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)	Biomedical Authors' awareness of publication ethics: An international survey
AUTHORS	Schroter, Sara; Roberts, Jason; Loder, Elizabeth; Penzien, Donald; Mahadeo, Sarah; Houle, Timothy

VERSION 1 – REVIEW

REVIEWER	Joeri Tijdink VU University Amsterdam, the Netherlands
REVIEW RETURNED	29-Jan-2018

GENERAL COMMENTS	I have read this manuscript with great interest and I find the aim, methods and results of the study very interesting. Publication ethics is an understudied subject, as the authors already commented in their introduction. My overall impression is that the study is well written, complete and thorough and guides you with care through the methods and results. They managed to include a high number of respondents for a survey that may be perceived as long. (did your survey program measured the amount of time per respondent to complete the survey?). Furthermore, the outcomes are relevant for readers inside and outside the EU (I still consider the UK as part of the EU), although differences between countries are omnipresent. Below you will find some remarks and suggestions that may improve the manuscript. Introduction: The introduction is all about publication ethics, but there is no formal or informal description of publication ethics. What is it? What does it consist of? What are the ethical aspects covered by publication ethics? Methods: 1. How come the study was conducted in 2010 but only until now submitted to a journal? 2. The incentive to participate and enter into a prize draw to win a donation is very elegant! 3. Was there a formal informed consent procedure? And was there a privacy policy that participants could read prior to participation? 4. Is there a study protocol? 5. Did you preregister your study? 6. Is there a data analysis plan? Now it looks like the survey is exploratory by nature. This should be underlined more in the abstract, the methods and results section. It is exploratory and cross-sectional, thus the results should be interpreted with caution and no causal relations should be concluded.

	7. Did you send out a non-response questionnaire with surveying reasons for non-response? 8. What was the median time of the survey, as in the pilot phase you tested the survey and concluded that it was too long? What did you omit? 9. The vignettes should be presented in the methods sections, as readers need some help in understanding the vignette-section of the survey. Now, it is a little struggle to follow this section of the methods. 10. Is there an appendix that presents the survey questionnaire that makes it easier to comprehend the survey (ie a pdf?) 11. What survey program did you use? Results: 1. Page 8, line 28: it should probagly be 75th instead of 7th. ? 2. Now the results section starts with the training in PE and Perceived knowledge. I wonder why you chose to start with these results, and not start with the vignettes as they came first in the survey and are the primary outcomes, right? Or is the level of education the pirmary outcome. Please clarify this in the abstract and method sections as well.. 3. The vignettes section of the results section is perceived long. 4. It might be self-explanatory that some text in the results section can be deleted. If it is already in the Tables (ie table 2 and table 3) you may consider it deleting it from the text, unless it is an essential ans meaningfull result. 5. Please elaborate a little more on page 9, r 19. What does the phi = 0.45 score say? 6. The experimental manipulations in the vignettes did not influence much, but due to a high number of responses was it significant in all cases. Is it necessary to mention this for every variable? 7. Since the study is exploratory by nature and correlations will be reported, it may be improve the readability if only the highlights will be presented in the vignette section. Discussion: 1. Can you tell something more about the influence on study results that you only included BMJ publishing Group Journals and their authors. Are UK scientists better trained? 2. Might there be an UK-effect, as over 600 respondents came from the UK, and had the UK-respondents higher response rates? Other country related influences? 3. Again, I struggle with the definition of publication ethics. When is it publication ethics and when is it misbehaviors? 4. The problem with using vignettes in a study is the multi-interpretability of such vignettes. This may be better addressed in the discussion section 5. P 12, r. 22-26. Great paragraph, I think it is a very important remark to make that everyone has their own beliefs and experiences to determine something is unethical… 6. Is it possible to elaborate a little more on a possible response bias? Were the persons that did respond comparable with the complete sample? 7. What was the average time to completion? Did a lot of respondents stopped halfway the survey? Iow, how come approx. 25% of respondents that visited the website did not complete the survey? How may have influenced this the results? 8. Could it be that, according to a recent review on education/interventions of Research integrity, ineffective education

	is a possible strong determinant? See: http://www.cochrane.org/MR000038/METHOD_preventing-misconduct-and-promoting-integrity-research-and-publication 9. It could be helpful to end with a conclusion and elaborate more on other aspects of publication culture should be improved (ie the reward system, the success of getting funded) or what other interventions may work. Furthermore, it could be helpful if the authors zoom in a little more on differences in countries. And what can journals do to improve this? I am really looking forward to hear your thoughts and opinions on this, as you are part of the publishing system and will have informed thoughts on improving these ethics... Competing interests 1. I am not sure how to address this.. it may be that authors that work for BMJ and submit their paper to the BMJ may have a conflict of interest. Should this be clarified a little more?

REVIEWER	Jigisha Patel Springer Nature, UK
REVIEW RETURNED	08-Feb-2018

GENERAL COMMENTS	1. The abstract is misleading. It states that 38% of researchers responded, but in the manuscript in the results section it states that 38% visited the website, while in the discussion it states that, 'The response rate of 31% is low.' From the figures provided in the results section, 29% (3090/10582) completed the whole survey. The authors should be more precise about what the percentages mean. The authors should state that 28% completed the whole survey in the abstract and also include this figure in the strengths and limitations box. 2. It would be useful authors to explain their rationale for choosing the topics of the 5 vignettes in the survey. The study does not cover any aspects of research ethics, such as researchers' understanding/misunderstanding of consent to publish identifiable details, issues related to the conduct of unethical research, written v verbal consent etc. In my experience, these are very common in medical journals and this study was a good opportunity to explore attitudes and beliefs around these issues. Was there an active decision not to explore these? Discretionary 3. It would make the manuscript more useful to readers if the BMJ editors gave their own opinions on each vignette. Assuming all journals have the same core editorial policies, can the authors comment on how each vignette, with each contextual variation, would be viewed from the editorial perspective?

REVIEWER	Melissa S. Anderson, Professor University of Minnesota, USA
REVIEW RETURNED	13-Feb-2018

GENERAL COMMENTS	Review Biomedical authors' awareness of publication ethics: an international survey The paper presents interesting information on researchers' training in publication ethics, knowledge of publication ethics, and attitudes about certain publication-related behaviors. Note: the authors refer to "beliefs" about behaviors in their abstract and their statement of purpose (p. 4, bottom), but there are no beliefs represented in the results. The authors claim that opportunities for biomedical researchers to learn about ethical issues related to authors are "uncommon" (p.4) but such training is mandatory in the U.S. Researchers from the U.S. account for over 10% of the respondents (Figure 2 and Table 3). The authors note that they use medians to report interval data (p.6). In fact, they report results from ratio data. The authors do not present the actual survey response rate in the text when they are discussing survey completion statistics (p.8), but they do note response rates of 33.7% and 34.5% for the (presumably exhaustive) categories of those with triaged articles and those with reviewed articles in the journal in question, implying an overall rate between those percentages. On p.13, however, they cite a response rate of 31%. Figure 1 suggests a response rate of 3,090/10,582 = 29% --- or possibly 4,043/10.582 = 38%. I recommend that the authors specify and justify their response rate. It is not clear how the items in Table 1 were coded. The modal response for years of research experience was "9 to 15 years" (p.8), suggesting that 9 to 15 was a response category. Table 1, however, has categories 6 to 10 and 11 to 15. Likewise, the modal response for peer reviews was "2 to 4" (p.8), but Table 1 displays a median, suggesting that responses were by number, not by category. I don't see a need for "modal" figures. Table 1 also displays numbers and percentages of missing values, which do not vary a great deal by variable. The problem is that the valid percentages are not adjusted for missing values. Thus, for

