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 a research article's country of origin affect perception of its quality and relevance? A national trial of US public health researchers.
AUTHORS	Harris, Matthew; Macinko, James; Jimenez, Geronimo; Mahfoud, Maen; Anderson, Chloe

VERSION 1 - REVIEW

REVIEWER	Peter Rockers Assistant Professor, Department of Global Health, Boston University School of Public Health, United States
REVIEW RETURNED	13-Jul-2015

GENERAL COMMENTS	The authors aimed to test a form of cultural bias by readers of public health research studies. They conducted an interesting randomized trial that asked respondents to gauge the quality of an abstract with randomized information on the source country and institution. While the idea of the paper is interesting, I have several comments on the study itself: 1. Most importantly, the authors essentially find that there is no bias in respondents' evaluations of the quality of abstracts based on source country and institution. While one effect estimate turns out to be statistically significant, applying even a conservative Bonferroni correction to deal with the multiple comparisons would make it disappear. Despite this, the authors spend most of the paper discussing the causes and implications of biases that they do not find empirically. 2. The authors note that they purposefully designed the survey as 'speed reading' to encourage anchoring, a form of cognitive bias. While this is interesting, it does not seem to be the most relevant design for determining how readers determine the quality of a study. We know that determining the quality of a study based on the limited information in an abstract is never a good idea. Artificially inducing an environment where respondents are forced to rely on biases and then seeing which biases dominate has limited implications for understanding how readers of scientific literature actually interpret evidence. 3. The low response rate is a problem, and the authors do not do enough to make the case that their sample is representative. While there is mention of responder and invitee similarities and differences, it would be useful to include a more formal comparison table. We might actually expect responders to be more considerate of the merits of scientific evidence than invitees (i.e., less prone to bias), suggesting a potential bias toward the null.

	4. While the authors frame their study as one concerned with bias related to high- and low-income countries, in fact survey respondents were presented with specific universities. It is very likely that many respondents would interpret this information more specifically than just high- or low-income country. In other words, respondents may have had biases against certain universities presented that have nothing to do with the income level of the country the university is in. 5. A minor comment: the Ivy League variable is strange. It is not entirely clear what it is meant to capture. If it is meant as a proxy for prestige or quality of the institution, certainly there are more sensitive measures.

REVIEWER	Stephan H. Winnik, MD, PhD University Heart Centre Zurich, Zurich, Switzerland
REVIEW RETURNED	19-Jul-2015

GENERAL COMMENTS	In the field of scientometry, where the majority of studies are retrospective and uncontrolled, this study stands out with regard to its design. The manuscript is well written and the statistics are sound. Strengths are 1) the successful randomization of the two sets of abstracts to the study participants, and consecutively no differences btw. the groups regarding demographics 2) a high proportion of actual peer reviewers among both groups of respondents, suggesting representativeness for the scientific community 3) inclusion of a multivariable analyses, adjusting for a number of important co-variates Potential drawbacks are 1) the fact that only four abstracts were rated, which may disguise a selection bias 2) the somewhat arbitrary choice of originating countries and institutions 3) the restriction to healthcare faculty of a single state, which may, however, be regarded as representative for the U.S. 4) the rather low response rate, which, however, suffices with regard to the statistical power Minor: 1) The predictive values of the time spent reading the abstract for abstract referral to peers are somewhat contradictive among the different abstracts and their implication remains unclear. The relevance of these findings in context of the main question regarding a putative reviewer bias is questionable. 2) The authors state that "… income and development level of … source countries seem to determine, whether a manuscript is selected for publication", and cite Tutarel et al. (2004) on the composition of editorial boards. It appears more appropriate to undermine this statement with a representative work, we performed earlier on the wealth of nations and the dissemination of research (International Journal of Cardiology 169 (3): 190–95. doi:10.1016/j.ijcard.2013.08.101).

	3) p13, paragraph 1, line 8: Correct “We group…” to “We grouped…”

REVIEWER	Jeff Bakal University of Alberta, Canada
REVIEW RETURNED	14-Aug-2015

GENERAL COMMENTS	Overall I think this is a well presented study. However I feel that the authors have hinged the results on what is likely a type I error. In table 2 there are 24 comparisons and one with a marginally significant p-value. As such I think the presentation of this result is a little over played. I do think the study has merit and even though the authors didn't get the result they may have hypothesized initially this study can still ad value to the literature. As a minor point the Table 1 would benefit from N's in addition to percentages and the mean time in table 2 should also have an SD or appropriate measure of spread associated with it. In table three I am also concerned about the bias associated with the cutoffs chosen for the time spend reading. Can a continuous measure be used as well?

