Author's response to reviews

Title: Researchers' perceptions of research misbehaviours: A mixed methods study among academic researchers in Amsterdam

Authors:

Tamarinde Laura Haven (t.l.haven@vu.nl)
Joeri Tijdink (jk.tijdink@vumc.nl)
H.Roeline Pasman (hrw.pasman@vumc.nl)
Guy Widdershoven (g.widdershoven@vumc.nl)
Gerben ter Riet (g.terriet@amsterdamumc.nl)
Lex Bouter (lm.bouter@vu.nl)

Version: 1 Date: 12 Jul 2019

Author's response to reviews:

Dear Dr Kowalczyk,

Many thanks for your and your reviewers’ thorough assessment of our manuscript. Below we insert the reviewers’ comments in their original order. We respond to the reviewers’ comments in italics. Below our response, we copy-pasted the pieces of the revised manuscript with changes accepted that we changed in accordance with the reviewers’ requests. Those pieces are marked with quotation marks.

Reviewer #1: This is a well-designed study that compares perceptions of the importance of research misbehaviors in different fields. The results are interesting and shed some light interdisciplinary differences in research integrity. The paper could use some editing for English grammar and usage. For example, the first sentence of the papers says: "Most researchers believe to possess a considerable degree of integrity, honesty and objectivity." I think what the authors meant to say is "Most researchers are believed to possess..."

Response: Thank you for your encouraging words. We have asked a native colleague to proof-read the manuscript. The first sentence in the revised manuscript with changes accepted now reads as follows, see lines 64-66:
“Most researchers think of themselves as honest and consider their work to be conducted with integrity (1–3). In spite of this, there is increasing evidence that researchers misbehave quite frequently in their work (4–6).”

Reviewer #2: Thank you for proving me the opportunity to review this interesting manuscript. The manuscript has many positive qualities - it addresses a timely and highly relevant issue, it employs an ecologically valid and innovative methodological approach, the goals are clearly articulated, the presentation is concise and transparent. Further, the authors provide useful supplementing materials.

Response: Thank you for your kind words.

(1) A main concern revolves around the multimethod design, specifically the order of the survey and the focus group study. It goes without saying that it is perfectly fine to do focus groups to better understand qualitative aspects of the data. In the present case, however, it seems that in the present study focus groups were conducted to resolve ambiguities in survey items that should have been resolved before conducting the survey. I am aware that the authors address the ambiguity issue in the general discussion, but for me their elaborations raised more questions than answers.

Response: Thank you for the opportunity to clarify the aim of the focus groups. The aim of the focus groups was not to resolve ambiguities in the survey, but to better understand the survey results. Our aim was twofold. On the one hand, we wanted to know whether researchers recognized the top 5 research misbehaviours we identified based on the survey as relevant for their disciplinary field. On the other hand, if they did not recognize (some of) these 5 research misbehaviours, we wanted to give participants the opportunity to present and discuss other research misbehaviours that they considered (more) relevant to their disciplinary field.

We now explain this explicitly in the methods section, see lines 190-195 and further discuss this in the discussion section, see lines 418-431 in the revised manuscript with changes accepted:

[METHOD SECTION]

“We extended the survey results with focus group interviews. Our aim was twofold. First, we wanted to know whether researchers recognised the top 5 research misbehaviours we identified
based on the survey as relevant for their disciplinary field. Second, if they did not recognize (some of) the research misbehaviours, we gave participants of the focus group interviews the opportunity to present and discuss other research misbehaviours that they considered (more) relevant to their disciplinary field.”

[DISCUSSION SECTION]

“This mixed-method study, involving a survey followed by focus groups, aimed to develop insight into what academic researchers in Amsterdam from different disciplinary fields considered to be the most detrimental research misbehaviours. There are a few important takeaways from our study. First, based on the survey results, we found insufficient supervision and various forms of sloppy science to score highly on aggregate impact throughout all disciplinary fields. Researchers from the natural sciences and humanities also perceived nepotism to be of major impact on the aggregate level. The natural sciences regarded fabrication of data of major impact as well. The focus group interviews helped us to understand how researchers interpret ‘insufficient supervision’. Besides, the focus group participants added insight into sloppy science in practice. Second, researchers from the natural sciences and humanities added new research misbehaviours concerning their disciplinary fields to the list, such as the stealing of ideas before publication. This improves our understanding of research misbehaviour, or ‘questionable research practices’ beyond the social and biomedical fields.”

(2) With regard to the conclusions that can be derived from the present data discussing potential item ambiguities remains dissatisfying because this issue poses a serious threat to the study's validity (see Fielder and Schwarz (2016) for a more general analysis of theoretical and methodological challenges in the prevalence estimation of QRP from self-reports). The authors suggest tackling the ambiguity issue in future research (p. 17, l. 346). I fully agree that an improved study is an important next step. I think it is necessary, in fact, to address the limitations of the present work.

Response: Thank you for suggesting the work by Fielder and Schwarz (2016). In our study, we asked how frequently researchers had witnessed a particular behaviour. This does not yield a valid estimate of the prevalence, which we now explicitly state in the manuscript. We now suggest a potential interpretation of the perceived frequency below the table in Additional file 6. Also, we have integrated the work of Fielder and Schwarz into the discussion of the main findings, see lines 432-443 in the revised manuscript with changes accepted and additional file 6.
“The frequency response scale ranged from 1 (‘never’), 2 (‘once or twice’) to 3 (‘three times or more’). A mean value of 2.03 (“Insufficiently mentor or supervise junior co-workers”) thus means that on average, our respondents stated seeing this research misbehaviour once or twice within the last three years.”