	example, the authors note that 50.2% of the respondents are male and 29.7% of the respondents are female. The problem is compounded in the text, where the authors note that 30% of the respondents are female, leaving the reader to assume that 70% are male. Similar problems are associated with other results that are not adjusted for missing values. The percentages should be reported as valid percentages of those who responded. The missing values need not be reported in Table 1, though the authors could choose to indicate, perhaps in a footnote, the range of percentages missing. The authors claim that "Previous training was positively associated with perceived knowledge scores, indicating that individuals with higher levels of previous training endorsed ..." The variable names "previous training" and "level of training" do not match the actual measure, which is based on the respondent's own perception of the quality of training received (Table 2). Perception of quality of training received would be a more accurate variable name. The authors created a measure of this variable as "each respondent's highest rating from any of their previous training scores" (p.9). This measure needs justification. It would seem that a mean of the four training-quality scores would be a better measure. It would distinguish between (hypothetical) scores 4-4-4-4 and scores 4-0-0-0, which the "highest rating" measure does not. The authors note that "The majority of participants reported that they had 'some knowledge' of most issues (37.2% to 59.9%) ..." (p.9). A percentage below 50% does not indicate a majority. Perhaps the authors mean to indicate "at least some knowledge," which would be accurate, though the parenthetical numbers would need to be corrected. The results related to the vignettes all note the percentage of variation unaccounted for, which is redundant as they also give the percentage of variation accounted for. The results all note that the experience of the researcher did not influence responses and that there were no higher-order interactions, which could be simply noted for all cases at the beginning of this section. On this point, however, see the last comment below. The second column heading in Table 1 is ambiguous. People can "visit" a survey site without answering any questions at all. The

	people represented in this column must have actually responded to the items noted here. The authors do not present the actual questions or the actual response categories for the variables represented in Tables 2 and 3, though we can infer (perhaps incorrectly) the response categories from the tables. The authors do not indicate what instructions they gave with regard to the vignette items, though they give some indication of the question asked and the response categories, which ranged from 0 to 10 (p.6). Without labels on these values, respondents' interpretations of, say, a score of "5" versus a score of "7" may have varied, which should be noted. Figure 1 is a useful rubric for the authors' use, but it is unnecessary in the paper. A simple summary (such as a more precise paragraph on p.8) would suffice. Figure 2 is likewise unnecessary. It presents imprecise information that could be presented in a simple four-line table as: Countries with over 350 respondents: UK; 200-250 respondents: US, Italy, Netherlands, 100-200 (etc.), 30-100 (etc.). The images in Figure 3 are unnecessarily complex. The authors write, "Figure 3 displays the unethical ratings for each vignette as a function on the experimental manipulations" (p.9). Figure 3 is therefore just a series of frequency distributions. The visual style suggests a continuous measure, whereas the actual discrete ratings are indicated (0 to 10). The drawings require the reader to puzzle them out and, even so, provide only comparative indications of the distributions, not actual numbers. All of Figure 3 could be represented in a single table showing the mean, standard deviation and range of each of the 20 distributions, which would actually present more information than the figures do. More generally, it is not clear why the authors have not chosen to do a multivariate analysis (e.g. regression) relating vignette scores to respondents' perceptions of quality of training, perceived knowledge of ethics, the demographic and experience variables. The only linear models used have "perceived 'unethicalness' as the outcome variable and randomised condition as the predictors" (p.6). That means, for example, that a respondent's ethical estimation of author omission is analyzed as a function of level of

	acknowledgement (Box 1 and p.10), but the respondent is actually assigning a specific ethical estimation to *author omission with a specific level of acknowledgement*, not just to author omission. That is, the dependent variable is inseparable from each of the predictors. If this point is correct, it suggests that the results of the linear models should be deleted. If incorrect, it suggests that the method needs to be explained and justified more carefully.

REVIEWER	Mark Bahr Bond University, Australia
REVIEW RETURNED	09-Mar-2018

GENERAL COMMENTS	I have somewhat mixed feelings in regards to the analysis presented here. The study itself is admirable. It is undoubtedly a large study for this type of research and shows in places a sensitive understanding of data analysis with in other places rather odd handling odd some issues. The initial concern I had was the obvious potential for conflict of interest with authors professional roles. This is acknowledged at the end of the paper perhaps an earlier acknowledgment would be less distracting. The findings in regard to explicit training are useful in a broad sense. The finding that people disagree on what constitutes a breach is not especially remarkable. Even amongst the highly trained there is considerable disagreement. I suspect the vignettes do not provide sufficient context to allow proper consideration and this contributes somewhat to differences. Large samples are wonderful but with them comes the problem of overpowered research where small differences are seen as significant. There is something of a sense of that throughout. Perhaps more problematic are two other aspects. The reporting of order effects is uninformative effectively the message is communicated that there are in some cases significant order effects which raise doubt as to the interpretation of the provided results without adequate discussion of those order effects. on occasions magnitudes of difference of non-significant findings are reported when these lack any real meaning. Yet more important details such as the IV's involved in interactions (albeit non-significant) are not identified. Consequently the results in places lack adequate clarity. The approach of treating some data as interval and other as categorical is in a sense understandable but contributes to lack of clarity or at least consistency. Given the large n it would be acceptable to reduce to the less powerful technique. The plots are interesting but I wonder if Ci's in the form of white error bars could be plotted over the centroids to provide better side by side comparison.

REVIEWER	Mario Malicki University of Amsterdam,Netherlands
REVIEW RETURNED	04-Apr-2018

GENERAL COMMENTS	Dear Authors, Thank you for the opportunity to review this manuscript. I enjoyed reading it, and I would like to offer my suggestions for its improvement: 1) In the introduction you mention topics related to authorship, so would advise citing a sys. review on authorship, as well as ghost writing:

	A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. Doi: 10.1371/journal.pone.0023477 Systematic review on the primary and secondary reporting of the prevalence of ghostwriting in the medical literature 10.1136/bmjopen-2013-004777 2) Methods – please list the exact dates the surveys were sent, and if you are making the invite email and the full survey available as appendix or depositing them somewhere. 3) ethics of undertaking and publishing s. reserach – were these separated question, as many medical schools may have ethics, but will in much smaller extend deal with publication ethics? From the tables, it seems this was one question, so maybe mention this fact in limitations. 4) In methods you state - Perceived knowledge scores were transformed into a T score, please explain in detail, does this mean that (no knowledge=0, some knowledge=1, substantial knowledge=2) and the sum were then transformed into t, or was it 1, 2, 3 coding? How is it you have a mean of 10 and SD of 10, can the score be negative? 5) In methods you say you compared respondents to non-respondents based on acceptance or rejection of their articles, but you compared to triaged or reviewed – plz update, and list the test used for the comparisons, with p values for all in the results in the first paragraph, or state data not shown. 6) For correlation of training and knowledge scores, I would advise you check for differences between those that have received no training and those have received any kind of training, instead of using the top score for one of the categories – or maybe, no training, mentor training, any type of training – and compare both total score, and individual categories, as some question i.e. PP and AO have a large percentage of those declaring no knowledge. 7) Another alternative for that would be instead of what you mentioned: “To estimate this association, we coded each respondent’s highest rating from any of their previous training sources, and estimated an association with their perceived knowledge total score.” You could calculate the total score of the training, by assigning for each type of training a score / no training mentor 0, poor 1, average 2, and then the same for o training course 0, poor 1…etc. and sum those all up and correlate that score with the total knowledge score. 8) Have you done a regression analysis on T scores based on participants characteristics? Please report the overall scores for participants, and the possible differences based on the received training, gender and other characteristics. 9) In the results section I feel that you are missing a correlation between the scores and vignettes, and the regression analysis of the vignettes, based on the respondents characteristics and total scores, not just the variations of the vignette. Additionally, as the knowledge questions were for 7 topics, and there were 5 vignettes, perhaps the knowledge of a specific event in question is more related then the total score. Tables and figures:- 1) Figure 2 – advise putting actual numbers on top of each column 2) Figure 3 legends – omit :- sign for b to d, and explain does the width of the figure present the percentage of authors choosing that option

