Response to Reviewers

PCI-RR submission entitled “Mapping Cross-Disciplinary Perspectives on Responsible Conduct of Research: A Delphi Study”
26 Mar 2024

Reviewer 1

Thank you for the opportunity to review this Stage 1 Registered Report for PCI-RR. I enjoyed reading this manuscript proposing a Delphi study to map responsible conduct of research. I think it is important and interesting work, but I have some concerns regarding the mapping of research aims and methods and some other aspects of the methodology. Please see below for my comments and suggestions.

To contextualise my review: I am a psychologist and metascientist who has worked on several aspects of open science (replicability, reproducibility, generalisability, big team science, uptake of open science practices in undergraduates and researchers, and diversity in [open] science). However, I do not have any experience in mixed methods research, Delphi studies, or existing frameworks underpinning individual and institutional codes of research conduct. I have carefully read the Stage 1 manuscript, but have not read all the referenced work. I will therefore provide my review from this positionality, and hope that in combination with other reviews the authors will have feedback on all aspects of their proposed study.

Thank you for reading our protocol and for providing suggestions for improvement!

1.1. I was surprised to read that the idea is to make one framework, and for the “bullseye” and more central circles to include facets that span multiple disciplines and the outer rings to contain aspects that are more “niche”. I’m not sure that this is in keeping with the aim of allowing for disciplinary differences. The fact that the majority of disciplines agree with one particular aspect of responsible conduct of research doesn’t mean that it’s necessarily the most important part of RCR, and similarly, something which might be very specific to only some disciplines may be one of the most important aspects of RCR for those disciplines. In addition, if there is only one person participating from each discipline, having a rule that “items where only one person considers them important will fail validation and be excluded from the framework” seems odd if this is a really key aspect of RCR for this discipline.

Thanks for pointing this out. We realise this was not sufficiently clear in the original manuscript. Upon reflection from reviewer comments and further discussion, we have decided to separate the construction of the framework from this Delphi project. This comment made us realise we had not communicated our idea of the visuals very well; note that we did not mean to imply that those dimensions that would end up in the “bullseye” should be considered the most important, but rather that they are the most widely shared. We had the bullseye model in mind when imagining what we might do with the outputs of the Delphi, but comments like these helped us clarify that our primary goal in the Delphi is simply to collect insights from multiple disciplines, full stop. The question of how to visually present them in a framework is a separate step. We have revised the manuscript to reflect this.

Additionally, We have since updated the procedure of the Delphi process so that no dimensions will be "dropped" on the basis of their panel-judged importance, and we will simply report the judged importance of all dimensions in the study.

1.2. I am not sure about the feasibility of the snowball sampling method of requesting "that each existing panelist provides us with other possible participants". This feels like it may not work as well for getting one participant per discipline (i.e. panellists may be more likely to know people within their own discipline). For this reason, I might suggest instead using an approach that is more similar to the process sometimes used when seeking reviewers for a manuscript or speakers for a symposium, whereby you ask for recommendations from people who decline the invitation for people who could take their place. This could also help with issues of diversity (see next point), if you had some text to include something like "if you decide to decline this opportunity, we would be very grateful if you could recommend someone that fulfils X criteria, prioritising researchers who are Y".

Thank you for pointing this out, as well as providing a suggestion to address it, which we have kindly followed.. We have deleted the plans concerning snowball sampling, and added the following line to our protocol:

"In addition, we will also ask for recommendations of other participants who fulfill the above-mentioned criteria, from those who declined to participate. Potential participants pointed out by declining participants will be vetted as to whether they meet the inclusion criteria before they are invited to participate."

1.3. I would like to see more detail regarding the diversity of participants, particularly how you will prioritise diversity of participants and what information will be collected about them. It is clear that the proposed methods (if successful) will result in disciplinary diversity. How do you plan to "work to ensure that as much of that diversity as possible filters into the final sample" and ensure the sample is "as diverse as possible"? Which aspects of diversity will you be prioritising (only gender and geographical region are mentioned), and how? In addition, will readers have access to information about the diversity of the participants (for example: methodological background, career stage, ethnic background, et cetera)? I know you plan to collect some of this information already, but will it be shared?

