or base rate errors, etc.).

REVIEWER	Penny Whiting Bristol Medical School, University of Bristol, England
REVIEW RETURNED	23-Nov-2017

GENERAL COMMENTS	The substantial changes made to this paper by the authors have greatly improved it. I have no further comments.

REVIEWER	Clare Davenport University of Birmingham UK
REVIEW RETURNED	29-Nov-2017

GENERAL COMMENTS	Overall I would like to congratulate the authors on extensive

	revisions that I think have improved the manuscript. Major: 1) Although the authors do provide direction for future research I would like to have seen some discussion of study limitations. In particular: -Only one numeric format was compared to the experiential presentation. Future studies could use different numeric presentation formats to see if the effect is maintained, particularly numeric formats that have been shown to facilitate understanding (eg natural frequencies). -I would suggest clinical experience is a key potential variable rather than setting, particularly if fictitious examples are used. -Aspects of the presentation of the numeric format may have exacerbated problems in understanding. I personally find the numeric presentation confusing and I think it could have been more simply presented (particularly the use of decimal places when describing people) without compromising the key features of the numeric presentation for comparison. For example "If a person has Disease X, it is not certain whether he or she will have a positive result on the genetic test. More precisely, only 83.33 of every 100 such people will have a positive result on the genetic test." as: Of 100 individuals with Disease X, eighty four will test positive OR Of every 100 individuals with Disease X eighty four will test +ve with the genetic test 2) Appendix: Although the authors state that participants were requested to present their estimates in a frequency format the appendix does not include the question posed to participants. For example for PPV were participants asked to calculate the PPV / positive predictive value or to calculate the probability of testing positive if they had disease? Minor: Below are specific further minor comments that I consider would improve readability of the manuscript further. Abstract/Results: "Estimation accuracy was quantified by the mean absolute error (MAE; absolute difference between estimate and true predictive value). PPV estimation errors were higher in the numerical format…" The term 'higher' here could be misinterpreted for 'more positive'. Would 'larger' be clearer? Introduction: The authors refer to 'numerical formats' as if they are homogeneous which simplifies important known differences in the way numerical information is provided and interpreted (such as the difference between normalised and natural frequencies). This may result in erroneous conclusions being drawn by readers. The last paragraph includes the following: "In summary, there is significant evidence suggesting an advantage of experience over numerical formats in the context of diagnostic inference. However, all studies to date have been conducted with medical non-clinicians"

	When you refer to 'significant' are you referring to statistical significant results? If not I would avoid this term. I am not sure what is meant by 'medical non-clinicians'. Do you mean individuals with medical knowledge but not qualified Doctors? When you refer to ''all studies to date have been conducted with medical non-clinicians" would it be more appropriate to say 'we are not aware of any studies conducted with clinicians'? Results/Discussion: Generally a statement clarifying the direction and magnitude of estimation of error would be useful here (even if it is conveyed in the figures (I could not access the figures)); in particular the direction of error of estimation for NPV could be clearer (I think it was underestimated therefore lower than the actual value). An emphasis was placed on overestimation of the PPV (as opposed to error of estimation of NPV) the rationale for which I didn't understand. Although the experience format mimics presentation of accuracy in terms of naturally occurring frequencies the extent to which this is discussed could be misleading in the sense it suggests natural frequencies were presented.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Dr. Michelle McDowell
Institution and Country: Max Planck Institute for Human Development, Berlin, Germany.
Please state any competing interests: None declared.

Please leave your comments for the authors below
The authors have addressed my main concern from my initial review regarding the use of terminology. I have a few more comments that I believe can further strengthen the manuscript.

1. For unfamiliar readers, the authors should include in the abstract an example of PPV or NPV or both.

Response: We have added definitions of the PPV and NPV in the abstract.

2. In the first paragraph of the Introduction, the authors state "Overestimation of the PPV, for example, increases the risk of overdiagnosis…". The references refer to risks of overtreatment. Is there any support for the statement that overestimation of the PPV increases the risk of overdiagnosis?

Response: We thank the reviewer for this comment. We now more accurately state "Overestimation of the PPV, for example, may increase the risk of overtreatment such as unnecessary surgery or chemotherapy" in the introduction (see page 5).

3. The first sentence on page 6 states that "an experience format increased respondents' understanding of patients' knowledge of the risks and benefits…" To clarify, does it increase respondents' understanding of the risks and benefits, increase patients' knowledge of the risks and benefits, or is the interpretation as it is written?

Response: We apologize for this error which arose during the revision. We now say “In another study, an experience format increased patients’ knowledge of the risks and benefits of lung cancer screening.”

4. The last paragraph of page 4 and the first of page 5 could be combined, especially if reference 23 and 24 are part of the “series of studies”.

Response: We have combined these two paragraphs as suggested.

5. Perhaps the statement “there is significant evidence” is a bit too strong terminology given that only a few studies have been reviewed (first sentence, last paragraph of introduction).