	In hopes may comments can help you improve your manuscript,

VERSION 1 – AUTHOR RESPONSE

Comment	Response	Description of the location and wording of all revisions that have been made (clean version)
Editorial comments		
- Please include a copy of the STROBE checklist as a supplementary file, completed with page numbers indicating where each of these items can be found in your manuscript.	We respectfully point out that the STROBE checklist for observational studies does not apply to survey studies. As far as we are aware, there is no relevant reporting checklist for survey studies.	No change.
- Please provide another copy of your figures with better qualities and please ensure that figures are of better quality or not pixelated when zoom in. NOTE: They can be in TIFF or JPG format and make sure that they have a resolution of at least 300 dpi. Figures in PDF, DOCUMENT, EXCEL and POWER POINT format are not acceptable.	We have made several revisions to the figures in response to reviewer suggestions and have uploaded enhanced resolution figures (300 DPI).	We have uploaded revised figures in this submission. Please note that Figure 2 has 5 panels (2a to e) but we have submitted it as 5 separate files in high resolution.
Reviewer 1		
They managed to include a high number of respondents for a survey that may be perceived as long. (did your survey program measured the amount of time per respondent to complete the survey?).	We agree that the survey might have been perceived as long. Indeed, we did measure the time spent on the survey and we now present these data in the revised manuscript.	We have added the following sentence to the Methods section under survey instrument: "We recorded the elapsed time completing the survey and present this data using median [25th, 75th]"

The introduction is all about publication ethics, but there is no formal or informal description of publication ethics. What is it? What does it consist of? What are the ethical aspects covered by publication ethics?	We agree it is useful to provide a description of publication ethics.	We have added the following sentences to the introduction: "We define publication ethics as professional conduct that, in the words of COPE, "reflect[s] the current best principles of transparency and integrity." We chose to focus on some of the topics emphasised by COPE in its educational activities for authors and editors."
Methods: 1. How come the study was conducted in 2010 but only until now submitted to a journal?	Several key researchers experienced life transitions during the course of this project, including professional relocation and maternity leaves. These delayed completion of the paper.	No change.
2. The incentive to participate and enter into a prize draw to win a donation is very elegant!	Thank you.	No change.
3. Was there a formal informed consent procedure? And was there a privacy policy that participants could read prior to participation?	Consent was implied by survey completion. Potential respondents were informed of the privacy standards for the research.	We have added the following sentences to the methods section of the paper: "Consent was implied by completion of the survey. Respondents were told that their responses would be treated confidentially and held on a secure server. They were also told that editors would not see named individual responses."
4. Is there a study protocol??	We did not write a formal protocol, but did pre-specify planned analyses and shared this with the whole research team.	We have indicated in the statistical analysis section that "all reported analyses were pre-specified"
5. Did you preregister your study?	No, we did not pre-register the study in 2008.	No change.
6. Is there a data analysis plan? Now it looks like the survey is exploratory by nature.	We have considered this comment carefully. With all due respect to the reviewer, we don't agree that the analysis is	We have indicated in the statistical analyses section that

This should be underlined more in the abstract, the methods and results section. It is exploratory and cross-sectional, thus the results should be interpreted with caution and no causal relations should be concluded.	exploratory, as it was a planned analysis and we experimentally tested the effects of varying the information in vignettes. As such, we think it is appropriate to look at causal relationships.	"all reported analyses were pre-specified."
7. Did you send out a non-response questionnaire with surveying reasons for non-response?	We did not send out a non-response questionnaire.	We have added the following sentence to the Procedures section of the methods section: "We did not survey non-respondents to learn their reasons for nonresponse."
8. What was the median time of the survey, as in the pilot phase you tested the survey and concluded that it was too long? What did you omit?	We reduced the number of vignettes from 8 to 5 and within each vignette we reduced the number of questions. We previously asked if the "person" in the vignette should have done something differently and if so what.	We have added the following statement to the questionnaire development and piloting section in the methods: "The questionnaire was shortened by reducing the complexity and number of vignettes based on these results."
9. The vignettes should be presented in the methods sections, as readers need some help in understanding the vignette-section of the survey. Now, it is a little struggle to follow this section of the methods.	Box 1 shows the contents of each vignette and the variables and statements that were randomised for inclusion. We do already refer to Box 1 in the methods but have moved the reference to it to earlier in the methods to make it clearer.	We have moved the reference to Box 1 earlier in the methods section. Appendix 1 now shows the questionnaire which should help readers (see Survey instrument section).
10. Is there an appendix that presents the survey questionnaire that makes it easier to comprehend the survey (ie a pdf?)	We now include the questionnaire as Appendix 1.	We now refer to Appendix 1 in the section called survey instrument within the methods section.
11. What survey program did you use?	We did not use commercial survey software such as SurveyMonkey, since we needed to randomise submitting authors to receive different subsets of vignettes in order to reduce respondent	We have added the following sentence to the methods section of the paper: "We developed customised survey software for this project so that we could randomise submitting authors to receive different

	burden. Free software such as SurveyMonkey is not sophisticated enough to do this.	presentations of the vignettes."
Results: 1. Page 8, line 28: it should probagly be 75th instead of 7th. ?	We have revised this.	We have changed this to: "Respondents had a median [25th, 75th] age of 44 [37, 52], almost half reported their main language was not English,and 30% were female and 50% male (Table 1)."
2. Now the results section starts with the training in PE and Perceived knowledge. I wonder why you chose to start with these results, and not start with the vignettes as they came first in the survey and are the primary outcomes, right? Or is the level of education the pirmary outcome. Please clarify this in the abstract and method sections as well..	We prefer to report the level of training in PE and perceived knowledge ahead of the vignettes as it helps to contextualise the results, in the same way that many put the respondent characteristics first in the results section. We agree that the flow could be improved so have moved the sections around. This is a survey so there was no primary outcome. The experimental manipulations of the vignettes were of equal interest as the level of training received in PE and the perceived knowledge of PE. No attempt was made to adjust for multiple comparisons.	We have moved the perceived knowledge section ahead of the level of training section as this flows better and renumbered the tables.. We have added the following statement to the Methods: "We did not adjust for multiple comparisons."
3. The vignettes section of the results section is perceived long.	The vignette data is quite complicated to explain so takes up words. It is not possible to put all this information in a table and just guide the reader in the main text or nuances will be omitted. We have deleted some redundancy in the text. Please see comments to Reviewer 3 as well.	To reduce words we have deleted the statement that "There were no higher-order interactions among the experimental manipulations, allowing main effects to be interpreted" from within each vignette section and report this at the top as this result was consistent across all vignettes. Within each vignette we now just report the amount of variation accounted for and no longer report the amount unaccounted for as well.