This comment reflects that we have not been as clear as we could have been regarding our meaning of diversity. We were referring to diversity in terms of disciplinary (our first concern) and regional diversity (within the limits of the UK and EU primarily). All mentions in the protocol text to other dimensions of diversity have been removed to reflect this. In addition, we have added the following paragraph to explicitly explain this position:

"We also note that despite our goal of developing output that is diverse in terms of the scientific disciplines that are represented in it, this output will represent only a limited selection of countries, regions and cultures. While the broader project within which this study is situated concerns RCR in the UK and regions of Europe, and a Euro-centric approach is appropriate to those ends, we emphasise that our findings will be produced with the input of a largely Western participant sample. We discuss the impact of this on our findings further in the limitations section in the discussion."

1.4. You say that you "will continue to contact possible candidates until we receive consent to participate from 40 people" but does this mean making replacements when no one from a particular discipline says yes? Otherwise you might end up with a sample biassed towards the disciplines interested more in research integrity.

Thanks for pointing this out. Indeed, in the first protocol we had no safeguard to protect from disciplinary biases arising from one (set) of disciplines being overrepresented in the Delphi panel. Upon consideration, we decided to add both a minimum number of disciplines represented (15) and a maximum number of respondents within a discipline (3). Please note that we also lowered the minimum panel size at the start of the Delphi process to 30 panellists.We have rewritten the appropriate section of the manuscript to explain and justify these decisions (see "Panel Size" under the Method section). Relevant text is pasted below:

"*Recommendations and empirical studies on Delphi methods vary on desirable sample size. For instance, 20-30 participants seems to be a sufficient panel based on Melander (2018; note that this is more than typical consensus Delphi panels tend to require), while Turoff (2002) suggests that between 10 and 50 panellists is sufficient for a dissensus Delphi. Choosing a minimum number necessarily contains an arbitrary factor, as well as a pragmatic one. Considering our goal of disciplinary diversity in the expert panel, we elected to use a minimum on the higher end of the average that these two sources suggest. As such, we have set the minimum panel size at the start of the process to be 30 panellists. To avoid disciplinary bias in our sample, i.e., where a disproportionate amount of experts would have a background in a specific discipline, we also decided to include a minimum amount of disciplines present in our sample before we start the Delphi study. As such, the starting panel must represent a minimum of 15 disciplines, with no more than 3 participants from one single discipline.*"

1.5. As this is a Stage 1 RR, I would love to see some open materials included in the revision so that these can also be reviewed before data collection. This could include for example a visual example of what the RCR map could end up looking like, templates of emails you will send to recruit participants, participant study information, the initial reference document that will be provided to participants, demographic questionnaires, pilot data, et cetera.

We apologise that supporting materials were not part of the original submission. This was an oversight on the part of the authors. Please find the participant information and ethics forms, questionnaire, reference document, and other supporting materials on the OSF page for this project at: https://osf.io/7n8bh/. (Note that we do not intend to add a visual example of the eventual RCR framework, as we have separated that effort from this study, as outlined in 1.1. Also, A final version of the recruitment e-mail is not yet available, but will be uploaded to the project's OSF page before the recruitment starts, as per our data and material sharing policy. We will likely craft individualised versions of recruitment emails, as this can aid uptake. We will explain that we are running a Delphi study to investigate RCR across disciplines, and that we have contacted them as they have expertise in the field. We will invite them to participate both for the good of the community and the early career researchers involved in this project (as well as any intrinsic interest).