REVIEWER	Jesse Berlin Johnson & Johnson, USA I am a full-time employee of Johnson & Johnson, but I see no conflict in reviewing this methodological paper.
REVIEW RETURNED	25-Aug-2015

GENERAL COMMENTS	Major and Essential 1. Abstract: It’s unclear in the abstract, but clear in the methods, that the randomization was such that EACH abstract appeared half the time as high income and half the time as low income. The way you currently word the abstract could be interpreted as each abstract being classified “permanently” as high or low, and the randomization wasn’t necessarily “matched” / blocked /paired within each abstract. If you think carefully (if ONE thinks carefully), it’s clear that the only correct way to do the study you did was to be able to compare high vs. low WITHIN abstract, but it’s not clear on first reading of YOUR abstract. 2. Page 6 of 46, lines 13-14: “Government regulators consider the reliability of an innovation more positively than industrial scientists”. I know there’s a reference, so assume there’s appropriate support for this claim, but I’m having a hard time understanding what it means. What is the “reliability of an innovation?” Is it effectiveness? Predictability of the response? And do you mean “industry-based scientists?” (That would be those employed by a for-profit industry?) This is just a matter of clarification. 3. Page 13 of 46, lines 11-14: Sorry for the multi-part comment, but this is all about statistical methods: a) Assuming you used the 1-10 response as your outcome variable

	(which I'm not sure I could find stated explicitly), then why use Poisson regression? That would typically be used for count data (numbers of events, for example), which you do not have. The multivariable analog to the t-test would be a linear regression. You're calculating rate ratios in Table 3, but I'm not sure I know what the rate is a rate OF. If you really believe the Poisson is the correct model, then you need an explicit justification. b) Presumably, because of the randomization, the unadjusted analysis would be primary (the t-test). Was the t-test, in fact, the primary analysis (of the primary outcome)? These details need to be specified. c) Technically, the 1-10 outcome is an ordinal variable. Why not use (either instead of, or at least in addition to, the t-test) a Wilcoxon test? The multivariable analog would be some kind of ordinal logistic model (ideally), but IF you show that the t-test and Wilcoxon test give very similar results, I would (personally) be OK with using linear regression. (That's not true of all statisticians, so again, a clear justification is needed for whatever choice you make). I'll also freely admit that I don't believe your conclusions will change dramatically when using different methods, but that's an empirical question. d) You had 4 abstracts. Table 2 suggests you just looked separately at each abstract (which is fine). Was that the analysis that was pre-specified in your analysis plan? I ask because you could do an analysis separately of each abstract, comparing high vs. low income, or you could lump everything together, or you could first lump, but include a categorical variable for "abstract number" and include that as a set of indicator variables, or you could include indicator variables and test the "abstract X income" interaction. My guess is that lumping everything together might have diluted the difference you see for the one abstract where there was an "effect" of income. This way you get to conclude that income may matter SOME of the time, but on average (if you were to analyze all 4 abstracts together), I suspect your conclusion would be that income doesn't matter. e) For the secondary analyses using categories of response (high, middle, low?), what statistical model did you use? 4. Page 15 of 46, lines 8-10: You report IRRs in Tables 3 and 4, but (following from my question about the use of Poisson models), I don't know what "rates" are being compared, when the outcome variable is an ordinal (or continuous, if you prefer to think of it that way, having used the t-test). 5. Page 16 of 46: You should probably provide a brief discussion of the MMR example, for those who are not familiar with it. I don't think it's fair to just refer people to another paper or to assume everyone knows the example. 6. Discussion: I found the order of presentation in the Discussion a bit unexpected. You might consider changing things around so that you focus first on the research findings, then get into the discussion about diffusion / adoption of technology, which is a bit less directly related to the specific findings of your research. 7. In looking at the research findings, I would advise caution in interpreting findings around speed. Since this was framed as a speed reading study, people may not have read things as they

	normally would. (They may have expressed a "need for speed," which may have affected their interpretation of the abstracts.) MINOR but essential: 8. Page 6 of 46, line 21: Spell out OECD when first used. 9. Page 15 of 46, line 20: There is a typographical error. The IRR is reported here as 0.09 when it should be 0.90. (I hope this becomes irrelevant, unless you can justify the use of a Poisson model.) 10. Page 20 of 46, lines 8-11: "We cannot speculate as to the triggers individuals identify with when reading each individual abstract under relatively rapid, timed conditions but it is encouraging that, despite the wide variation in scores given to the abstracts, that overall there were few differences between the two survey groups". Actually, the wide variation might be contributing to the lack of ability to detect differences. Your two questions may simply not provide much ability to discriminate. You could explore the measurement qualities of your questions in a separate validation study, in which you compare abstracts that are deliberately constructed to be of "good" or "poor" quality (and I know I'm being vague about what I mean by "quality" here.) 11. Given the ambiguity of your findings, you might be hesitant to make a specific recommendation about the implications of the findings. Having said that, would you suggest that reviewers of submitted manuscripts be masked to country of origin? (It's an odd question for a paper in a journal that uses a completely non-anonymous review process.) I'm just thinking "out loud" about this.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Reviewer Name Peter Rockers

Institution and Country Assistant Professor, Department of Global Health, Boston University School of Public Health, United States

The authors aimed to test a form of cultural bias by readers of public health research studies. They conducted an interesting randomized trial that asked respondents to gauge the quality of an abstract with randomized information on the source country and institution. While the idea of the paper is interesting, I have several comments on the study itself:

Most importantly, the authors essentially find that there is no bias in respondents' evaluations of the quality of abstracts based on source country and institution. While one effect estimate turns out to be statistically significant, applying even a conservative Bonferroni correction to deal with the multiple comparisons would make it disappear. Despite this, the authors spend most of the paper discussing the causes and implications of biases that they do not find empirically.