4. It might be self-explanatory that some text in the results section can be deleted. If it is already in the Tables (ie table 2 and table 3) you may consider it deleting it from the text, unless it is an essential and meaningful result.	We have carefully gone back through the results section and removed any duplication/ redundancy between the text and tables. The reviewer has not indicated specifically where there is duplication.	We have revised the presentation of the results.
5. Please elaborate a little more on page 9, r 19. What does the phi = 0.45 score say?	We have rephrased this statement so it is clearer.	We have revised this to "Perceived quality of previous training was positively associated (phi =0.45, p<0.001) with perceived knowledge scores, indicating that individuals with higher levels of perceived quality of previous training endorsed higher perceptions of knowledge about ethical issues. To estimate this association, we coded each respondent's highest perceived quality rating from any of their previous training sources, and estimated an association with their perceived knowledge total score."
6. The experimental manipulations in the vignettes did not influence much, but due to a high number of responses was it significant in all cases. Is it necessary to mention this for every variable?	For each vignette the experimental manipulations accounted for only a small amount of total variability but the amount did vary between 1.5% and 16% so we think it is important to report this for each vignette.	No change.
7. Since the study is exploratory by nature and correlations will be reported, it may be improve the readability if only the highlights will be presented in the vignette section.	Please see our response to point 6 in the methods above. We have only reported the highlights of our pre-specified analysis.	No change.
Discussion 1. Can you tell something more about the influence on study	BMJ Publishing Group journals attract submitting authors from all over the world. As such	We have indicated that respondents were working in 101 countries.

results that you only included BMJ publishing Group Journals and their authors. Are UK scientists better trained?	respondents to the survey reported they worked in 101 countries.The study was not designed to evaluate differences in levels of training between countries, and some of the country samples were very small. A subsequent paper is planned wherein we will examine country-level differences in vignette responses for those countries with sufficiently large author sample sizes.	We have added this to the respondent characteristics section of the results "Respondents reported they worked in 101 countries."
2. Might there be an UK-effect, as over 600 respondents came from the UK, and had the UK-respondents higher response rates? Other country related influences?	This is a fascinating question, but not one that we had planned to address. Response heterogeneity is an important issue but requires careful consideration of responses that are conditional on a host of factors including training, years of experience, etc. In short, this would require a multivariable model and this study was not designed for such an effort.	No change.
3. Again, I struggle with the definition of publication ethics. When is it publication ethics and when is it misbehaviors?	We agree that we could be clearer about the definition of publication ethics. We think that publication ethics involves avoiding specific misbehaviors, which we have identified at the COPE website and included in our survey.	See response to previous reviewer. We have added the following sentences to the introduction: "We define publication ethics as professional conduct that, in the words of COPE, "reflect[s] the current best principles of transparency and integrity." We chose to focus on some of the topics emphasised by COPE in its educational activities for authors and editors."
4. The problem with using vignettes in a study is the multi-interpretability of such vignettes. This may be better addressed in the discussion section	We agree that vignettes can be interpreted in many different ways and that this can be a problem. We addressed this by piloting the vignettes and revising them for clarity.	We have added a sentence to the study limitations section of the discussion: "Although we piloted and revised the vignettes based on feedback, it remains possible that respondents might not have interpreted them as intended."
5. P 12, r. 22-26. Great paragraph, I think it is a very important remark to	Thank you.	No change.

make that everyone has their own beliefs and experiences to determine something is unethical…		
6. Is it possible to elaborate a little more on a possible response bias? Were the persons that did respond comparable with the complete sample?	This is a very important question and one that we considered in pre-study planning. Unfortunately, we had to weigh the crucial need for providing respondent anonymity against collecting respondent-identifying information that could be used to address this potential bias. We decided to refrain from collecting additional data and have presented all of the information we had in the first submission. We have elaborated on this issue in the Discussion.	We have revised the Discussion to include the following statements in the strengths and limitations section: "Response bias, in any variety of forms, is always of concern in a survey study of this type. Although we could examine several obvious sources of responder bias (e.g. author experiences in submission), we took great care in blinding participant identities to best ensure anonymity, so we could not collect extensive information on non-responders for the purposes of comparison with responders."
7. What was the average time to completion? Did a lot of respondents stopped halfway the survey? low, how come approx. 25% of respondents that visited the website did not complete the survey? How may have influenced this the results?	For those who completed the entire questionnaire, the median time to complete was 8 [5, 12] minutes. 4043 completed at least some of the questionnaire and 3090 of these 4043 (76%) completed every question in the survey.3668 (91%) completed at least one vignette. We think these figures show a high level of engagement with the survey.	We have added the time taken to complete to the results section.
8. Could it be that, according to a recent review on education/interventions of Research integrity, ineffective education is a possible strong determinant?	We now mention this review in the Discussion.	We have added this to the study implications section of the Discussion: "The authors of a recent Cochrane review evaluating the effectiveness of

See: http://www.cochrane.org/MR000038/METHOD_preventing-misconduct-and-promoting-integrity-research-and-publication		educational or policy interventions addressing research integrity and responsible conduct of research concluded that the effectiveness of these interventions on reducing misconduct is uncertain owing to the very low quality of the available evidence. [23]"
9. It could be helpful to end with a conclusion and elaborate more on other aspects of publication culture should be improved (ie the reward system, the success of getting funded) or what other interventions may work. Furthermore, it could be helpful if the authors zoom in a little more on differences in countries. And what can journals do to improve this? I am really looking forward to hear your thoughts and opinions on this, as you are part of the publishing system and will have informed thoughts on improving these ethics...	We concur that it would be valuable to be able to elaborate more on other aspects of publication culture as they might relate to our topic, but unfortunately, we believe doing so would be highly speculative and would extend too far beyond the scope of the evidence we are presenting here. As stated above, this study was not designed to evaluate differences in levels of training between countries, and some of the country samples were very small. Nevertheless, a subsequent paper is planned wherein we will examine country-level differences in vignette responses for those countries with sufficiently large author sample sizes.	No change.
Competing interests 1. I am not sure how to address this.. it may be that authors that work for BMJ and submit their paper to the BMJ may have a conflict of interest. Should this be clarified a little more?	We have revised the COI statement.	We have added the following statements to the COI declaration: "None of the authors work directly for BMJ Open or are involved in the decision-making process for articles submitted to BMJ Open. This paper was sent out for peer review in the usual way and treated in the same way as all submissions to the journal."
Reviewer 2		

1. The abstract is misleading. It states that 38% of researchers responded, but in the manuscript in the results section it states that 38% visited the website, while in the discussion it states that, 'The response rate of 31% is low.' From the figures provided in the results section, 29% (3090/10582) completed the whole survey. The authors should be more precise about what the percentages mean. The authors should state that 28% completed the whole survey in the abstract and also include this figure in the strengths and limitations box.	Thank you, we agree this is confusing and have revised the text accordingly. Whilst 3090/10582 (29%) completed the entire survey, we feel that it is reasonable to report the overall response rate as 4043/10582 (38%) in the abstract as it is not typical to report a response rate for only the subset who completed every single question (including the demographics). BMJ Open asks authors not to report results in the strengths and limitations box and response rates are results.	We have revised the abstract and results to make it clearer that 38% completed at least some of the questionnaire. We have revised the Discussion to say that" the response rate of 38% is low". We have clarified the meaning of the percentages throughout the results.
2. It would be useful authors to explain their rationale for choosing the topics of the 5 vignettes in the survey. The study does not cover any aspects of research ethics, such as researchers' understanding/misunderstanding of consent to publish identifiable details, issues related to the conduct of unethical research, written v verbal consent etc. In my experience, these are very common in medical journals and this study was a good opportunity to explore attitudes and	We had hoped to include more vignettes in the study, but our piloting indicated that the respondent burden was too high with more than 5. We chose the topics for the vignettes addressing issues we most commonly encounter and which COPE emphasises in its educational activities for authors and editors. Whilst we agree with this reviewer on the importance of the issues they have raised, we simply could not include all topics in this survey. These would be a good focus for another survey.	We have added the following explanation to the last paragraph of the Introduction "We define publication ethics as professional conduct that, in the words of COPE, "reflect[s] the current best principles of transparency and integrity." We chose to focus on some of the topics emphasised by COPE in its educational activities for authors and editors. Under the questionnaire development and piloting section in the results we now indicate: "To reduce respondent burden, the questionnaire was shortened by reducing the complexity and number