Reviewer 2

This proposed registered report is part of a larger project aimed at developing a new, inclusive framework for responsible research. I had a bit of a chuckle at footnote 3 about the distinction between people involved in scientific reform vs. those involved in RCR, as I am firmly part of the former group. Accordingly, my comments here are coming from outside of the RCR ecosystem, and thus at times may reflect some level of ignorance. Alternatively, I think my outsider perspective may be helpful for improving some aspects of the proposal.
My feedback pertains to two broad issues that I think need to be addressed before the proposal can be assessed further: properly motivating the study and providing a sufficient level of detail for a registered report. I will take each, in turn.

Thank you for taking the time to evaluate our study protocol, and for providing useful feedback.

2.1. The Introduction section did not provide me with a strong understanding of the authors' framework, how it improves upon the existing frameworks, and thus why this project is necessary. As written, it assumes a lot of common ground knowledge that a naïve reader will not have. The many different terms and acronyms-- ELSA, RRI, RR(I), RCR—make it all the more difficult to follow. The authors need to take some more time explaining these different terms and frameworks and their interrelations. Doing so would then provide a foundation for the reader to understand the need for a new framework. The authors make reference to their newly developed framework, including reference to "dimensions," but do not explain the framework in any detailed way. For example, the authors state that there is a need for, "a new RCR framework that balances breadth and specificity with feasibility and practicality," but don't explain how existing ones fail at this or how theirs achieves it. I understand that the current project is just one piece of a much larger project, but nevertheless this paper needs also to stand alone.

Thanks for this point. For the sake of clarity, we have reformulated this specific study to be completely stand-alone; separate from both the scoping review that precedes it and the practical framework that will follow it. We also redrafted large parts of the introduction to supply more background information, and clarify and contextualise the acronyms named. In essence, the redrafts to the introduction can be summarised with the following (new) excerpt from the abstract:

"Currently, many approaches to research and training in RCR are either generalized across all disciplines, at a high level (e.g., international frameworks on research integrity) or at the other extreme, discipline-specific. Relying on the expertise and knowledge of a carefully selected multidisciplinary panel of RCR scholars and practitioners, this Delphi study aims to expand the current (underspecified) frameworks of RCR to develop a more diverse and comprehensive concept of what constitutes RCR across disciplines, along with a mapping that captures this updated understanding"

2.2. A somewhat related concern is that I was not clear on what the context for this work was, or to put it another way, who the audience would be. The authors rely on the REF units of assessments and reference the European Commissions' Frameworks Program, which suggest that the context is the UK and/or

Europe, but other aspects of the text suggest that this is a global framework. Either is fine, but it should be explicitly stated. As part of this, it would be helpful to know how this framework would be put into action. Who will enforce this, or who will pay it any mind? How will this framework be successful? Who will take it up?

Thank you for this comment. This is something we should have been clearer about in the original protocol.

Regarding who will use the framework: As mentioned in the previous point, we have since reformulated the outputs of the current study to be separate from the framework. However, for the reviewer's information we will briefly describe our aim with the framework. This study is designed to help the author team gather diverse perspectives on RCR from multiple disciplines that we can later use to develop a framework for our own use to guide the embedding of responsible research practices in communities of practice across both the UK and Europe. As such, this is not a prescriptive framework that will be 'enforced', but a descriptive framework that we will use to help both local research groups and larger organisations, such as institutions and existing communities of practice around particular practices, see the overall context of how RCR manifests differently across disciplines, and therefore better understand how to create initiatives to apply, embed and improve RCR practices. For instance, seeing the overall picture may help researchers and organisations appreciate how RCR practices may need to evolve within a field, how examples may be gleaned from similar fields, or where various initiatives for improvement may be best targeted across disciplines.