Whilst we agree that a Bonferroni correction would be required if we had made multiple comparisons, it is not needed in this case because we only conduct within---abstract comparisons, comparing the scores given for high--- vs low---income sources of the same abstract. Each analysis point is therefore a different outcome and although we have three outcomes for each abstract, we are still only comparing the high---vs low---income sources for each abstract, not across abstracts.

The authors note that they purposefully designed the survey as 'speed reading' to encourage anchoring, a form of cognitive bias. While this is interesting, it does not seem to be the most relevant design for determining how readers determine the quality of a study. We know that determining the quality of a study based on the limited information in an abstract is never a good idea. Artificially inducing an environment where respondents are forced to rely on biases and then seeing which biases dominate has limited implications for understanding how readers of scientific literature actually interpret evidence.

We thank the reviewer for this important comment. It is certainly true that the quality of a study can only be determined by thorough critical appraisal of the full manuscript. We took the view that, in practice, reading an abstract is frequently how researchers decide whether to continue to read the full manuscript. Indeed, this is certainly the case for those engaged in systematic reviews. Our research design therefore was based on the how people habitually consume research. There is a trade---off also between the accuracy with which people are able to rate research, and the ease of completing an on---line research survey. Our design sought to find a balance between these competing requirements. As Reviewer 4 also notes, the wide Standard Deviations of the mean scores given to each abstract, may be explained by the fact that the participant only has the abstract upon which to base his/her evaluation. This is a limitation of the research because it is then more challenging to establish significant differences between the groups' ratings of the research. Nonetheless, in the context of this randomised controlled trial, where both groups are evaluating the same abstracts, the challenge in determining the quality of the abstract is shared between the two groups.

The low response rate is a problem, and the authors do not do enough to make the case that their sample is representative. While there is mention of responder and invitee similarities and differences, it would be useful to include a more formal comparison table. We might actually expect responders to be more considerate of the merits of scientific evidence than invitees (i.e., less prone to bias), suggesting a potential bias toward the null.

Thank you for these interesting comments. As noted by Reviewer 2, the power of the study was adequate to detect a relatively small difference in mean scores between the two groups, so although the response rate seems low for surveys it has not posed any problems for the analysis of the data. Furthermore, with regards representativeness, the respondents are 10% of the entire universe of Public Health researchers in the US, from all CEPH accredited institutions, across all 50 states. Our respondents are therefore a very significant proportion of that population. Finally, although we could not collect more information on the invited participants than that which was provided in their institutional websites we could identify the gender (for the majority), the region and the institution type. Based on these characteristics, we note that there tended to be more females, and more respondents from CEPH Programmes in Public Health, than would be predicted by the characteristics of the

entire population of invitees. This has already been noted in the manuscript but we provide a detailed breakdown in the table below for the reviewer's interest. It is difficult to state whether, based on this difference in respondent type, there would be any substantial selection bias, and in which direction. We note, however, that far more relevant, is that the characteristics between the group that responded to the high---income source abstracts and the group that responded to the low---income source abstract were identical. The respondents might be more motivated to respond to this type of survey, but this would, again, be shared across the two groups and therefore not influence the findings. If respondents were more motivated to respond because of an interest and sympathy towards global health issues or research from low---income countries, then this might bias the findings. However, it would still be shared across the two groups, and we avoid the possibility by framing the study as a Speed---Reading Survey.

	Whole panel (n=9,421)
Gender	
Males	48.98%(4,614)
Females	47.48%(4,473)
	DK = 3.55% (334)
Age	
21---40	n/a
41---50	n/a
51---60	n/a
61+	n/a
Qualifications	
Academic	n/a
Professional	n/a
Academic & Professional	n/a
Birth country	
US born	n/a
Non---US born	n/a
Research familiarity	
Daily	n/a
Weekly, or less frequently	n/a
CEPH type	
CEPH School	68.48%(6,451)
CEPH Program	29.47%(2,776)
Ivy league	
Yes	12.93%(1,218)
No	85.02%(8,009)
US region	
Northeast	27.86%(2,624)
South	40.68%(3,832)
Midwest	15.84%(1,492)
West	13.02%(1,226)

While the authors frame their study as one concerned with bias related to high--- and low--- income countries, in fact survey respondents were presented with specific universities. It is very likely that many respondents would interpret this information more specifically than just high--- or low---income country. In other words, respondents may have had biases against certain universities presented that have nothing to do with the income level of the country the university is in.

The abstract source did indeed mention the university, and we cannot rule out the possibility that an individual has specific prejudices against one of the selected institutions. We minimised the elements of the anchoring to just country and institution and it would indeed be interesting to explore biases to specific institutions in another study. However, in this study, the country was included in at least three locations throughout the abstract. We believe that even if a particular respondent has certain attitude towards an institution, this would be as likely to occur in either of the two study groups and therefore will influence the analysis in equal measure.