beliefs around these issues. Was there an active decision not to explore these?		of vignettes based on these results."
Discretionary 3. It would make the manuscript more useful to readers if the BMJ editors gave their own opinions on each vignette. Assuming all journals have the same core editorial policies, can the authors comment on how each vignette, with each contextual variation, would be viewed from the editorial perspective?	Whilst we like this idea, adding our own opinions on the vignettes would significantly increase the length of the paper. We feel this would be best placed in a different opinion based article.	No change.
Reviewer 3		
Note: the authors refer to "beliefs" about behaviors in their abstract and their statement of purpose (p. 4, bottom), but there are no beliefs represented in the results.	We agree that it was confusing to refer to beliefs where we meant opinion and have revised this in the relevant sections.	We have revised the following statements to remove reference to beliefs: Abstract "The extent to which biomedical authors have received training in publication ethics, and their attitudes and opinions about the ethical aspects of specific behaviours, have been under-studied.". Introduction (last para) "The goal of this study was to evaluate the prevalence and quality of formal training in publication ethics among biomedical authors, and to elicit their attitudes and opinions about specific behaviours."

		Results -Perceived knowledge of publication ethics section: We have revised this to "Only 8.8% of participants indicated that they possessed "substantial knowledge" on all seven topics."
The authors claim that opportunities for biomedical researchers to learn about ethical issues related to authors are "uncommon" (p.4) but such training is mandatory in the U.S. Researchers from the U.S. account for over 10% of the respondents (Figure 2 and Table 3).	We appreciate the reviewer's point that research training is mandatory for US researchers, but in most cases this includes relatively little training about the matters covered in this survey. For example, the commonly used CITI training program used by academic institutions (including that of one of our US authors) is quite basic and is focused more on ethical matters pertaining to research conduct and has only a small amount of material devoted to authorship matters.	We have changed the wording in Introduction to say that "opportunities for biomedical researchers to learn about these ethical issues are not always available or required. If available, they often do not focus in-depth on such matters."
The authors note that they use medians to report interval data (p.6). In fact, they report results from ratio data.	We sincerely appreciate the reviewer's careful consideration of the descriptions of the type of data under consideration. We allow that reasonable people might disagree about the level of measurement so we have adjusted our statement accordingly. We do wish to retain the use of medians [25th, 75th].	We have revised the statement in the statistical analysis section to: "...medians [25th, 75th percentile] are used for data with at least ordinal properties"
The authors do not present the actual survey response rate in the text when they are discussing survey completion statistics (p.8), but they do note response rates of 33.7% and 34.5% for the (presumably exhaustive) categories of those with triaged articles and those with reviewed articles in the journal in question, implying an overall rate	We have revised the first paragraph under respondent characteristics to make the response rate clearer.	We have included the following statements: "After correcting for delivery failures, 10,582 people were sent an invitation. 4043/10582 (38%) completed at least some of the survey. Of those responding 3090 (76%) completed the entire survey, 3668 (91%) rated at least one vignette. Having an article peer reviewed (34.5%) versus not peer reviewed (33.7%)

between those percentages. On p.13, however, they cite a response rate of 31%. Figure 1 suggests a response rate of 3,090/10,582 = 29% --- or possibly 4,043/10.582 = 38%. I recommend that the authors specify and justify their response rate.		was not related to the response rate, p = 0.339. We have removed Figure 1 (flow chart) as suggested by Reviewer 3.
It is not clear how the items in Table 1 were coded. The modal response for years of research experience was "9 to 15 years" (p.8), suggesting that 9 to 15 was a response category. Table 1, however, has categories 6 to 10 and 11 to 15. Likewise, the modal response for peer reviews was "2 to 4" (p.8), but Table 1 displays a median, suggesting that responses were by number, not by category. I don't see a need for "modal" figures.	The categories in Table 1 are for reporting purposes only. We agree this was confusing and have removed the statements about the modal values.	We have revised the third paragraph of results to: “Respondents had a median [25th, 75th] age of 44 [37, 52], almost half reported their main language was not English,and 30% were female and 50% male (Table 1). Roughly 17% of the 3,222 respondents who disclosed their country of training and country of work reported that they received postgraduate education in a country that was different to their current country of work. Respondents ranged in research experience; 254 (6%) had less than 10 years of experience and 510 (13%) had over 25 years. Respondents completed a median of 5 [2, 10] peer reviews a year and had published a median of 30 [10, 70] articles in their career.”
Table 1 also displays numbers and percentages of missing values, which do not vary a great deal by variable. The problem is that the valid percentages are not adjusted for missing values. Thus, for example, the authors note that 50.2% of the	We agree that this was confusing but with a large number of missing responses we feel it is important to record this in the table. We have revised the text to make it clearer, for example, we report both the proportion of males and females (see our response above).	See above for comments on this. We have revised Table 1 to make it easier to follow..

respondents are male and 29.7% of the respondents are female. The problem is compounded in the text, where the authors note that 30% of the respondents are female, leaving the reader to assume that 70% are male. Similar problems are associated with other results that are not adjusted for missing values. The percentages should be reported as valid percentages of those who responded. The missing values need not be reported in Table 1, though the authors could choose to indicate, perhaps in a footnote, the range of percentages missing.	If we report the valid percent as suggested this will overinflate the percentages so we prefer not to do this.	
The authors claim that "Previous training was positively associated with perceived knowledge scores, indicating that individuals with higher levels of previous training endorsed ..." The variable names "previous training" and "level of training" do not match the actual measure, which is based on the respondent's own perception of the quality of training received (Table 2). Perception of quality of training received would be a more accurate variable name.	We agree this was confusing and have rephrased this.	We have revised the title of Table 2 to “Receipt of and perceived quality of ethical training (n=4043)” We have changed the text in the results section to: “Perceived quality of previous training was positively associated (phi =0.45, p<0.001) with perceived knowledge scores, indicating that individuals with higher levels of perceived quality of previous training endorsed higher perceptions of knowledge about ethical issues. To estimate this association, we coded each respondent’s highest

		perceived quality rating from any of their previous training sources, and estimated an association with their perceived knowledge total score.”
The authors created a measure of this variable as "each respondent's highest rating from any of their previous training scores" (p.9). This measure needs justification. It would seem that a mean of the four training-quality scores would be a better measure. It would distinguish between (hypothetical) scores 4-4-4-4 and scores 4-0-0-0, which the "highest rating" measure does not.	We respectfully disagree with this insightful interpretation. If each of the previous training items represented a correlated measure of a central construct, then averaging them would represent a more reliable estimate of the construct. However, in the case of previous training, each of the items is assumed independent (uncorrelated), and obtaining a score of 4 in any one of them represents a high level of the construct. We contend that 4-0-0… is indistinguishable from 4-4-4.. We now explicitly state these assumptions.	We have included the following statements in the results section under the Training in publication ethics section : “The highest score was used because it was not expected that participants would receive training from all sources and high levels of perceived quality from any single source could impact perceived knowledge”.
The authors note that "The majority of participants reported that they had 'some knowledge' of most issues (37.2% to 59.9%) ..." (p.9). A percentage below 50% does not indicate a majority. Perhaps the authors mean to indicate "at least some knowledge," which would be accurate, though the parenthetical numbers would need to be corrected.	Thank you for pointing out this mischaracterisation of the data.	We have changed this statement in the Perceived knowledge section to:: “Participants reported substantial variability in the perception of their own knowledge about seven ethical topics (Table 2). Substantial knowledge in the seven topics ranged from 21.3% for author omission to 60.5% for conflicts of interest.”
The results related to the vignettes all note the percentage of variation unaccounted for, which is redundant as they also give the percentage of variation accounted for.	We have revised the text to remove duplication.	We have removed the statements about the proportion of variance unaccounted for within each vignette.