Response: We have replaced the term “significant” with the word “strong”.

6. One argument that is not so clear from the Introduction is why the authors anticipate the experience advantage would (or would not) extend to clinicians? The study uses hypothetical scenarios, so it cannot be that clinicians could have pre-existing ideas of PPV or NPV for tests that could influence responses irrespective of format. Have other studies shown a difference between format effects for experts versus non-experts?

Response: We are not aware of studies showing a difference in format effects for experts and non-experts, but given the evidence of expertise effects on skilled performance and strategy selection (see work by Ericsson, Charness, Newell, and others), it was reasonable to question whether the format effect would generalize to an expert population. While clinicians had no prior knowledge of the specific scenarios used in this study, they did have training and expertise in clinical diagnosis. The fictitious scenarios included tests that were representative of those used in clinical practice (i.e., gold standard test with high sensitivity and high specificity), as well as tests that were less representative (i.e., low sensitivity and low specificity). If clinicians’ pre-existing knowledge of typical PPVs and NPVs biased their judgments, one may have expected to see differences in accuracy between tests; however, none emerged.

7. To facilitate understanding of the experimental conditions, Figure 1 could also include the text of one of the problems as presented in the numerical condition rather than having the text in an Appendix. I think this would be important for understanding the type of problem presented in the numerical condition.

Response: Thank you for this suggestion. We have revised Figure 1 and updated the caption to include both an image and description of the numerical and the experience format.

8. Could the authors include a definition of sensitivity and specificity in the Introduction?

Response: We have added the definition of test sensitivity and specificity to the introduction (p. 5).

9. Could the authors provide a little more description of the counterbalancing. Each condition (numeric vs. experience x test type) were randomly presented? So participants did not first do the three numeric followed by the three experience tasks (or in a counterbalanced order) but could do any of the 6 problems in any order?

Response: We now describe the counterbalancing scheme in more detail (pp. 7-8).

10. I still think that there is not enough description of the outcome metric used (MAE). The other reviewers also question the use of this metric over others (e.g., percentage correct estimate) and I am

not sure this is adequately addressed. Figure 2 now presents the percentage correct PPV and NPV values whereas the results present the MAE. Could the authors discuss both, and justify more clearly why they employed MAE and emphasize that lower values are better for this metric.

Response: We now describe the MAE metric, its properties, and the rationale for its use in more detail (pp. 8-9). We also say explicitly that Figure 2 does not show the MAE (which is reported in the body of the Results section) but the PPV and NPV estimates, as well as the true PPV and NPV, for each test (p. 8). The figure thus provides complementary information that is not already included in the results section. (Please note that Figure 2 does not show the percentage correct as the reviewer suggests.)

11. One final question related to the results. What kinds of errors did clinicians make in each format? This could be an interesting additional analysis (e.g., that experience reduced hit rate errors or base rate errors, etc.).

Response: Prior research shows that experts and laypeople confuse the sensitivity with the PPV (e.g., Kramer & Gigerenzer, 2005) and neglect the base rate (e.g., Bar-Hillel, 1980). Unfortunately, the current study was not designed to assess the source of errors. However, because the sensitivity of the three tests varied widely in the current study (from 50% to 100%), yet no statistical difference in PPV estimation error surfaced across tests, we can infer that sensitivity (i.e., the hit rate of the test) did not significantly affect the accuracy of PPV estimates. While outside the scope of the current paper, further research into format effects on systematic estimation errors is an interesting topic for future research.

Reviewer: 2
Reviewer Name: Penny Whiting
Institution and Country: Bristol Medical School, University of Bristol, England
Please state any competing interests: None

Please leave your comments for the authors below
The substantial changes made to this paper by the authors have greatly improved it. I have no further comments.

Reviewer: 3
Reviewer Name: Clare Davenport
Institution and Country: University of Birmingham, UK
Please state any competing interests: None declared

Please leave your comments for the authors below
Please see attached file
Overall I would like to congratulate the authors on extensive revisions that I think have improved the manuscript.
Major:
Although the authors do provide direction for future research I would like to have seen some discussion of study limitations. In particular:
1. Only one numeric format was compared to the experiential presentation. Future studies could use different numeric presentation formats to see if the effect is maintained, particularly numeric formats that have been shown to facilitate understanding (eg natural frequencies).

Response: We agree, and we now state this in the last paragraph of the discussion (pp. 10-11): ”In particular, it would be important to contrast the experience format with a numerical format in which decision-relevant information is presented in natural, rather than in normalized, frequencies.”

2. I would suggest clinical experience is a key potential variable rather than setting, particularly if fictitious examples are used.

Response: We examined whether clinical experience made a difference in estimation accuracy across formats. The results showed no difference between residents' or practicing clinicians' experience (see the last sentence of the results section, p. 9).