A minor comment: the Ivy League variable is strange. It is not entirely clear what it is meant to capture. If it is meant as a proxy for prestige or quality of the institution, certainly there are more sensitive measures.

In order to control for the possibility that some individuals might have prejudices towards other institutions we included the Ivy League variable as the best proxy that we had for institutional prestige. The rationale was that people from institutions with higher prestige might be more critical of other institutions. Not only were these respondents equally distributed between the two groups, but also we found no evidence to suggest that their responses were different from the other respondents.

Reviewer 2

Reviewer Name Stephan H. Winnik, MD, PhD

Institution and Country University Heart Centre Zurich, Zurich, Switzerland

In the field of scientometry, where the majority of studies are retrospective and uncontrolled, this study stands out with regard to its design. The manuscript is well written and the statistics are sound.

Strengths are

1) the successful randomization of the two sets of abstracts to the study participants, and consecutively no differences btw. the groups regarding demographics

2) a high proportion of actual peer reviewers among both groups of respondents, suggesting representativeness for the scientific community

3) inclusion of a multivariable analyses, adjusting for a number of important co---variates

We thank the reviewer for noting that the randomization of the two sets of abstracts to the study participants was conducted adequately, and that the characteristics of the two groups that respond to each abstract were similar. This is an important feature of this research ensuring that the difference in the rating of the abstracts between the two groups can be explained only by the one attribute that differed between the two groups i.e. the source of the abstract.

Potential drawbacks are

1) the fact that only four abstracts were rated, which may disguise a selection bias

We agree that the number of abstracts might disguise a selection bias, however there is an important trade off between the length of the survey and the likelihood of finding a bias in one of the abstracts. In order to account for the full range of possible interests that the respondents have, including also to chose abstracts that sufficiently represent the types of abstracts that possible respondents are likely to draw upon in their daily activities, we decided to use these four abstracts that cut across a range of interest areas, methodologies and study designs. In future research, it would be interesting to evaluate the extent to which responses vary by study design, topic, methodology etc.

2) the somewhat arbitrary choice of originating countries and institutions

We thank the reviewer for raising this important point because it is certainly a challenge to select candidate countries and institutions. We actually used, and have described in detail in the manuscript, some explicit basic criteria to choose the institutions and countries, representing different income levels whilst also paying special attention to regional representativeness. Again, with reference to the preceding comments, there is a trade off to be made between the number of abstracts used in the survey (and therefore the number of sources), and the ease of completing the survey. As a result of the successful randomisation, any specific prejudices to some regions, countries or institutions that individuals may have will be distributed equally between the two groups.

3) the restriction to healthcare faculty of a single state, which may, however, be regarded as representative for the U.S.

Actually, the survey was distributed to all professors (full, associate and assistant) in every CEPH---accredited School or Programme of Public Health in each of the 50 states in the USA. The respondents reflected the regional distribution of these professionals.

4) the rather low response rate, which, however, suffices with regard to the statistical power

As per our comment above, we agree that despite the response rate of 10%, as this was of the entire universe of Public Health researchers in CEPH---accredited institutions, the response rate is empirically significant. The number of respondents (899) permitted statistical analysis to 80% power.

Minor:

1) The predictive values of the time spent reading the abstract for abstract referral to peers are somewhat contradictive among the different abstracts and their implication remains unclear. The relevance of these findings in context of the main question regarding a putative reviewer bias is questionable.

We thank the reviewer for raising this point. The time spent reading the abstract was calculated so that we could adjust for this in the analysis. It is however a secondary finding that warrants further investigation particularly as the time spent on the abstract does not consistently influence the rating given to it (for some abstracts the rating is higher, and in others it is lower). This may be due to the type of study described in the abstract (whether an RCT or a cross---sectional design). We ensured that the abstracts were presented in random order for each participant in order to avoid a 'fatigue---factor' across all four abstracts.

2) The authors state that "… income and development level of … source countries seem to determine, whether a manuscript is selected for publication", and cite Tutarel et al. (2004) on the composition of editorial boards. It appears more appropriate to undermine this statement with a representative work, we performed earlier on the wealth of nations and the dissemination of research (International Journal of Cardiology 169 (3): 190–95. doi:10.1016/j.ijcard.2013.08.101).

We thank the reviewer for highlighting the relevance of this interesting study and we have included it in the introduction as part of the literature review. Consequently, we have also adjusted all the references and their citation numbers.

3) p13, paragraph 1, line 8: Correct "We group…" to "We grouped…"

Reviewer: 3

Reviewer Name Jeff Bakal

Institution and Country University of Alberta, Canada

Please leave your comments for the authors below

Overall I think this is a well presented study. However I feel that the authors have hinged the results on what is likely a type I error. In table 2 there are 24 comparisons and one with a marginally significant p---value. As such I think the presentation of this result is a little over played.