The results all note that the experience of the researcher did not influence responses and that there were no higher-order interactions, which could be simply noted for all cases at the beginning of this section. On this point, however, see the last comment below.		Where applicable we have deleted the statements about the experience of the researcher not influencing responses from within each vignette and report it at the start of the results as a common finding for 4 of the 5 vignettes. Similarly we report the a common statement about no higher order interactions in all the vignettes. "For all except the conflict of interest vignette (p=0.006), the level of experience of the researcher described did not significantly influence responses (p>0.05)."
The second column heading in Table 1 is ambiguous. People can "visit" a survey site without answering any questions at all. The people represented in this column must have actually responded to the items noted here.	We agree this was confusing and have relabelled it and revised Table 1.	We have revised Table 1.
The authors do not present the actual questions or the actual response categories for the variables represented in Tables 2 and 3, though we can infer (perhaps incorrectly) the response categories from the tables. The authors do not indicate what instructions they gave with regard to the vignette items, though they give some indication of the question asked and the response	We have now included Appendix 1 showing the actual questions in the survey. We also now describe how perceived quality was scored as a footnote to Table 2. It is common practice to not use anchors on 10-point scales like this.	We have added the following footnote to Table 2: "*Measured on a 4-point Likert scale (0=poor quality, 1=average quality, 3=good quality, 4=excellent quality)." We have added the following to the perceived knowledge section in the methods:

categories, which ranged from 0 to 10 (p.6). Without labels on these values, respondents' interpretations of, say, a score of "5" versus a score of "7" may have varied, which should be noted.		"Respondents were given a short definition of seven ethical topics and asked to indicate their level of knowledge (0=no knowledge, 1=some knowledge, 2=substantial knowledge) of each topic:...."
Figure 1 is a useful rubric for the authors' use, but it is unnecessary in the paper. A simple summary (such as a more precise paragraph on p.8) would suffice.	We have revised the text and removed Figure 1 (flow chart).	We have removed Figure 1.
Figure 2 is likewise unnecessary. It presents imprecise information that could be presented in a simple four-line table as: Countries with over 350 respondents: UK; 200-250 respondents: US, Italy, Netherlands, 100-200 (etc.), 30-100 (etc.).	We have revised Figure 2 to include the actual number of cases as suggested by Reviewer 5. Hopefully Reviewer 3 will agree that this increases the precision of the data displayed.	We have revised Figure 2 (now labelled Figure 1).
The images in Figure 3 are unnecessarily complex. The authors write, "Figure 3 displays the unethical ratings for each vignette as a function on the experimental manipulations" (p.9). Figure 3 is therefore just a series of frequency distributions. The visual style suggests a continuous measure, whereas the actual discrete ratings are indicated (0 to 10). The drawings require the reader to puzzle them out and, even	We appreciate the reviewer's suggestion about this figure. During manuscript preparation, we toiled with several versions of this figure, trying to find the most parsimonious method to display the effect sizes observed in the manipulation. We attempted to use box plots and several discrete category plots, but we can assure the reviewer that each of these was more complex than the submitted figure. In the text, we report effect sizes with 95%CI and believe that this obviates the value of an additional table. With the reviewer's permission, we would like to retain the figure as we think it is the best of several difficult choices. We have added a more informative footnotes for	We have added a footnote to help readers interpret Figure 3 and made the labels clearer.

so, provide only comparative indications of the distributions, not actual numbers. All of Figure 3 could be represented in a single table showing the mean, standard deviation and range of each of the 20 distributions, which would actually present more information than the figures do.	the figures to help readers interpret them.	
More generally, it is not clear why the authors have not chosen to do a multivariate analysis (e.g. regression) relating vignette scores to respondents' perceptions of quality of training, perceived knowledge of ethics, the demographic and experience variables.	This is an excellent point, but proscribes analyses that are well beyond our initial planned analyses. We absolutely agree that these associations might be of interest, but contend that the planned analyses are quite complex and additional analyses should be saved for a later "exploratory" analysis.	No change.
The only linear models used have "perceived 'unethicalness' as the outcome variable and randomised condition as the predictors" (p.6). That means, for example, that a respondent's ethical estimation of author omission is analyzed as a function of level of acknowledgement (Box 1 and p.10), but the respondent is actually assigning a specific ethical estimation to author omission with a specific level of acknowledgement, not just to author omission. That is, the	Indeed, this interpretation is entirely incorrect. However, if the thoughtful reviewer was confused by our descriptions, our readers will be too. In response, we have added clarifying sentences to the statistical analysis section. We sincerely hope that this is now easier to follow.	We have added this text to the statistical analysis section: "This resulted in a fully crossed design where all combinations of conditions in Box 1 were presented across participants (each participant completed only one version of each vignette). For the model, the three between-subjects main effects…"

dependent variable is inseparable from each of the predictors. If this point is correct, it suggests that the results of the linear models should be deleted. If incorrect, it suggests that the method needs to be explained and justified more carefully.		
Reviewer 4		
The initial concern I had was the obvious potential for conflict of interest with authors professional roles. This is acknowledged at the end of the paper perhaps an earlier acknowledgment would be less distracting.	We have revised the COI statement. See comments to Reviewer 1. We have reported the COI statement where the journal requires it.	We have revised the COI statement.
The findings in regard to explicit training are useful in a broad sense. The finding that people disagree on what constitutes a breach is not especially remarkable. Even amongst the highly trained there is considerable disagreement. I suspect the vignettes do not provide sufficient context to allow proper consideration and this contributes somewhat to differences. Large samples are wonderful but with them comes the problem of overpowered research where small differences are seen as significant. There is something of a sense of that throughout.	We appreciate that there are disagreements across individuals regarding what constitutes a breach and our providing evidence to this effect may seem face valid and commonsensical. Nevertheless, we believe there is value in providing empirical evidence regarding the extent, nature, and scope of such disagreements, and we hope and anticipate that the evidence this study provides will support additional discussion of these issues. We agree with this reviewer that examination of key factors contributing to the observed differences are, in most instances, beyond the scope of this effort. The need to debate statistical vs. practical significance is endemic to the research enterprise and hardly unique to this work. Regardless, we believe this work reflects a healthy balance between empirical rigor and the pragmatic utility of its findings.	No change.
Perhaps more problematic are two other aspects. The reporting of order	Order effects: Indeed, we treated the order effects as a nuisance variable, and adjusted	We have added the following statements to the

effects is uninformative effectively the message is communicated that there are in some cases significant order effects which raise doubt as to the interpretation of the provided results without adequate discussion of those order effects. on occasions magnitudes of difference of non-significant findings are reported when these lack any real meaning. Yet more important details such as the IV's involved in interactions (albeit non-significant) are not identified. Consequently the results in places lack adequate clarity. The approach of treating some data as interval and other as categorical is in a sense understandable but contributes to lack of clarity or at least consistency. Given the large n it would be acceptable to reduce to the less powerful technique.	for them in the models. In this setting, the order effects are a function of the survey paradigm and offer little additional information concerning the constructs under study. We have expanded our description of this important issue and thank the reviewer for their excellent point. Interaction effects: We assure the reviewer that for the vignettes each of the predictors (except order of presentation) was coded as a fixed factor in the analysis. The effect sizes are reported using apparent differences in the 0-10 rating scale (i.e., the outcome), but each of the predictors in the factorial model are categorical. We have attempted to better introduce this fact in the statistical methods section. Of note is that it is notoriously difficult to interpret factorial models with three levels and this underscores the need to plot these effects in Figure 3.	statistical analysis section in the methods: Order effects: “The rank order of presentation of each vignette was adjusted as an additional covariate to control for order effects.” Interaction effects: We have added, ”For the models, the three between-subjects categorical main effects …”
The plots are interesting but I wonder if Ci's in the form of white error bars could be plotted over the centroids to provide better side by side comparison.	This is a great suggestion and we actually considered this when creating the plots. However, due to the sample size, the model-based estimates of central tendency for each of the conditions is associated with very precise CI bounds. This is one of the reasons that we opted to plot the actual data spread in addition to the point estimate. We have taken great care to report the contrasts between effects with point estimates of difference and 95%CI as these are the parameters of greatest importance.	No Change.
Reviewer 5		