We have made the broader context of the project a larger part of the manuscript so that readers can contextualise the use of the framework that we intend. We include the following paragraph in the introduction section:

"Lastly, it is also important to consider the role of this study as it forms a component of a larger multi-year project, which aims to broadly develop a diverse understanding of how RCR is conceptualised and applied across different research disciplines. The conceptual mapping we will have co-produced with the help of RCR experts during this Delphi study will form a scaffold for interaction with communities of practice in the project's latter half, helping them to contextualise where particular disciplines and practices sit in relation to others in the overall ecosystem of RCR. Its broad remit will also help to spotlight the perspectives of disciplines that have been more peripheral in discussions and evaluative frameworks on RCR so far. Thus, we aim for our mapping to stimulate a more nuanced understanding of cross-disciplinary conceptions of RCR within the communities that work to embed practices in situ. While our output can be used more broadly to assist other interested entities (such as individuals or research groups) in conceptualizing and applying RCR principles for their own needs, that is a secondary purpose. Therefore in our aim to generate a tool that presents a wide perspective on the RCR sphere, we have cast our epistemic net broadly."

So, primarily, this framework is for internal use to achieve the greater goals of the wider project. Ideally, we develop a framework that is also usable outside of the project, and outside of our remit (i.e., outside of the UK/EU) but that is a secondary consideration.

Regarding geographical context: This comment also reflects that we have not been as clear as we could have been regarding our meaning of diversity. We were referring to diversity in terms of disciplinary (our first concern) and regional diversity (within the limits of the UK and EU primarily). All mentions in the protocol text to other dimensions of diversity have been removed to reflect this. In addition, we have added the following paragraph to explicitly explain this position:

"*We also note that despite our goal of developing output that is diverse in terms of the scientific disciplines that are represented in it, this output will represent only a limited selection of countries, regions and cultures. While the broader project within which this study is situated concerns RCR in the UK and regions of Europe, and a Euro-centric approach is appropriate to those ends, we emphasise that our findings will be produced with the input of a largely Western participant sample. We discuss the impact of this on our findings further in the limitations section in the discussion.*"

2.3. My second broad concern is the lack of specific detail in places, which is expected for a registered report. I like the idea of using the registered report for developing a framework vs. testing hypotheses, and I understand that this project is largely exploratory, but you should still nail down as many details as possible and avoid vague decision criteria. A few examples of unclear procedures are as follows: The sampling strategy is unquestionably complicated for this kind of project. However, the authors should have clear criteria for what would constitute a sufficient sample for the project to proceed as agreed. This is critical given that the IPA comes with a guarantee* to publish the final paper. The authors state that they "hope" to retain a sample of 20 at the final round, but this is not a commitment. Moreover, nothing in the procedure precludes this sample of 20 from coming from a small slice of disciplines and/or countries. The authors should be much more specific about what the minimum acceptable sample will be, both in terms of numbers and characteristics. Statements that the authors will monitor the diversity of the sample until they are content should be avoided in favor of more formal criteria.

Thank you for sharing these concerns. We agree that our original protocol was insufficiently specific, and we are very grateful for you and your fellow reviewers for pointing out where improvements were necessary. Our sampling and recruitment strategy, our Delphi methodology , and our expected results (essentially, the entire Methods section) have been redrafted with much greater detail, including explicit inclusion criteria, stopping criteria, minimum sample size, format of reporting the results, and more.

2.4. The dissensus approach is a strength of the project, but it was unclear whether a dimension mentioned by a single participant would be included in a subsequent round, or whether there would need to be a higher frequency of mentions.

We have since changed and clarified our Delphi process (see "Delphi procedure" heading in the Methods section) so that suggested additions are only solicited in the initial modification phase and after this, no dimension will be "dropped" on the basis of the panel-judged importance. Dimensions will only be removed from subsequent rounds if they are considered "stable", at which point they will be reported as such, and for which we have predefined a measure of stability. For your convenience we have pasted the relevant paragraph below:

**"*An important aspect of the Delphi process is the concept of stability, or when we consider the answers to be similar enough between two or more subsequent rounds that we can consider the answer "settled" or "definite". Since panel responses can vary greatly between rounds - as per the explicit aim of the Delphi process - it is important to assess whether the panel's response on any given dimension is still developing or whether it can be considered settled. In fact, different authors have argued that assessing the level of consensus in a Delphi study is meaningless without having assessed stability of responses, since the response may not be an accurate reflection of the conclusive judgment of the panel (Dajani et al., 1979, Scheibe et al., 2002). To reduce participant burden, where stability is reached, the item will be considered 'set', and not feature in subsequent rounds.*
We use a simple metric for stability: for each dimension, we will take the absolute value of the change in ratings for each participant. If the mean of these absolute-value change scores is less than the equivalent of 1 point on our rating scale (i.e., 16.66% of the total breadth of the rating scale), we will consider the dimension stable. This tracks closely with the recommended cutoff for stability of 15% difference recommended by Scheibe et al. (2002), which is based on an empirical estimation of the random change between rounds. However, we will also temper this quantitative stability judgement with qualitative analysis: if the qualitative data contain novel arguments for the importance or unimportance of a dimension that we have reason to believe may sway the panel substantially in the following round, we will not consider the dimension to be stable."**

2.5. I appreciated that the authors included the IQR and median values for the different levels of the "bullseye," although the low ratings will lead to the dimension either being placed on the outer ring or being dropped entirely—these are very different outcomes that should have clear criteria. Moreover, how will the authors determine that there is "no change" in these ratings across rounds? What amount of change constitutes a level of meaningful change?

As mentioned above, we have since separated the Delphi study from the creation of the framework, as the framework will likely need significant further consultation in a separate process. As the study currently stands, we plan to simply report and discuss the results from the Delphi study (see the final, revamped section, "Expected results" for details), without pre-specifying how these results will later inform any framework. The role that the Delphi study plays in the development of the later framework is hopefully now clear in the manuscript.

Regarding our plan to determine whether "no change" has occurred between rounds, we now have operationalised this fully, and motivate and describe it in the section titled "Stability." A relevant paragraph of that section is pasted below:

"We use a simple metric for stability: for each dimension, we will take the absolute value of the change in ratings for each participant. If the mean of these absolute-value change scores is less than the equivalent of 1 point on our rating scale (i.e., 16.66% of the total breadth of the rating scale), we will consider the dimension stable. This tracks closely with the recommended cutoff for stability of 15% difference recommended by Scheibe et al. (2002), which is based on an empirical estimation of the random change between rounds. However, we will also temper this quantitative stability judgement with qualitative analysis: if the qualitative data contain novel arguments for the importance or

unimportance of a dimension that we have reason to believe may sway the panel substantially in the following round, we will not consider the dimension to be stable."

2.6. The authors reference an, "Initial Reference Document," which appears to be central to the study, but was not included for review or discussed in any detail (related to my first broad point). This, along with other study materials, should be included for the next round of review.

Copied from response to reviewer comment 1.5: We apologise that these materials were not part of the original submission. This was an oversight on the part of the authors. Please find the participant information and ethics forms, questionnaire, reference document, and other supporting materials on the OSF page for this project at: https://osf.io/7n8bh/. The initial reference document in particular can be found at https://osf.io/jrf47.

2.7. As a final, somewhat distinct point, the authors indicate in parentheses that the Delphi method is sometimes considered to be a sequential mixed methods design. How exactly this is the case should be described in text.

We understand that this may have been needlessly confusing. Since we have outlined clearly what our Delphi procedure will be, we have simply deleted the text in parentheses.

Reviewer 3

This is a highly interesting study that can help the academic world better grasp responsible research. In general, the plan has been very carefully crafted and the design is promising. In terms of my reviewer point of view, although I'm familiar with Delphi studies and have participated in them many times, I haven't published any myself so my pragmatic know-how can be lacking in this regard. Below, I try to provide feedback that is helpful for further improving the plan. I list the comments one by one to make it easier to read. My writing style is sometimes a bit blunt so please do not see it as an adverse signal, I really like the study plan.

We are glad to hear that you're generally positive about our study plan, and are very grateful for your detailed and helpful feedback.