We address this comment in the preceding section (Reviewer 1) but to reiterate we have only conducted within---abstract analysis and so although we cannot exclude the possibility, the likelihood of a type 1 error is small.

I do think the study has merit and even though the authors didn't get the result they may have hypothesized initially this study can still ad value to the literature.

Thank you for this comment. We believe that as the first, large---scale randomised assessment of research evaluation in the US, this is an important empirical contribution to the literature.

As a minor point the Table 1 would benefit from N's in addition to percentages and the mean time in table 2 should also have an SD or appropriate measure of spread associated with it.

Table 1 does in fact include already the N's in the top line of the table, however as the reviewer suggests we have added in the SD's for the time taken to complete the abstracts in table 2.

In table three I am also concerned about the bias associated with the cutoffs chosen for the time spend reading. Can a continuous measure be used as well?

The cut---offs chosen for the time spent on reading the abstract are meaningful (less than a minute, one minute, two minutes etc) and were chosen also to divide the respondents into fairly equal groups.

Reviewer: 4

Reviewer Name Jesse Berlin

Institution and Country Johnson & Johnson, USA

Please leave your comments for the authors below Major and Essential

1. Abstract: It's unclear in the abstract, but clear in the methods, that the randomization was such that EACH abstract appeared half the time as high income and half the time as low income. The way you currently word the abstract could be interpreted as each abstract being classified "permanently" as high or low, and the randomization wasn't necessarily "matched" / blocked /paired within each abstract. If you think carefully (if ONE thinks carefully), it's clear that the only correct way to do the study you did was to be able to compare high vs. low WITHIN abstract, but it's not clear on first reading of YOUR abstract.

We are grateful to the reviewer for pointing out this issue with the abstract. We have amended the abstract accordingly to better reflect the methods used, and that the comparison was within---abstract.

2. Page 6 of 46, lines 13---14: "Government regulators consider the reliability of an innovation more positively than industrial scientists". I know there's a reference, so assume there's appropriate support for this claim, but I'm having a hard time understanding what it means. What is the "reliability of an innovation?" Is it effectiveness? Predictability of the response? And do you mean "industry---based scientists?" (That would be those employed by a for---profit industry?) This is just a matter of clarification.

To clarify this point, Dearing et al (1994) refer to 'reliability' as the degree to which an innovation is communicated as being consistent in its results. They contrast this with many other attributes of an innovation, such as economic advantage, effectiveness, observability, complexity, trialability etc and noted that different communities of practice value different attributes to varying degrees. Their research included industrial scientists that were employed by corporations, government regulators and consulting engineers.

3. Page 13 of 46, lines 11---14: Sorry for the multi---part comment, but this is all about statistical methods:

a) Assuming you used the 1---10 response as your outcome variable (which I'm not sure I could find stated explicitly), then why use Poisson regression? That would typically be used for count data (numbers of events, for example), which you do not have. The multivariable analog to the t---test would be a linear regression. You're calculating rate ratios in Table 3, but I'm not sure I know what the rate is a rate OF. If you really believe the Poisson is the correct model, then you need an explicit justification.

b) Presumably, because of the randomization, the unadjusted analysis would be primary (the t---test). Was the t---test, in fact, the primary analysis (of the primary outcome)? These details need to be specified.

c) Technically, the 1---10 outcome is an ordinal variable. Why not use (either instead of, or at least in addition to, the t---test) a Wilcoxon test? The multivariable analog would be some kind of ordinal logistic model (ideally), but IF you show that the t--- test and Wilcoxon test give very similar results, I would (personally) be OK with using linear regression. (That's not true of all statisticians, so again, a clear justification is needed for whatever choice you

make). I'll also freely admit that I don't believe your conclusions will change dramatically when using different methods, but that's an empirical question.

Taking all of these points together, our strategy to use the Poisson initially, was because we considered the scale of 1---10 to not be continuous (respondents could not chose fractions, for example). The Poisson model was used to count the number of respondents that gave each number on the scale 1---10. The IRR is also called the Prevalence Ratio i.e. the prevalence of responding that particular number on the scale in the high versus the low group. We could, if it makes the results clearer, call the IRR the Prevalence Ratio. However, we have followed the reviewer's advice and also conducted a newer analysis using a generalised ordered logit model. We did not originally use an ordered logit because some covariates violated the parallel trends assumption. We have reworked the results using a generalized ordered logit model (gologit2 in Stata) that allows for relaxing this assumption, but only for those covariates that violate it. We have updated our results tables (3 and 4) to reflect these results, noting that the overall interpretation remains mostly unchanged from our original analyses. Even though the main findings of the study remain unchanged, we have made the necessary changes to the text in the manuscript to reflect this newer analysis

(p.13 l.14, and p.15 l.8---19), assuming that the editorial team and reviewer prefer to retain this analysis compared to the Poisson variation. The primary analysis of the unadjusted outcome was indeed done using a t---test and this was noted in the manuscript on p.13 (l.15).

d) You had 4 abstracts. Table 2 suggests you just looked separately at each abstract (which is fine). Was that the analysis that was pre---specified in your analysis plan? I ask because you could do an analysis separately of each abstract, comparing high vs. low income, or you could lump everything together, or you could first lump, but include a categorical variable for "abstract number" and include that as a set of indicator variables, or you could include indicator variables and test the "abstract X income" interaction. My guess is that lumping everything together might have diluted the difference you see for the one abstract where there was an "effect" of income. This way you get to conclude that income may matter SOME of the time, but on average (if you were to analyze all 4 abstracts together), I suspect your conclusion would be that income doesn't matter.