[DISCUSSION SECTION]

“When comparing our findings to the literature, it is important to keep in mind that our findings are not prevalence estimates. Equating the self-reported proportion of a research misbehaviour with its prevalence has been criticized, see Fielder & Schwarz (23). Moreover, in our survey we asked respondents to report how often they had witnessed a particular research misbehaviour, not how often they had engaged in such behaviour themselves. We then combined this with the degree of impact respondents assigned to that item to obtain the ‘aggregate impact’.”

(3) Keeping the limitations of the present items in mind, I would also comment on one of its main findings. It appears that across disciplines insufficient supervision of junior co-workers is considered a key reason for research misbehaviors. From a psychological perspective, there seems to be a "double-natured" sense of directing responsibility from self to others: First, other colleagues failing to supervise junior co-workers sufficiently; second, junior researchers who employ QRP due to a lack of supervision (which by the way depicts junior researcher as rather dependent and naïve subjects). Such other-directed attributions may have a rather self-serving quality (e.g., expressing solidarity with ill-supervised junior researchers to make oneself feel good). There may be other psychological processes involved that deserve attention (e.g., participants' responses reflect aspects of current social representation of their field that cannot simply equated with objective realities; see Stürmer et al., 2018). In sum, I would like to encourage the authors to reflect a bit more about the "psychology" implicated in participants' judgements.

Response: Thank you for your interesting comment and for giving us the opportunity to clarify one of our main findings. We agree that the item “insufficient supervision of junior co-workers” is double-natured. However, based on our findings we cannot conclude that insufficient supervision is a key reason for research misbehaviour by the supervisees. In our study, we regard insufficient supervision a research misbehaviour in itself. In principle, a researcher must comply with the professional norms of her field. Adequately supervising and mentoring junior co-workers is one of those professional norms. Non-compliance with this norm then refers to the
situation where a researcher is negligent on their supervision duties. We consider this a research misbehaviour, as many other research misbehaviours are considered misbehaviours since they describe noncompliance with the professional norms in the field.

However, we agree with the reviewer that insufficient supervision could lead to research misbehaviour by the supervisee, although we did not intend to portray junior researchers as naïve. We now reflect on the double-layered nature of insufficient supervision (why we think insufficient supervision is both a research misbehaviour and a potential reason for research misbehaviours by supervisees) in our general discussion, see lines 443-458 in the revised manuscript with changes accepted.

“Somewhat surprising is the consistent recognition of insufficient supervision and mentoring. We would like to reiterate that we regard insufficient supervision a research misbehaviour in itself. Like many other research misbehaviours, insufficient supervision describes non-compliance with one of the professional norms in academic research (adequate mentoring).

Yet, it seems plausible that insufficient supervision could, in some cases, lead to the supervisees unintentionally engaging in sloppy science because they were not socialized well into responsible conduct of research (24). However, we believe that the influence of insufficient supervision may go further. If a supervisor fails to create a safe learning climate, this could lead to situations where PhD students do not feel confident to share their concerns about a mistake (e.g. in the data-analysis) or to oppose their supervisor’s interpretation. Similarly, Roberts and colleagues (25) put forth the speculation that when the supervisor creates an environment where only spectacular outcomes are valued, supervisees may engage in sloppy science because that yields the desired outcomes. Nevertheless, in our study we did not investigate the possible reasons for research misbehaviours and investigating this would require a different research design.”

Regarding the “psychology” implicated in participants’ judgements: let us imagine that participants were in fact keen to accept insufficient supervision as an explanation for the occurrence of research misbehaviour. Social representation theory, as explained in Stürmer et al. (2018), tells us that this still need not be the best explanation, let alone an accurate reflection of reality. Applied to our field: It could be that researchers coined insufficient supervision since it fits the narrative that research misbehaviour is incentivised by the way academic science is organized. That way of making sense of the occurrence of research misbehaviour is understandable and not uncommon (e.g., Davies, 2018). Yet, in our study we did not investigate the possible reasons for research misbehaviours. Therefore, proper reflections on participants’ judgements to the question whether A is a key reason for B is outside the scope of this paper.
Minor point: Reference 16 indicates that parts of the data have been published elsewhere (PlosOne). For the reader some information concerning the distinctive and novel contributions of the present paper would be helpful.

Response: Thank you for directing our attention to the relevance of our research reported here in regards to the previous findings. The data were gathered as part of the same survey questionnaire, but the data we discuss here have not been included in any other publication. We now clarify this in the Methods section and indicate what these data contribute, see lines 155-160 in the revised manuscript with changes accepted.

“The survey consisted of three parts, one of which was the list of 60 research misbehaviours described here. The remainder comprised two instruments, one about the research climate for integrity (17) and another about the degree of perceived publication pressure (18). The data described here concern perceptions of research misbehaviours and thus extend the previous findings by describing which research misbehaviours impact the research climate and go further by exploring this in a qualitative focus group study.”

Reviewer #3: In this study, the authors report on the results of a survey and follow-up focus groups to understand how researchers at universities in Amsterdam in different disciplines and career stages view research misbehaviors. While the findings may offer some contributions to the field, and I appreciate the work that went into collecting these data, my view is that the current formulation of the manuscript lacks details to ensure the clarity and rigor of the study purpose, approach, findings, and conclusions. I would recommend the following edits, details, and rationale to be addressed in each section:

Title:

(1) I recommend editing the title, as it is not an accurate reflection of the study. The study does not examine whether research misbehaviors differ by field. Maybe edit to be “perceptions of research misbehaviors,” or something like this.