3.1. Perhaps my largest comment concerns the reference document, which is the starting point of the work and is narratively described in the MS. Because the data and information for the document are already available, I was surprised not to find this document as a supplement. I believe Stage 1 RR would have been an excellent opportunity to gather external feedback for this document, the structure of which represents the core of the study. One can comment on the narrative descriptions too, but I personally find it a challenging without seeing how the document is fully structured. I would really encourage attaching this document for the review round 2 (although it might be already too late because reviewers / recommender might find it not practical to suggest major revisions anymore at that point).

We apologise that this document was not part of the original submission. This was an oversight on the part of the authors. Please find the reference document on the OSF page for this project at: (https://osf.io/jrf47).

3.2. Related to the above, it remains a bit unclear how the reference document was or will be constructed. I understand the review serves as a basis, but there are also mentions of interviews, thematic analysis, etc. This seems like a gap methodologically, as there is no further information. E.g., how many interviews were carried out, what were they like (questions list as a supplement?), are these data open or are there reasons for not sharing, etc. I was also unable to find out what kind of thematic analysis was applied (there are dozens of different TAs!), who was involved in that analysis, or what that analysis process was like in general (are any of the coding materials shared?). I don't want to unnecessarily complicate this study which is already going to be a laborious enterprise, but it would also be unfair to leave this central element uncommented. I will be happy with many explanations or solutions, but adding more information about this step would be necessary IMO.

Thank you for pointing this out. Indeed, such detail should be made clear for readers. We have added the following paragraph to the manuscript, briefly outlining the data collection and analysis steps that were taken to come to the initial reference document. As pointed out in that text, we have also added an appendix (https://osf.io/jrf47) that describes the analysis process in greater detail.

"Before drafting the reference document, i.e., a proposed list of RCR dimensions, the authors conducted a scoping review of the existing RCR literature (Field et al., 2024) and interviews with RCR scholars and practitioners (see the interview guide at https://osf.io/xv98y). The interviews consisted of 10 one-on-one interviews and one focus group which included two moderators and eight participants. The articles included in the scoping review and interview transcripts were subject to a thematic analysis, conducted by SMF, which involved coding topically salient sections of text and combining related codes into themes (see the appropriate appendix for a detailed description of this process at https://osf.io/jrf47). This analysis generated a series of overarching themes which reflected salient dimensions of RCR from the literature on the topic. These dimensions are the core of the initial dimension list, to which dimensions taken from existing, older RCR frameworks were added (i.e., the Singapore Statement on Research Integrity, the Australian Code for Responsible Conduct of Research, and the All European Academies European Code of Conduct for Research Integrity)."

3.3. The introduction is clear and explains the background well to someone who hasn't been directly involved in related program development. That said, the topic appears to be tackled largely from a Western perspective, and I don't see many cross- or multicultural aspects addressed (before footnote 2). Different countries and cultures have different ethical and legal approaches, and it would be good to discuss this diversity explicitly in the introduction. I know there are limits to content, so I leave it for the authors to negotiate to what degree they wish to integrate this aspect in the study.

This comment reflects that we have not been as clear as we could have been regarding our meaning of diversity. We were referring to diversity in terms of disciplinary (our first concern) and regional diversity (within the limits of the UK and EU primarily). All mentions in the protocol text to other

dimensions of diversity have been removed to reflect this. In addition, we have added the following paragraph to explicitly explain this position:

"We also note that despite our goal of developing output that is diverse in terms of the scientific disciplines that are represented in it, this output will represent only a limited selection of countries, regions and cultures. While the broader project within which this study is situated concerns RCR in the UK and regions of Europe, and a Euro-centric approach is appropriate to those ends, we emphasise that our findings will be produced with the input of a largely Western participant sample. We discuss the impact of this on our findings further in the limitations section in the discussion."

We must emphasise that our scope is intentionally narrow. Our author team is embedded in these regions and are best able to develop RCR research from these perspectives. Although a broader perspective would be good in general, it is outside the scope of the project for this to be part of its design. Moreover, as another reviewer has pointed out, this study is ambitious enough even given the narrower cultural focus, and so feasibility has been a priority for the project also.