The reviewer is correct that combining everything together dilutes the differences and that the more appropriate analysis is the within---abstract type, controlling therefore for the type of abstract. Empirically, lumping the data together conflates the different variables of abstract type, institution and country.

e) For the secondary analyses using categories of response (high, middle, low?), what statistical model did you use?

We used a univariate logistic regression model containing the binary outcome (i.e. above/below a certain threshold) and a binary indicator of the abstract's country of origin (whether the person (randomly) received the high income or low income abstract). The corresponding test is a Wald test of the beta coefficient for the abstract country of origin.

4. Page 15 of 46, lines 8---10: You report IRRs in Tables 3 and 4, but (following from my question about the use of Poisson models), I don't know what "rates" are being compared, when the outcome variable is an ordinal (or continuous, if you prefer to think of it that way, having used the t---test).

We refer the reviewer to our response above (points a---c) regarding the re---analysis.

5. Page 16 of 46: You should probably provide a brief discussion of the MMR example, for those who are not familiar with it. I don't think it's fair to just refer people to another paper or to assume everyone knows the example.

Thank you for pointing this out – we have added in a sentence to describe the problem that occurred in this case (p.16, l.5---9)

6. Discussion: I found the order of presentation in the Discussion a bit unexpected. You might consider changing things around so that you focus first on the research findings, then get into the discussion about diffusion / adoption of technology, which is a bit less directly related to the specific findings of your research.

We thank the reviewer for this suggestion and defer to the editorial team to advise on the necessity of this change.

7. In looking at the research findings, I would advise caution in interpreting findings around speed. Since this was framed as a speed reading study, people may not have read things as they normally would. (They may have expressed a "need for speed," which may have affected their interpretation of the abstracts.)

Thank you for pointing this out – we have added a sentence to this effect (p.19 l.4---6)

MINOR but essential:

8. Page 6 of 46, line 21: Spell out OECD when first used.

Thank you – we have added in the meaning of OECD (p.7 l.23)

9. Page 15 of 46, line 20: There is a typographical error. The IRR is reported here as

0.09 when it should be 0.90. (I hope this becomes irrelevant, unless you can justify the use of a Poisson model.)

All of the results are now presented in accordance with the newer analysis.

10. Page 20 of 46, lines 8---11: "We cannot speculate as to the triggers individuals identify with when reading each individual abstract under relatively rapid, timed conditions but it is encouraging that, despite the wide variation in scores given to the abstracts, that overall there were few differences between the two survey groups". Actually, the wide variation might be contributing to the lack of ability to detect differences. Your two questions may simply not provide much ability to discriminate. You could explore the measurement qualities of your questions in a separate validation study, in which you compare abstracts that are deliberately constructed to be of "good" or "poor" quality (and I know I'm being vague about what I mean by "quality" here.)

Thank you for pointing this out. We have added in a sentence to highlight this point on p.20 l.4---6.

11. Given the ambiguity of your findings, you might be hesitant to make a specific recommendation about the implications of the findings. Having said that, would you suggest that reviewers of submitted manuscripts be masked to country of origin? (It's an odd question for a paper in a journal that uses a completely non---anonymous review process.) I'm just thinking "out loud" about this.

This is an important point. As this is the first study of its kind, to the best of our knowledge, it is difficult to know whether, at a population level, this sort of finding will be replicated.

The study sets a benchmark for what to expect empirically at least, and we have added in a statement to this effect in the final paragraph of the discussion (p. 21 l. 4---8). In a related research study, also conducted this year in the US, we undertook a qualitative exploration of the barriers to Reverse Innovation. Our findings were, as the Reviewer notes, related to the biases invoked in response to country---of---origin, and our recommendations were precisely to consider the value of revealing country---of---origin in an innovation process. The manuscript is currently being revised (minor revisions only) for the journal Globalization and Health which has a rolling series on Reverse Innovation. We have taken the liberty therefore of referencing this study in the Discussion.

VERSION 2 – REVIEW

REVIEWER	Peter Rockers Assistant Professor Department of Global Health Boston University School of Public Health
REVIEW RETURNED	21-Oct-2015

GENERAL COMMENTS	The authors' response to my first comment (and the same comment made by the 3rd reviewer) is insufficient. The main point of the comment is not whether a Bonferroni correction is needed, but rather that while the empirical results seem to refute the original hypothesis of the study and suggest limited bias on the part of their study subjects, the framing and discussion of the paper seem to ignore this fact. The authors seem set on making an argument around a behavior that they do not find empirically.