Response: Thank you for your suggestion, we agree that the title may not be the best summary of the study. We now propose, see lines 1-2:
“Researchers’ perceptions of research misbehaviours: A mixed methods study among academics in Amsterdam”

Introduction:

The introduction indicates that the study examines what researchers view at the most important research misbehaviors. Later we find that "importance" was computed through scores given regarding the frequency and impact of the behaviors. While this method for computing "importance" may be justified, I'm not sure it represents a measure of "what the researchers view as the 'most important' research misbehaviors." More justification of this metric is needed.

Response: Thank you for your valuable comment that helps us in providing a more coherent justification of our metric. Our reasoning was as follows and is also present in Bouter et al. (2016): falsification, fabrication and plagiarism (FFP) are most detrimental to science. If a research falsifies or fabricates data, the impact on the validity of the findings is major. Since FFP are most likely relatively rare, especially compared to milder research misbehaviours, it is not useful to overemphasize the importance of FFP. Instead, we decided to focus on research misbehaviours that were perceived frequently and that impacted the validity of the findings. We agree with the reviewer that, in the survey, it was not the participants themselves that classified the misbehaviour as important. We followed the methodology of Bouter et al. (2016) by combining participants’ two answers and assuming that this product score can be interpreted as a proxy for aggregate impact. We have adjusted this in the revised manuscript with changes accepted, see lines 167-176:

“Second, although falsifying data, fabricating data or committing plagiarism are most detrimental to science, they are relatively rare and therefore it is not useful to overemphasize the importance of FFP. To get a sense of which research misbehaviours were most detrimental at the aggregate level, we followed Bouter et al. (11) and multiplied the impact score of each research misbehaviour with its perceived frequency. In particular, we use the product score (multiplication) of impact and frequency as a proxy for aggregate impact throughout this manuscript.”
Furthermore, more conceptual clarity of what an "important research misbehavior" means is necessary, as it seems confusing to refer to a misbehavior as important. What is meant by "important"? Please be specific.

Response: Upon closer inspection, we agree that it is indeed confusing. We meant important as something one should not neglect, but a more sensible interpretation of our metric would be ‘aggregate impact’. We have now changed “important” to “impactful at the aggregate level” or “aggregate impact” throughout the manuscript. To follow-up on this clarification, we have changed the manuscript accordingly, see the quotation under the first comment above.

Similarly, in relation to your measurement and methods, if the most frequent are also not the most impactful—and the introduction specifically mentions that misbehaviors that are more frequent could be impactful because they are frequent—why multiply the ratings of frequency and impact together to create a new variable? What precisely does this new variable offer us in interpreting the findings? I'm not saying this approach can't be potentially justified, just that the rationale isn't clear enough.

Response: In line with the above explanations, we believe that frequency ratings and impact ratings both present a somewhat distorted picture. Our product score is an attempt to present a metric that represents aggregate impact by combining both aspects. We believe that aggregate impact thus operationalized can be helpful to the research integrity area in singling out which research misbehaviours could be priority candidates for intervention. To investigate whether the research misbehaviours had actual impact, we explicitly asked this in the focus groups. As the focus groups in general confirmed the results of the survey, our notion of ‘aggregate impact’ is supported by the qualitative findings. We now emphasize potential concerns about our aggregate impact metric in the limitations section, see lines 505-521:

“Another limitation regards the interpretation of aggregate impact. Participants did not rate the research misbehaviours to have major impact on the aggregate level, but we used the product of the perceived frequency of a research misbehaviour and the potential impact on the validity as a proxy for aggregate impact. Hence, we labeled these scores as ‘aggregate impact’ scores. The validity of this metric has no exact (mathematical) justification but is intuitively similar to e.g. the well-known QALY (Quality-Adjusted LifeYear) metric, which multiplies the subjective quality score of a state of living by the time spent in that state (36). In the focus groups, we explicitly asked whether research misbehaviours had actual impact. As the focus groups in
general confirmed the results of the survey, our notion of ‘aggregate impact’ is supported by the qualitative findings.”

Methods and Results:

(4) Several additional sentences should be added to explain and justify the planned missing data design, and how the assignment of 20 items to each person ultimately yield satisfactory sample sizes across groups (discipline; career stage) and items. Also, in this type of design, how does the research team know if the participants’ response patterns might be affected by the 20 items they see? If you see 20 items that vary greatly from the 20 items another participant sees, could this affect responding? Additionally, please justify the sample size for individual items, especially in some disciplines where the n size is quite small. This to me seems like a design concern that has to be discussed. How can we know that 40 responses, for example, on a particular misbehavior represents how researchers in humanities view that behavior?

Response: Thank you for this valuable comment. We try to answer your questions in turn, splitting them into 4a, 4b, and so on, so please continue reading.

(4a) Several additional sentences should be added to explain and justify the planned missing data design, and how the assignment of 20 items to each person ultimately yield satisfactory sample sizes across groups (discipline; career stage) and items.

Response: We presented only 20 items to each respondent because the questionnaire would otherwise be too long. As we did not plan to statistically test whether any differences were significant, we initially thought this would not be a problem. We expected the response rate to be at least 15% (7548 * .15 / 3 = 377). This means that each item is answered by at least 300 participants. The width of the 95% confidence interval around the mean would then be .11 * SD to both sides, which would yield fairly reliable estimates. Yet, this only holds for our overall interval and as our manuscript focuses on differences between groups, we conducted additional exploratory analyses below (see our answers on your comments 4c, 7b and 8). We now explain some of the shortcomings of our missing data design in the Discussion, see lines 505-512:

“A further limitation is that we presented participants with a random selection of 20 research misbehaviours because we feared that presenting them the full list of 60 would be too time-consuming. This type of design is sometimes called missingness by design, as all participants
have missing values for some items. Based on similar surveys in the field, we estimated our response rate to be at least 15%. Since our population consisted of 7548 researchers, 15% of them answering 1/3 of our items would mean at least 300 responses per item. Initially we expected more than 300 responses would be sufficient to compute reliable standard deviations, standard errors and confidence intervals.