1) In the introduction you mention topics related to authorship, so would advise citing a sys. review on authorship, as well as ghost writing: A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. Doi: 10.1371/journal.pone.0023477 Systematic review on the primary and secondary reporting of the prevalence of ghostwriting in the medical literature 10.1136/bmjopen-2013-004777	We have added these two references to the Introduction.	We have added these two references to the Introduction: Marušić A, Bošnjak L, Jerončić A. A Systematic Review of Research on the Meaning, Ethics and Practices of Authorship across Scholarly Disciplines. PLoS One 2011;6(9):e23477. doi: 10.1371/journal.pone.0023477. Stretton S. Systematic review on the primary and secondary reporting of the prevalence of ghostwriting in the medical literature *BMJ Open* 2014;4:e004777. doi: 10.1136/bmjopen-2013-004777
2) Methods – please list the exact dates the surveys were sent, and if you are making the invite email and the full survey available as appendix or depositing them somewhere.	The survey was live between 01 August and 30 September 2011. We have included the survey as an Appendix.	We have added the dates the survey was live to the methods section of the paper.
3) ethics of undertaking and publishing s. reserach – were these separated question, as many medical schools may have ethics, but will in much smaller extend deal with publication ethics? From the tables, it seems this was one question, so maybe mention this fact in limitations.	We asked: "How would you rate the quality of the training/guidance you have received on the ethics of publishing scientific research? (Excellent/Good/Average/Poor/I have never received this type of training/guidance)	We have revised the statement in the methods under Respondent characteristics to: "...and to rate the perceived quality of the training or guidance they had received on the ethics of publishing scientific research"
4) In methods you state - Perceived knowledge scores were transformed into a T score, please	We apologise for the confusion about these methods. We originally performed a linear transformation (i.e., did not	We have deleted the following statement in the

explain in detail, does this mean that (no knowledge=0, some knowledge=1, substantial knowledge=2) and the sum were then transformed into t, or was it 1, 2, 3 coding? How is it you have a mean of 10 and SD of 10, can the score be negative?	change the distribution of scores, just rescaled them to have a mean of 50 and SD of 10). However, there was a typo in our manuscript (mean = 10) and we do not actually use this transformation in the manuscript so now report the scores in their original scales using the values as you suggested, (no knowledge=0, some knowledge=1, substantial knowledge=2)	methods under statistical analysis: "Perceived knowledge scores were transformed into a T score (mean: 10, SD: 10)."
5) In methods you say you compared respondents to non-respondents based on acceptance or rejection of their articles, but you compared to triaged or reviewed – plz update, and list the test used for the comparisons, with p values for all in the results in the first paragraph, or state data not shown.	The comparisons were conducted using chi-squared testing. We have clarified the text to avoid confusion.	We have revised the Methods to include this statement: "We compared respondents with non-respondents by country in which they were based, the journal to which they submitted, and whether the paper they had submitted to the journal was peer reviewed or not". We have revised the results section to include the following: "Having an article peer reviewed (34.5%) versus not peer reviewed (33.7%) was not related to the response rate, p = 0.339."
6) For correlation of training and knowledge scores, I would advise you check for differences between those that have received no training and those have received any kind of training, instead of using the top score for one of the categories – or maybe, no training, mentor training, any type of training – and compare both total score, and individual categories, as some question i.e. PP and AO have a large	This is a really interesting suggestion but it does require the creation of post hoc subgroups that were not planned in the original analysis. We hope the reviewer will be amenable for us to decline this suggestion in favor of retaining our prespecified plan of analysis.	No Change.

percentage of those declaring no knowledge.		
7) Another alternative for that would be instead of what you mentioned: “To estimate this association, we coded each respondent’s highest rating from any of their previous training sources, and estimated an association with their perceived knowledge total score.” You could calculate the total score of the training, by assigning for each type of training a score / no training mentor 0, poor 1, average 2, and then the same for o training course 0, poor 1…etc. and sum those all up and correlate that score with the total knowledge score.	This is another intriguing suggestion and is similar to something suggested by Reviewer #3. If each of the previous training items represented a correlated measure of a central construct, then averaging them would represent a more reliable estimate of the construct. However, in the case of previous training, each of the items is assumed independent (uncorrelated), and obtaining a score of 4 in any one of them represents a high level of the construct. We contend that 4-0-0… is indistinguishable from 4-4-4.. We now explicitly state these assumptions in the manuscript.	We have clarified our assumptions underlying the approach to the analysis in response to this insightful comment (see Results: Training in publication ethics section).
8) Have you done a regression analysis on T scores based on participants characteristics? Please report the overall scores for participants, and the possible differences based on the received training, gender and other characteristics.	This is an excellent point, but proscribes analyses that are well beyond our initial planned analyses. We absolutely agree that these associations might be of interest, but contend that the planned analyses are quite complex, that this sample was not designed for such modeling, and thus we would like to defer these additional analyses for a later “exploratory” analysis.	No change.
9) In the results section I feel that you are missing a correlation between the scores and vignettes, and the regression analysis of the vignettes, based on the respondents characteristics and total scores, not just the variations of the vignette. Additionally, as the knowledge questions were for 7 topics, and there were 5 vignettes,	This is another excellent point, but proscribes analyses and post hoc evaluations that are well beyond our initial planned analyses. We absolutely agree that these associations might be of interest, but contend that the planned analyses are quite complex and additional analyses should be saved for a later “exploratory” analysis.	No change.

perhaps the knowledge of a specific event in question is more related then the total score.		
1) Figure 2 – advise putting actual numbers on top of each column	We have revised Figure 2 to include actual numbers.	We have revised Figure 2 (now Figure 1) as suggested.
2) Figure 3 legends – omit :- sign for b to d, and explain does the width of the figure present the percentage of authors choosing that option	We have omitted Figure 1 so Figure 3 is now Figure 2. These are two excellent suggestions and we have revised the figure caption to better describe the plots.	Figure 2: Plots are revised and figure caption has been enhanced to allow better interpretation.

VERSION 2 – REVIEW

REVIEWER	Joeri Tijdink VU University, the Netherlands
REVIEW RETURNED	26-May-2018

GENERAL COMMENTS	Dear Authors, Great to see that the paper has improved substantially. Just one minor thing. As you have stated: 'all reported analyses were pre-specified." This sounds to me that it is well possible to define this in more detail in an Appendix and publish the document with these pre-specified and planned analyses. Best Wishes

REVIEWER	Jigisha Patel Springer Nature UK
REVIEW RETURNED	22-Jun-2018

GENERAL COMMENTS	Thank you for your responses. I don't have any futher comments for the authors. Note to the editor: I do not have sufficient expertise to judge whether there is a need for a satistical review. Given that I am forced to give a 'yes' or 'no' answer in order to submit this review and have stated 'no'. However, I leave it to the editor to determine whether there is a need for a statistical review.