REVIEWER	Jesse Berlin Johnson & Johnson, USA I am a full-time employee of Johnson & Johnson, but see no competing interest in the review of this methodological paper.
REVIEW RETURNED	04-Oct-2015

GENERAL COMMENTS	Major and Essential 1. I know you responded to this, but I'm not sure I made myself clear in my initial comments, based on your response. You had 4 abstracts and looked separately at each abstract (which is still fine), using within-abstract comparisons. When I said "lump," I didn't mean to imply ignoring the within-abstract nature of the design. Essentially, I was suggesting that you "stack up" the within-abstract comparisons and include a categorical variable for "abstract number" and include that as a set of indicator variables. That model, in fact, does retain the within-abstract nature of the comparison. What you've done assumes that there is an "abstract X income" interaction, which then allows you to look separately at each abstract (more on this below). My suggestion was equivalent to taking a weighted average, across all abstracts, of the within-abstract differences. That would also partly address the concern from the other reviewers about the multiple comparison issue. If you're arguing, as you seem to be, that the abstracts are so different from each other that it makes little sense, conceptually, to do a single analysis, that's fine, but you should be more explicit about that argument.

	2. I hadn't intended to get into the discussion about multiple comparisons explicitly, but again, I suspect that the analysis I proposed would affect your significant finding. If you think of the problem as similar to how a randomized clinical trial would be analyzed, then there could be an argument against restricting the analyses to looking separately at each abstract. In a clinical trial, one typically expects to see an overall test of the treatment comparison. In some instances, the investigators will then go on to do separate analyses within subgroups of participants. (Here, "abstract" is equivalent to "subgroup of participants.") Most clinical trial statisticians would argue that you need to show a significant treatment X subgroup interaction test, before doing the treatment comparisons within subgroups, in order to protect against possibly spurious subgroup findings related to multiple comparisons. You have not provided this overall test. As I noted, if there's a strong justification NOT to do the overall test, then please provide it. 3. As in my first comment, I appreciate your response about the secondary analyses using categories of response (high, middle, low). Again, I was hoping you would mention the method you used in the methods section. MINOR but essential: 4. Page 6, around lines 33-35 (the text doesn't quite line up with the line numbers): You still say, "Government regulators consider the reliability of an innovation more positively than industrial scientists". I appreciate the explanation you provided in your response and I was hoping you would include that definition in the text of the paper. It was a new term for me. (As an aside, there's a subtlety in the definition that intrigued me, which I might be over-interpreting: "the degree to which an innovation is communicated as being consistent in its results." That's different from an intervention "being consistent in its results," because someone, with a point of view, has to do the communication, whereas a government regulator could judge the consistency for him or herself. It's off topic, and I don't mean to send you off on a tangent, but I do still think it would be helpful to add something explaining that "reliability" means "consistency of results," or "communicated consistency," whichever you mean.)

VERSION 2 – AUTHOR RESPONSE

Reviewer Name Peter Rockers
Institution and Country Assistant Professor, Department of Global Health, Boston University School of Public Health, United States

The authors' response to my first comment (and the same comment made by the 3rd reviewer) is insufficient. The main point of the comment is not whether a Bonferroni correction is needed, but rather that while the empirical results seem to refute the original hypothesis of the study and suggest limited bias on the part of their study subjects, the framing and discussion of the paper seem to ignore this fact. The authors seem set on making an argument around a behavior that they do not find empirically.

Thank you for raising this issue. We think that the conclusions of the research study cannot be

distilled easily into a simple 'reject or accept' of the null hypothesis. We agree that we were expecting to find evidence of bias across all or most of the abstracts, and we only found it for one of the abstracts, and only for the rating of relevance for this particular abstract. This finding, although seemingly small, does not warrant a wholesale rejection of the null hypothesis for a number of reasons:

1. We can be fairly certain that this is not a type 1 error because the participants were balanced for known and unknown confounders, the statistically significant finding was still present despite controlling for known socio-demographic and behavioural characteristics (such as time spent on the abstract) but also we had a large sample that provided sufficient statistical power to detect the small differences that we found between the two groups.
2. The fact that we found this detectable difference for one of the four abstracts is significant from a 'clinical' perspective. Given the number of abstracts that are read and reviewed annually from all over the world, even this apparently very small difference in rating may, at scale, have significant impact on the way that research is consumed and viewed. It is beyond the scope of our article to speculate on the impact of this, but it certainly raises the concern and cannot be simply ignored.
3. The fact that our finding was less than expected does not mean we did not find anything empirically – only that our expectations need revising. This study's contribution is that it sets a benchmark for the kinds of differences one can expect to find in rating abstracts under controlled, experimental conditions.

We are, however, concerned to not give the impression that we are 'set' on making an argument – this is not our intention. So we have revised the Discussion to give a more balanced appraisal of the findings from the research. Whilst we discuss the fact that we did find a difference in one of the abstracts, we also very clearly note that this difference was very small and that by and large the respondents rated the abstracts equally irrespective of source. We hope that this addresses the reviewer's concerns, and we thank him for pointing it out.