Unfortunately, a quick glance at the width of the standard deviations in Table 2 revealed that the distribution of our scores was not normal. In fact, more than 90% of the aggregated impact variables have a skewed distribution. Consequently, we must be careful in the interpretation of the top 5. The ranking is purely based on point estimates. In fact, labeling the ranking as a top 5 may be dangerous as “top” suggest that the #1 misbehaviour ranks absolutely higher than #2. Based on our explorative analyses, it can be concluded that this only holds for biomedicine, see additional file 11. The top 5 presented in Table 2 simply lists 5 research misbehaviours that were impactful on the aggregate level and one should not overinterpret differences in the places on the list.”

(4b) Also, in this type of design, how does the research team know if the participants' response patterns might be affected by the 20 items they see? If you see 20 items that vary greatly from the 20 items another participant sees, could this affect responding?

Response: The 20 items themselves as well as the order of those 20 items was randomly generated for each participant, precluding systematic effects of order of presentation. We now explicitly write this in the methods section, see lines 147-150 in the revised manuscript with changes accepted:

“To reduce the overall length of the survey and decrease the risk of participant fatigue(16), participants were randomly presented 20 out of 60 items from the list by Bouter et al (11). To preclude order effects, the order of presentation of the 20 items was also randomised.”

(4c) How can we know that 40 responses, for example, on a particular misbehavior represents how researchers in humanities view that behavior?”

Response: The reviewer is right that the average of 40 responses may not reliably represent how humanities view a particular behaviour. To get a better sense of the precision of our estimates, we drew 1000 stratified bootstrap samples and calculated the 95% bias-corrected confidence
interval around the initial mean estimates. We then investigated how precise we were able to estimate our mean values on the rankings. As will become more evident below (see responses to point 7 and 8), we were least precise when it comes to humanities and natural sciences. A more precise picture is presented in additional file 10 that lists the top 12 for the humanities. In addition, we explain the adjusted rankings in the revised manuscript with changes accepted, see lines 297 and 309 and additional files 10 and 11:

“In biomedicine, ‘Insufficient supervision’ ranked #1 and inspection of the confidence intervals indicated that no other misbehaviour could be ranked highest. There was no overlap between the confidence interval around the mean estimate for ‘Insufficient supervision’ and the confidence intervals of the research misbehaviours listed second and third. For the natural sciences, the confidence interval around ‘Insufficient supervision’ overlapped with confidence intervals up to misbehaviours ranked twelfth. The top 12 for natural sciences can be found in additional file 10. Besides sloppy science, the top 12 for natural sciences also listed data fabrication (#7) and nepotism (#11). In the social sciences, the confidence interval around ‘Insufficient supervision’ overlapped with the confidence intervals up to the misbehaviour ranked sixth, see additional file 10. The confidence interval around ‘Insufficient supervision’ in the humanities overlapped with research misbehaviours up to rank #12. Hence, the top 12 for the humanities can be found in additional file 10. Besides sloppy science, the top 12 for researchers in the humanities included nepotism (#6).”

(5) Even though Bouter is referenced, and the items shown in an additional file, I would advise a paragraph describing the development and rigor of the survey items used in the survey. How were they developed? Are they reliable? Are the response scales used in this study from Bouter, or adapted for the current study? More details about the survey are needed. Also, were they translated? Or, was the study performed only in English?

Response: We agree with the reviewer that this will help readers to better understand our study. Therefore, we have added a short summary: the list was not developed with the intention of creating a measurement instrument. Bouter et al. (2016) first compiled a long list of more than 100 research misbehaviours. They then reviewed all misbehaviours to decrease doublings and shortlisted 60 items. These 60 items were then tested for comprehensibility on a group of 15 researchers that was commissioned to design a funding program for studies on fostering responsible research practices. This resulted in a modified list of 60. This adapted list was then distributed among 60 presenters of the World Conference on Research Integrity for review. After incorporating these expert revisions, the list was then used as input for a workshop during the 4th World Conference on Research Integrity where 26 conference attendees provided final input for
the phrasing of the items and the response scales. The full development procedure can be found in the manuscript by Bouter and colleagues.

We slightly adapted the response scales for our study. Specifically, we added the time period that respondents should keep in mind to prevent inflated findings (in line with Fielder and Schwarz (2016)). We now elaborate on the development of the survey in the Methods section and contrast the initial response scales with the response scale used in our study, see lines 112-134 in the revised manuscript with changes accepted. Lastly, the items were not translated as the study was performed in English.

“We presented participants with research misbehaviours from a list of 60 major and minor misbehaviours as composed by Bouter et al. (11). For a thorough description of the development of the list, the reader is referred to Bouter et al. (11). The list can be found in additional file 1.

In brief, they compiled an extensive list of over 100 research misbehaviours based on the existing literature on research misbehaviours. After removing duplications, 60 items remained which were tested for comprehensibility on 15 researchers. These 60 items were then distributed among keynote speakers and oral presenters of the World Conference on Research Integrity for review. Finally, the list of 60 was used in an invitational workshop at the 4th World Conference on Research Integrity (2015) which provided final input for the phrasing of the items and the relevant response scales. The list was developed and used by us in English.