REVIEWER	Mark Bahr (Ph.D.) Bond University, Australia
REVIEW RETURNED	12-Jun-2018

GENERAL COMMENTS	The paper addresses an interesting issue and for the most part is clearly articulated. The reporting of statistical analysis around the vignettes could still be clearer but that may be personal preference. The violin plots are interesting but the information to ink ratio may be questionable. More importantly its not absolutely clear how the cited obtained p's from page 9 to 11 were generated it appears that these are comparisons of obtained values using perhaps a Z test with CI's reported to demonstrate the likely range of values. That's fine but the actual determination of the values is unstated. On page 6 the first part of the opening line 49 "TH.." is not required. By all means state that the analysis was blind (although it seems unlikely that is accurate in terms of being blind to the hypotheses given that TH was involved in writing according to the declaration). Who conducted the analysis is not relevant to the argument.

REVIEWER	Mario Malicki University of Amsterdam,Netherlands
REVIEW RETURNED	

GENERAL COMMENTS	28-May-2018

GENERAL COMMENTS	Dear authors, thank you for this revised version, I find it much improved, however: You mention that: All available data were used for the analysis and all reported analyses were pre-specified. - however you did not make the protocol available to the reviewers or the readers. Additionally, even thought they were pre-specified - you are missing a lot of exploratory statistics and therefore the pre-specification was maybe insufficient in quality, and reviewers including myself have suggested additional analysis you should conduct, so without this done - I feel the paper should not be published. Your sample size is large enough, for these analysis to be conducted. In the response to reviewers you mention this should be done in an exploratory paper - but I am not sure this was pre-specified in your plan before, nor why would there be a reason to salami slice this publication. Please conduct the asked analysis, and then reshape and resubmit the manuscript accordingly. STROBE can be applied to cross sectional studies, as well as CHERRIES reporting guidelines for surveys. I find the following explanation goes against your data:The highest score was used because it was not expected that participants would receive training - your data shows that multiple sources of education were given to participants, and therefore even though you did not expect this, you should adjust accordingly and do the sum or none vs any, as suggested before.

VERSION 2 – AUTHOR RESPONSE

Comment	Response	Description of the location and wording of all revisions that have been made (clean version)
Reviewer 1		
Great to see that the paper has improved substantially. Just one minor thing. As you have stated: 'all reported analyses were pre-specified." This sounds to me that it is well possible to define this in more detail in an Appendix and publish the document with these pre-specified and planned analyses.	This is an excellent point but does imply that our plan of analysis consisted of more documentation than what is presented in the actual manuscript; it does not. It is not uncommon for the statistical analysis plan to be extremely brief. We hope the reviewer will accept the written plan as submitted in the statistical methods section in the actual manuscript.	No change.
Reviewer 2		
Thank you for your responses. I don't have any further comments for the authors.	-	No change.
Reviewer 4		
The paper addresses an interesting issue and for the most part is clearly articulated. The reporting of statistical analysis around the vignettes could still be clearer but that may be personal preference. The violin plots are interesting but the information to ink ratio may be questionable. More importantly its not absolutely clear how the cited obtained p's from page 9 to 11 were generated it appears that these are comparisons of obtained values using perhaps a Z test with CI's	We thank the reviewer for additional comments about the violin plots and concerns about the reporting of the analysis. If it is agreeable to the reviewer, we would like to retain our violin plots as we think they do reflect the distribution of scores by condition along with the measures of central tendency. In regards the p-values from the contrasts reported in the text, these values were generated using the linear mixed model introduced in the statistical methods section. This approach contrasts the	We have added this to the vignettes section on p9 where the figures are described: "The p-values reported in the text below were generated using the linear mixed model described in the statistical analysis section. This approach contrasts the fixed-effects (i.e. experimental conditions) to generate point estimates of the difference between conditions, 95%CI around these differences,

reported to demonstrate the likely range of values. That's fine but the actual determination of the values is unstated.	fixed-effects (i.e., experimental conditions) to generate point estimates of the difference between conditions, 95%CI around these differences, and p-values for this contrast. We now state this more clearly in the Figure captions.	and p-values for this contrast."
On page 6 the first part of the opening line 49 "TH.." is not required. By all means state that the analysis was blind (although it seems unlikely that is accurate in terms of being blind to the hypotheses given that TH was involved in writing according to the declaration). Who conducted the analysis is not relevant to the argument.	We have revised this.	p6: "All statistical analyses were conducted blinded to the identities of the respondents."
Reviewer 5		
You mention that: All available data were used for the analysis and all reported analyses were pre-specified. - however you did not make the protocol available to the reviewers or the readers. Additionally, even thought they were pre-specified - you are missing a lot of exploratory statistics and therefore the pre-specification was maybe insufficient in quality, and reviewers including myself have suggested additional analysis you should conduct, so without this done - I feel the paper should not be published. Your sample size is large enough, for these analysis to be conducted. In the response to reviewers you	The authors sincerely appreciate the spirit from which this criticism is based. We thank the reviewer for their thoughtful consideration of our manuscript and for their efforts in improving it. However, there are always a large number of additional analyses that can be completed on any dataset. Scientists do not differ in the regard that they attempt to learn from their data but often do differ in respect to how much analysis they deem appropriate at the risk of type-I error. More importantly, there will always be subgroup analysis, effect modification, sensitivity analyses, etc that could be applied to every analysis, but if	No change.

mention this should be done in an exploratory paper - but I am not sure this was pre-specified in your plan before, nor why would there be a reason to salami slice this publication. Please conduct the asked analysis, and then reshape and resubmit the manuscript accordingly.	these analyses are undertaken AFTER examination of the data, their interpretation is fraught with difficulty. For the p-values to have proper meaning, and for the confidence intervals to have the proper coverage, the choice of test statistics (or estimation procedures) must not be conditional on the observed data. Stated differently, if certain analyses are chosen because of an observed pattern in the data, the test statistics lose their intended meaning. The renowned statistician, Andrew Gelman, has labeled this behavior “Forking Paths” and we refer the reviewer to the problem here: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf It is of note that the reviewer has made these recommendations after evaluating our data (in manuscript form), and as such cannot be thought to be making recommendations that are independent of the data. Finally, in regards to our original plan of analysis, we respectfully contend that our original hypotheses were complex, requiring focused inferences, and were not underdeveloped in regards to our objectives. We sincerely hope that the reviewer can appreciate our point of view on this issue. Furthermore, we fully intend on pursuing additional *exploratory* analyses, based on what we learned from these primary analyses. For these analyses we will model the responses using the ideas generated by	

	the reviewer, and examine the covariances between training and ratings to examine the relationships between them. We contend that these analyses can be prosecuted in the context of a second paper where the reader is made aware that the statistical inferences were designed post hoc and that they should be interpreted with extreme caution.	
STROBE can be applied to cross sectional studies, as well as CHERRIES reporting guidelines for surveys.	We have completed CHERRIES and uploaded it with this resubmission.	No change. See supplementary file.
I find the following explanation goes against your data:The highest score was used because it was not expected that participants would receive training - your data shows that multiple sources of education were given to participants, and therefore even though you did not expect this, you should adjust accordingly and do the sum or none vs any, as suggested before.	If we understand this comment correctly, we believe that we have not adequately communicated to the reviewer what we expected in the study planning. We did in fact anticipate that respondents would have received training from multiple sources. However, we anticipated that their perception of this training would be best characterized by the most esteemed source of this training (i.e., if two sources of training we not equally beneficial, the respondents would still perceive they received the level of training in accordance with the best version they received). In regards to revising the analysis using post hoc considerations, we refer the reviewer to our response to Reviewer #5 (above). We sincerely hope this is acceptable to the reviewer.	No change.