Reviewer Name Jesse Berlin
Institution and Country Johnson & Johnson, USA

Please leave your comments for the authors below
Major and Essential

1. I know you responded to this, but I'm not sure I made myself clear in my initial comments, based on your response. You had 4 abstracts and looked separately at each abstract (which is still fine), using within-abstract comparisons. When I said "lump," I didn't mean to imply ignoring the within-abstract nature of the design. Essentially, I was suggesting that you "stack up" the within-abstract comparisons and include a categorical variable for "abstract number" and include that as a set of indicator variables. That model, in fact, does retain the within-abstract nature of the comparison. What you've done assumes that there is an "abstract X income" interaction, which then allows you to look separately at each abstract (more on this below). My suggestion was equivalent to taking a weighted average, across all abstracts, of the within-abstract differences. That would also partly address the concern from the other reviewers about the multiple comparison issue. If you're arguing, as you seem to be, that the abstracts are so different from each other that it makes little sense, conceptually, to do a single analysis, that's fine, but you should be more explicit about that argument.

2. I hadn't intended to get into the discussion about multiple comparisons explicitly, but again, I suspect that the analysis I proposed would affect your significant finding. If you think of the problem

as similar to how a randomized clinical trial would be analyzed, then there could be an argument against restricting the analyses to looking separately at each abstract. In a clinical trial, one typically expects to see an overall test of the treatment comparison. In some instances, the investigators will then go on to do separate analyses within subgroups of participants. (Here, “abstract” is equivalent to “subgroup of participants.”) Most clinical trial statisticians would argue that you need to show a significant treatment X subgroup interaction test, before doing the treatment comparisons within subgroups, in order to protect against possibly spurious subgroup findings related to multiple comparisons. You have not provided this overall test. As I noted, if there’s a strong justification NOT to do the overall test, then please provide it.

We are grateful to the reviewer for bringing this to our attention, and apologize if we did not address the query adequately after the first revision. We always intended to analyze the data using within-abstract comparison because we believe there is a strong argument to do so. Table 2 shows quite clearly that abstracts that have a stronger research design are rated higher overall, for strength of evidence, and referral to a peer, than abstracts with a weaker design and so there is a clear abstract-effect. The overall strength of each abstract differs based on its design and quite probably topic. Some are uniformly higher and some are uniformly lower - so by combining these into one overall measure of "strength" or "referral" we risk obscuring the effect of the high versus low country of origin.

We attempted the sensitivity analysis suggested by the reviewer by 'stacking up' the within-abstract comparisons and including a categorical variable for abstract number as a set of indicator variables. Ordered logistic regression using robust standard errors showed no significant difference for high-income vs low-income abstracts for strength of evidence (1.05, 95% CI 0.95-1.15) and referral to a peer (1.00, 95% CI 0.91-1.09). Due to the fact that there is considerable variation in the mean response for each abstract, based most likely on the differences in study design within each abstract, we do not believe such an analysis to adequately capture this inter-abstract variation. Interpreting results of such a pooled analysis relies on the assumption that the mean score for each abstract would be more or less equal, except for the manipulated source of the study (high or low income). Since we instead have a mix of study types, combining them together simply obscures the fact that the stronger studies always score higher than the weaker (cross-sectional) ones. The only way to ensure that this does not interfere with the findings is to control for this effect using within-abstract analysis. We do not view abstract as equivalent to subgroup of participants. To use a pharmaceutical analogy, we consider the abstracts to be separate treatments, not doses of the same treatment, and in effect we have conducted four separate test-control studies.

3. As in my first comment, I appreciate your response about the secondary analyses using categories of response (high, middle, low). Again, I was hoping you would mention the method you used in the methods section.

Thank you for pointing this out – we have now included this information in the methods section.

MINOR but essential:

4. Page 6, around lines 33-35 (the text doesn’t quite line up with the line numbers): You still say, “Government regulators consider the reliability of an innovation more positively than industrial scientists”. I appreciate the explanation you provided in your response and I was hoping you would include that definition in the text of the paper. It was a new term for me. (As an aside, there’s a subtlety in the definition that intrigued me, which I might be over-interpreting: “the degree to which an innovation is communicated as being consistent in its results.” That’s different from an intervention “being consistent in its results,” because someone, with a point of view, has to do the communication,

whereas a government regulator could judge the consistency for him or herself. It's off topic, and I don't mean to send you off on a tangent, but I do still think it would be helpful to add something explaining that "reliability" means "consistency of results," or "communicated consistency," whichever you mean.)

Thank you – we have included this clarification in the text on page 6.

Correction

Harris M, Macinko J, Jimenez G, *et al*. Does a research article's country of origin affect perception of its quality and relevance? A national trial of US public health researchers. *BMJ Open* 2015;5:e008993. The institutional affiliation of the last author of this paper is incorrect. Chloe Anderson's correct affiliation is: MDRC, New York, NY, USA; work supported and completed while at The Commonwealth Fund, New York, NY, USA.

BMJ Open 2016;**6**:e008993corr1. doi:10.1136/bmjopen-2015-008993corr1