We used two response scales from the initial list: frequency and impact, respectively. We altered these response scales slightly by specifying the time frame or unit respondents had to keep in mind when reading the items. The impact response scale, “How often will this misbehaviour occur?”, was changed into (italics stress our changes): “How often have you observed the behaviour stated above in the last three years?”. This question had to be answered in reference to respondents’ main disciplinary field. Answer options were 1 (“Never”), 2 (“Once or twice”) and 3 (“Three times or more”). The impact response scale, “If it occurs, how large will its impact be on the validity of knowledge?”, was changed into “If you were to observe this behaviour, how large would its impact to be on the validity of the findings of the study at issue?”. Responses ranged from 1 (“Negligible”) to 5 (“Enormous”).”

(6) Page 7 indicates that scores range from 0 to 15; elsewhere it is indicated they ranged from 1 to 15. Please correct.
Response: Thank you for noting this error, it should be 1 to 15. This has been corrected, see line 173 in the revised manuscript with changes accepted.

(7) Two additional questions about the computed variable:

(a) If the computed variable has one score with a scale of 1 to 3 and the other score a scale of 1 to 5, is the 1 to 5 scale given more weight in this computation? If so, is this intended, and is this conceptually the desired meaning of these scores?

(b) Are the standard deviations in Table 2 correct? They seem rather large. Please check, and please provide a brief interpretation of these large SDs.

Response 7a: We wanted to find out which behaviours were most detrimental on the aggregate level and the absolute highest value of our metric does not differ if the 1-5 scale had been recoded to be equal to the 1-3 scale; the highest product scores remain the highest product scores. We now clarify this in the Discussion section, see 437-441 in the revised manuscript with changes accepted.

“Because our aggregated impact metric is the product of impact (1-5) and frequency (1-3), one may wonder if we deliberately assigned impact more weight. Although this is true for the absolute score, this is not the case for the ranked aggregate impact product scores since the rank of a particular research misbehaviour does not change after recoding the impact scale.”

Response 7b: The standard deviations are correct, but the values of our 60 aggregated impact variables were not normally distributed (in fact, all were heavily skewed) so normal descriptive (standard deviations, standard error or confidence intervals) would not be reliable. Hence, we drew 1000 stratified bootstrap samples and calculated the 95% bias-corrected confidence interval around the initial mean estimates. These are presented in Additional file 10 and Additional file 11, we now write this explicitly in revised manuscript with changes accepted, see lines 512-520:

“Unfortunately, a quick glance at the width of the standard deviations in Table 2 revealed that the distribution of our scores was not normal. In fact, more than 90% of the aggregated impact variables have a skewed distribution. Consequently, we must be careful in the interpretation of the top 5. The ranking is purely based on point estimates. In fact, labeling the ranking as a top 5 may be dangerous as “top” suggest that the #1 misbehaviour ranks absolutely higher than #2.
Based on our explorative analyses, it can be concluded that this only holds for biomedicine, see additional file 11. The top 5 presented in Table 2 simply lists 5 research misbehaviours that were impactful on the aggregate level and one should not overinterpret differences in the places on the list.”

(8) Please provide a rationale for examining only the top 5 behaviors. Why 5? And, how are these interpretations of the data as showing that fields are similar or different being made? Are group comparison tests being performed? Are you adequately powered to make these comparisons given sample sizes? Even if the top 5 behaviors across disciplines sort out to the same, or similar, order, are the mean scores statistically different, if so, what does this imply? Please explain. Additionally, was an analysis done to conclude that career stage didn't influence perceptions, and therefore aggregation of career stage within discipline was appropriate?

Response: We agree that examining the top 5 is arbitrary. To statistically assess whether the top 5 made sense, we assessed the bias-corrected bootstrapped 95% confidence intervals around the mean estimates. This gives three possible scenarios: 1) The confidence intervals of X and Z do not overlap, 2) the confidence intervals overlap and the point estimate of variable X falls inside the confidence interval of variable Z, or 3) the confidence intervals overlap but the point estimate of X falls outside the lower or upper bound of Z (or vice versa). Since we report exploratory analyses, we take a conservative approach and only report a new ranking in case of scenario 1. Likewise, we only report on a difference between groups if the confidence intervals do not overlap. We now report this in the revised manuscript with changes accepted, see lines 292-336:

“The following analyses were not preregistered and should be treated as exploratory. We wanted to assess the precision of our mean estimates in Table 2. In what follows, we use bias-corrected bootstrapped 95% confidence intervals around the mean estimates.

In biomedicine, ‘Insufficient supervision’ ranked #1 and inspection of the confidence intervals indicated that no other misbehaviour could be ranked highest. There was no overlap between the confidence interval around the mean estimate for ‘Insufficient supervision’ and the confidence intervals of the research misbehaviours listed second and third. For the natural sciences, the confidence interval around ‘Insufficient supervision’ overlapped with confidence intervals up to misbehaviours ranked twelfth. The top 12 for natural sciences can be found in additional file 10. Besides sloppy science, the top 12 for natural sciences also listed data fabrication (#7) and nepotism (#11). In the social sciences, the confidence interval around ‘Insufficient supervision’
overlapped with the confidence intervals up to the misbehaviour ranked sixth, see additional file 10. The confidence interval around ‘Insufficient supervision’ in the humanities overlapped with research misbehaviours up to rank #12. Hence, the top 12 for the humanities can be found in additional file 10. Besides sloppy science, the top 12 for researchers in the humanities included nepotism (#6).