3. Aspects of the presentation of the numeric format may have exacerbated problems in understanding. I personally find the numeric presentation confusing and I think it could have been more simply presented (particularly the use of decimal places when describing people) without compromising the key features of the numeric presentation for comparison. For example
"If a person has Disease X, it is not certain whether he or she will have a positive result on the genetic test. More precisely, only 83.33 of every 100 such people will have a positive result on the genetic test."
as:
Of 100 individuals with Disease X, eighty four will test positive OR Of every 100 individuals with Disease X eighty four will test +ve with the genetic test

Response: We agree that the decimal places may have contributed to the difficulty of the numerical format. However, the format was designed to be as similar as possible to previous studies (e.g., Galesic, Gigerenzer, & Straubinger, 2009; Gigerenzer et al. 2011) to ensure that the results could be compared to those in the published literature.

4. Appendix: Although the authors state that participants were requested to present their estimates in a frequency format the appendix does not include the question posed to participants. For example for PPV were participants asked to calculate the PPV / positive predictive value or to calculate the probability of testing positive if they had disease?

Response: Thank you for this suggestion. We have now included the PPV and NPV questions posed to participants in the method section (see page 8).

Minor:
Below are specific further minor comments that I consider would improve readability of the manuscript further.
Abstract/Results:
"Estimation accuracy was quantified by the mean absolute error (MAE; absolute difference between estimate and true predictive value). PPV estimation errors were higher in the numerical format…"
5. The term 'higher' here could be misinterpreted for 'more positive'. Would 'larger' be clearer?

Response: We have replaced the term "higher" with "larger" in both the abstract and results section.

Introduction:
6. The authors refer to 'numerical formats' as if they are homogeneous which simplifies important known differences in the way numerical information is provided and interpreted (such as the difference between normalised and natural frequencies). This may result in erroneous conclusions being drawn by readers.

Response: The objective of the current study was to examine whether a number-free format would be superior to the previously used numerical formats. As such, we felt justified in grouping together these formats, even though we agree that there are important differences among them. We acknowledge the importance of normalized vs. natural frequencies in the discussion (pp. 10-11).

7. The last paragraph includes the following:
“In summary, there is significant evidence suggesting an advantage of experience over numerical formats in the context of diagnostic inference. However, all studies to date have been conducted with medical non-clinicians”

When you refer to ‘significant’ are you referring to statistical significant results? If not I would avoid this term. I am not sure what is meant by ‘medical non-clinicians’. Do you mean individuals with medical knowledge but not qualified Doctors? When you refer to ‘’all studies to date have been conducted with medical non-clinicians” would it be more appropriate to say ‘we are not aware of any studies conducted with clinicians’?

Response: We have now replaced the term “significant” with the word “strong” (p. 6). We have also changed the sentence “However, all studies to date have been conducted with medical non-clinicians” to “However, no study to date has tested this effect in clinicians.”

Results/Discussion:
8. Generally a statement clarifying the direction and magnitude of estimation of error would be useful here (even if it is conveyed in the figures (I could not access the figures)); in particular the direction of error of estimation for NPV could be clearer (I think it was underestimated therefore lower than the actual value).

Response: Page 7 of the results section states the direction of estimation error for both the PPV and NPV.
PPV: “…the extent to which PPVs were overestimated was reduced dramatically when information was experienced.”
NPV: “…with less underestimation and reduced variability in estimates when information was experienced.”
The magnitude of the estimation error is represented by the effect size, with a medium format effect size for PPV estimation errors (i.e., d=0.697), and a small effect size for NPV estimation errors (i.e., d=0.303), as stated in the abstract and results section.

9. An emphasis was placed on overestimation of the PPV (as opposed to error of estimation of NPV) the rationale for which I didn’t understand.

Response: Overestimating the PPV is a common error made in medicine (Gigerenzer et al., 2010; Wegwarth & Gigerenzer, 2011) that has important implications (i.e., overtreatment). The emphasis placed on the PPV is due to the “immediacy” of the situation compared to the NPV. For example, the reliability of a diagnostic test that produces a positive test result may be more important “in the moment” compared to a test that produces a negative test result. We included both the PPV and NPV because they both require similar forms of probabilistic inference (e.g., Bayesian inference), however the PPV was emphasized because prior research has largely documented PPV estimation errors, likely due to real-world implications of erroneous probabilistic judgments in medicine.

10. Although the experience format mimics presentation of accuracy in terms of naturally occurring frequencies the extent to which this is discussed could be misleading in the sense it suggests natural frequencies were presented.

Response: We thank the reviewer for this comment. We have clarified that participants were not presented with natural frequencies, but were able to derive natural frequencies from the “naturally occurring frequencies” with which they were presented in the experience format (p. 10). This is also stated in the introduction (p. 5) where we describe the experience format.

VERSION 3 – REVIEW

REVIEWER	Clare Davenport University of Birmingham UK
REVIEW RETURNED	05-Dec-2017

GENERAL COMMENTS	I have commented extensively on 2 occasions and I consider the authors have now addressed concerns raised by all peer reviewers adequately.