To see if the updated rankings differed between disciplinary fields, we again inspected the confidence intervals around the mean estimates. Biomedical sciences perceived ‘Insufficient supervision’ to have the greatest impact on the aggregate level, but this was not different from other fields. For the natural sciences, ‘Not report clearly relevant details of study methods’ ranked second (CI: 5.93 – 7.93). However, this rank differed significantly from the two other main disciplinary fields, i.e. the natural sciences perceived this to have a greater impact on the aggregate level than both biomedical researchers (#12, CI: 4.69 – 5.43) and researchers in the humanities (#51, CI: 2.88 – 3.97). In addition, insufficient attention to the expertise to perform the study (#5, CI: 5.23 – 7.36) ranked higher on aggregate impact in for natural sciences compared to the humanities (#36, CI: 3.03 – 4.9). Lastly, the presentation of grossly misleading information in a grant application (#9, CI: 4.54 – 6.5) as of greater impact than researchers in the social sciences (#47, CI: 3.11 – 4.00) and the biomedical sciences (#36, CI: 3.76 – 4.22).

For the social sciences, not publishing a negative study ranked second (CI: 5.71 – 7.29) and social science researchers were significantly more concerned about this than their colleagues in the humanities (#25, CI: 3.5 – 5.00). In addition, insufficient attention to the expertise to perform the study (#5, CI: 5.06 – 6.42) ranked higher on aggregate impact in for social sciences compared to the humanities (#36, CI: 3.03 – 4.9). Also, ‘Reporting an unexpected finding as being hypothesized from the start’ (#6, CI: 4.94 – 6.25) was perceived as having a greater impact on the aggregate level by social science researchers compared to researchers in the natural sciences (#34, CI: 3.24 – 4.83) and the biomedical sciences (#17, CI: 4.28 – 4.92).

Researchers in the humanities indicated selective citation to please editors, reviewers and colleagues (#5, CI: 5.13 – 7.03) to have more impact on the aggregate level compared to biomedical researchers (#23, CI: 4.11 – 4.78). Lastly, researchers in the humanities perceived the use of published ideas or phrases of others (#12) as of greater impact than biomedical researchers (#49, CI: 3.29 – 3.85) and the natural sciences (#36, CI: 3.09 – 5). All other comparisons between fields were nonsignificant, see additional file 11.”

(9) The survey response information suggests over 1000 people quit before completing the 20 survey items. This seems like an unusually high partial-completion rate. Do the authors have any idea of why this partial-response rate was so high? How long was the survey—the entire study, as I understand that this was part of a longer study? (And, in general, it would be helpful to make it clearer how the existing report is a subset of the larger project. It is hard to fully
understand this from the manuscript. I read the supplemental materials, which helped to some degree, but also created information overload.

Response: Thank you for pressing us to reflect on this. We can only guess what prompted people to stop, we received informal e-mails that respondents took longer to complete the questionnaire than we estimated in advance. It should also be noted that our percentage of participants that ‘started’ the survey includes those people that only clicked on the link. Those people did not fill in any items as they – presumably – perceived the first survey-page as too lengthy. We now elaborate on this in the Discussion section, see lines 497-503 in the revised manuscript with changes accepted:

“In addition, a high number of respondents that started answering the survey questions stopped before completing the 20 items. Before respondents were presented a random selection of 20 randomized items, they completed the Survey of Organizational Research Climate (henceforth: Source (34)). The number of participants that ‘started’ the survey included all researchers that opened the survey, even those who decided not to participate. In total, 18% of our invitees completed the Source and the later dropout rate of 3% during a survey questionnaire lies within the normal range (35).”

As for the second part of your question, the relation of the current manuscript with the rest of the project, we elaborate on this in the response to the third comment of reviewer#2 above. We explicitly describe what these data contribute to the larger study, see lines 155-160 in the revised manuscript with changes accepted.

“The survey consisted of three parts, one of which was the list of 60 research misbehaviours described here. The remainder comprised two instruments, one about the research climate for integrity (17) and another about the degree of perceived publication pressure (18). The data described here concern perceptions of research misbehaviours and thus extend the previous findings by describing which research misbehaviours impact the research climate and go further by exploring this in a qualitative focus group study.”

(10) The link between the survey portion of the study and the focus groups is confusing, and I think largely due to lack of details in the manuscript. As ‘sequential explanatory design’ would suggest, the focus groups would follow-up on quantitative survey responses, but the results
section focuses mostly on describing the new misbehaviors identified. So, was the purpose to understand the survey findings, identify new misbehaviors, or both? Page 7 states, "To better understand the relevance of the aggregated impact top 5 research misbehaviors established in the survey, we organized focus groups with researchers from three academic ranks and four disciplinary fields," however, it is confusing how the themes reported on from the qualitative interview were identified? Was this discussion of the top 5 research misbehaviors, the addition of new behaviors, or both? Also, it isn't clear what the participants were specifically asked to do in the Participative Ranking Methodology. I think this process involved reviewing all 60 behaviors, plus new ones, and ranking them. It needs to be clearer, as well, what photos were taken of. What exactly do the results of the focus groups tell us? Do they clarify why the top 5 from the survey are problematic? Do they add new behaviors to be concerns with? Also, the themes from the disciplines are fairly overlapping? Are their disciplinary differences that need to be clarified, or were the differences negligible? This isn't clear. Similarly, were there differences of note across career stage?

Response: Thank you for indicating the parts in our design that were unclear. You are right that the differences across disciplinary field, especially in the quantitative part, seem small. Differences across academic ranks were largely absent. The aim of the focus groups was twofold: to explore and understand the relevance of the research misbehaviour that seemed detrimental on the aggregate level. We now write the following on lines 189-194 in the revised manuscript with changes accepted:

“We extended the survey results with focus group interviews. Our aim was twofold. First, we wanted to know whether researchers recognized the top 5 research misbehaviours we identified based on the survey as relevant for their disciplinary field. Second, if they did not recognize (some of) the research misbehaviours, we gave participants of the focus group interviews the opportunity to present and discuss other research misbehaviours that they considered (more) relevant to their disciplinary field.”

We used participative ranking methodology that involved three steps. First, participants were presented with the 5 research misbehaviours that ranked highest on aggregate impact on post-its. Second, they were asked to reflect on the relevance of these behaviours for their disciplinary field and prompted to add new behaviours that we may have missed but that participants considered more relevant for their disciplinary field. All research misbehaviours were written down on post-its. Finally, participants were asked to reach consensus over a ranking of all the research misbehaviours based on severity. We now explicitly describe the procedure of the
ranking exercise in the methodology section, see lines 227-239 in the revised manuscript with changes accepted:

“We used a participative ranking method to structure the focus group discussion about the research misbehaviours. The procedure of the participative ranking method involved three steps. First, participants were presented with the 5 research misbehaviours that ranked highest on aggregate impact on post-its. Second, they were asked to reflect on the relevance of these behaviours for their disciplinary field and prompted to add new behaviours that we may have missed but that participants considered more relevant for their disciplinary field. All research misbehaviours were written down on post-its. Finally, participants were asked to reach consensus over a ranking of all the research misbehaviours. For that, we had created a provisional severity continuum/scale that ranged from ‘Minor’ to ‘Major’. When participants agreed on where each post-it had to be placed on the severity scale, we ended the exercise. In total, this took between 20 and 35 minutes. The remaining results of the focus groups will be part of another report. For an elaborate description of participative ranking methodology, the reader is referred to the guide by Ager, Stark & Potts (19).”

(11) More needs to be said in the methods about the coding approach and generation of themes. What made a theme a theme? Did it have to occur a certain number of times, etc.?

Response: Thank you for giving us the opportunity to describe our coding approach more clearly. For an issue to be identified as a theme, it first had to have a clear link with the research questions. Secondly, it should be either mentioned multiple times, or during the conversation be discussed as important and powerful. We added the following description detailing our coding approach, see lines 248-256 in the revised manuscript with changes accepted.

“The themes reported below are based on the qualitatively ranked research misbehaviours according to severity as well as the transcripts of the focus group conversations. Specific research misbehaviours, e.g. “reviewing without feedback, harsh reviewing, reviewers not up to scratch with developments” were clustered into broader issues, e.g. “sloppy reviewing”. For issues to be identified as emerging themes, the issue had to be related to the research question that involved research misbehaviours. Therefore, some issues that focused on political intricacies or personal integrity were disregarded. Moreover, it should be either mentioned multiple times, or during the conversation be discussed as important and powerful.”
(12) It would be good to indicate if focus group participants could have also participated in the survey, or if this was not the case.

Response: The focus group participants could have also participated in the survey. The survey was distributed one year prior to the focus groups, but participation in the survey was no requirement. We also did not ask participants specifically whether they had participated in the survey. We now clarified this, see lines 206-209, in the revised manuscript with changes accepted.

“Note that it could thus be the case that the focus group participants had also participated in the survey that was disseminated nine months prior to the start of the focus groups. Yet, we have no information to quantify this as we did not ask about it specifically.”

Discussion:

(13) I am left without a clear understanding of the key take-home points and contributions of this study in the discussion. Also, as noted above, did career stage make a difference? A brief comment might be worthwhile in the discussion.

Response: Thank you for your points. As noted above, there were no notable differences across career stages (academic ranks). We have now re-written the first paragraph of the Discussion, focusing more on the key points: Sloppy science and insufficient supervision are a concern across disciplinary fields and academic ranks. How sloppy science and insufficient supervision are understood depends on the disciplinary field. See also what we changed in the manuscript based on the second comment of Reviewer#2’, which we include here again, see lines 418-429 in the revised manuscript with changes accepted.

“There are a few important takeaways from our study. First, based on the survey results, we found insufficient supervision and various forms of sloppy science to score highly on aggregate impact throughout all disciplinary fields. Researchers from the natural sciences and humanities also perceived nepotism to be of major impact on the aggregate level. The natural sciences regarded fabrication of data of major impact as well. The focus group interviews helped us to understand how researchers interpret ‘insufficient supervision’. Besides, the focus group participants added insight into sloppy science in practice. Second, researchers from the natural sciences and humanities added new research misbehaviours concerning their disciplinary fields
to the list, such as the stealing of ideas before publication. This improves our understanding of research misbehaviour, or ‘questionable research practices’ beyond the social and biomedical fields.”

(14)  The discussion should connect the findings clearly to the research questions of interest, and engage a more lengthy discussion from how they connect or diverge from past findings. For example, on page 19, mentoring references should be provided.

Response: We have now rewritten the Discussion in an attempt to tailor it towards the research question and existing literature and added some references to mentoring programs, see lines 441-467 in the revised manuscript with changes accepted.

“Somewhat surprising is the consistent recognition of insufficient supervision and mentoring. We would like to reiterate that we regard insufficient supervision a research misbehaviour in itself. Like many other research misbehaviours, insufficient supervision describes non-compliance with one of the professional norms in academic research (adequate mentoring).

Yet, it seems plausible that insufficient supervision could, in some cases, lead to the supervisees unintentionally engaging in sloppy science because they were not socialized well into responsible conduct of research (24). However, we believe that the influence of insufficient supervision may go further. If a supervisor fails to create a safe learning climate, this could lead to situations where PhD students do not feel confident to share their concerns about a mistake (e.g. in the data-analysis) or to oppose their supervisor’s interpretation. Similarly, Roberts and colleagues (25) put forth the speculation that when the supervisor creates an environment where only spectacular outcomes are valued, supervisees may engage in sloppy science because that yields the desired outcomes. Nevertheless, in our study we did not investigate the possible reasons for research misbehaviours and investigating this would require a different research design.

The amount of literature on supervision and mentoring differs between disciplinary fields. Mentoring received extensive attention in medicine (26,27) and substantial attention in psychology (28). Mentoring and supervision have primarily been used as tools to foster diversity by encouraging minority groups to stay in science and engineering fields (29,30), but received little attention in themselves. One exception is a study by Green & Bauer (31) that linked mentoring to science students’ success. In the humanities, mentoring was coined as a way to improve the workplace culture (32). Interestingly, in our study, participants from the humanities
expressed concerns about the lack of supervision altogether, or a supervisor who is in fact not an expert in the field. Natural sciences researchers recognized this, but added that bad mentoring or a supervisor mentoring too many PhD candidates can make group relations sour and ultimately slow down research.”

(15) If the authors think it worth mentioning, do you think that the academic environment in Amsterdam would influence responding and results? Or, are these findings potentially universal? How far might they generalize? This might be worth noting, as your readership is likely to be international in scope.

Response: The low completion rate presses us to be humble about generalisation statements. Besides, we explicitly asked focus group participants to concentrate on research misbehaviours that they had witnessed themselves, so this might also decrease the generalisability. Nevertheless, since some research misbehaviours were recognized across ranks and fields, we are tempted to think that sloppy science and insufficient supervision could be generalized to other parts of the Netherlands. That being said, future research should establish the extent of their true generalization. We have now elaborated on this in our Limitations section, see lines 541-546 in the revised manuscript with changes accepted.

“Finally, note that we explicitly asked respondents to focus on research misbehaviours that they had witnessed themselves, so this could decrease the generalisability of our findings so that they might not even apply to the population of academics in Amsterdam. Nevertheless, since sloppy science and insufficient supervision were recognized by academic researchers across disciplinary fields, it seems plausible that this may concern researchers outside Amsterdam as well.”

(16) The limitations note response bias is not a problem if the respondents are representative. Do you think your respondents were representative? Do you have any evidence or way to confirm this? The comment is correct, but are you saying you think this point holds in your study? This is unclear as written.

Response: We now added a brief section where we contrast our respondents with the available statistics on academic researchers in The Netherlands. This leads us to conclude that, since we were primarily interested in differences across fields, our sample – although relatively small –
seems fairly representative, see lines 483-495 in the revised manuscript with changes accepted. We now write:

“We assessed the representativeness of our sample in two ways. First, we looked at our population that consisted of academic researchers in Amsterdam from two universities and two university medical centres. Those two university medical centres comprised 53% of the population. Biomedical researchers constituted 56% of our sample, indicating a small overrepresentation. Second, we compared our sample to national statistics on researchers in The Netherlands. As there are no national statistics on academic researchers in biomedicine, we filtered biomedical researchers out of our sample for this comparison. National statistics indicate that 32% of researchers work in the natural sciences, 41% work in social sciences and 27% in the humanities. In our sample, we find 25% of researchers to work in the natural sciences, 51% in the social sciences and 23% in the humanities. This indicates a moderate overrepresentation of the social sciences researchers and a slight underrepresentation of researchers in the natural sciences and humanities.”

(17) The discussion seems to bring in a new issue, namely whether the items as written in the survey are clear to participants. This raises concerns about the survey itself, and while may be worth mentioning, should not be a key focus in the discussion.

Response: You are right that this point received too much attention in the discussion, we have moved the point to later in the Limitations section, it now reads as follows, see lines 530-540 in the revised manuscript with changes accepted:

“Finally, since the list of 60 research misbehaviours is not formally validated, it remains possible that survey items were unclear to participants. Nevertheless, the list was tried out at length through workshops and other types of informal review. Yet, especially researchers from the natural sciences and humanities mentioned research misbehaviours that seemed missing or at least substantially different from the list of 60, such as referees or editors that abuse their power to steal original ideas. Properly assessing the relevance of these new items would require translating the qualitative data into items and a representative sample from all disciplinary fields. To facilitate such an attempt, we provide an updated list of research misbehaviours (additional file 13) in which items are reformulated, included as explanatory examples or added as new research misbehaviours. Validation of such a list could be an avenue for further research.”
Overall, I feel most of my comments relate to the need for greater rationale and detail in the manuscript. The issue of greatest concern to me is the measurement of "importance"—how this variable was constructed and what it means conceptually, along with the sample sizes for individual behaviors. It is important to strengthen this information in particular. In terms of manuscript writing, take your reader on a journey through each section, bearing in mind that we have no background context for your design, analysis, and interpretation choices. I appreciate all the additional supplements, but I do not advise leaving your reader to do too much work to understand your approach. Best wishes in this and your future work.

Response: Thank you for your elaborate review and specifically for pointing out the issues in our manuscript that needed to be solved. We hope to have provided a more compelling explanation of our constructed variable and clarified the use of the missingness-by-design setup of our survey. We added a number of extra connecting sentences to improve the legibility of the manuscript.

-------------------------------------------------------------------

Thank you again for the thorough reviews. Should our responses provoke any questions, please do not hesitate to get in touch.

Also on behalf of my co-authors,

Yours sincerely,

Tamarinde Haven

References


2. Anderson MS, Horn AS, Risbey KR, Ronning EA, De Vries R, Martinson BC. What do mentoring and training in the responsible conduct of research have to do with scientists’