That said, as you point out in the second part of your comment, such diversity should be more explicitly discussed. We will address this issue at length in the discussion/limitations section, as it is undoubtedly an important one.

3.4. Regarding the participating experts, I would prefer to have clear inclusion/exclusion criteria. There's currently a general description (p. 9-10), but it would be good to know explicitly what criteria the 95 listed experts meet and who have been excluded (for what reason). This is mainly a cosmetic note, as the information is distributed there to a large extent already. Having details like this noted at Stage 1 will add transparency to the process, even though changes may have to be done later in data collection or analysis.

Thanks for this comment. We have redrafted the sampling and recruitment strategy in much greater detail, including explicit inclusion criteria. These are too long to be pasted directly here, but all information can be found in the sections "Participant sample," "Panel size" and "Recruitment strategy" in the Methods section.

3.5. There is an important note about representation (p. 12) among experts. Especially related to my comment #3, I think it would be critical to somehow ensure significant representation of non-Western experts if the project aims at universal findings. Alternatively, it would be totally ok to clearly focus on specific, selected regional expertise. I just see a risk here that the results suggest a global concept when only a small proportion of experts come from countries that may have different cultural perspectives and produce half of the world's research (China, Japan, India, etc.). Again, I leave it for the authors to negotiate how to tackle this; the most important thing is that the authors are aware of it and will be able to consider the aspect critically in their analysis and reporting at Stage 2.

Thank you. We have made our regional framing more explicit in the manuscript, as noted in our answer to point 3.3.

3.6. I really like how the multidisciplinary dimension has been considered and how comprehensively it

has been integrated in the design. As an interdisciplinary researcher myself (having worked across all panel areas A-D), I was thinking whether a separate panel for interdisciplinary experts could even further improve the design. Such people might have distinct viewpoints. But you can also fully ignore this comment; I understand it could unnecessarily add to the already-hefty load of work.

This is a great idea. A follow-up study involving such a panel would be a great way to strengthen the findings, but as you intuit, we are facing a feasibility issue and would not be able to manage that in addition to the regular panel process.

3.7. A technical comment: will there be any measures for careless responding or other data quality checks? It's less common to have data issues in a Delphi, but it would be good to somehow plan to control data beforehand since it's an RR.

Good point. We address this in a paragraph under the "Feedback reports and analysis plans" heading:

"Since we selected our participants in this study on the basis of their individual expertise – which surpasses our knowledge of their disciplines – we will not perform any stringent quality checks on the content of the data, quantitative or qualitative. Data quality should be aided by the fact that participants are encouraged to write down free-text justifications to every question, which should prompt them to answer questions thoughtfully. If we notice suspicious patterns in the data, however (such as all items answered with the same choice) we will contact participants individually to check that they meant these."

We had originally included an "I don't understand" option in the ratings of importance for each dimension, but decided to remove this, as we felt this would either not be used much or would encourage panellists not to give ratings. We will therefore force panellists to give ratings of importance, but if they do not understand the question they can indicate so in the text box below.

3.8. Also, I didn't see data or document version sharing discussed anywhere in the MS. How will data sharing and document development be managed? I am assuming that everything will be shared as per TOP guidelines.

Thanks for pointing out that this information was missing. Indeed we plan to share all data and research materials. We have included the following paragraph:

"We believe in the importance of data sharing, both from the perspective of accountability, as well as the potential re-use of our data. As such, we will share all data and analysis, guided by the TOP Guidelines. We will do this by making the feedback reports openly available in a suitable repository. These feedback reports will include the pseudonymised 'raw' data, both quantitative and qualitative, along with our analyses of this data, subject to any redactions by study participants for their privacy. Final versions of all study materials will be uploaded to this page before the study starts."

3.9. Considering drop-outs, I am thinking whether it would make sense to recruit new participants at later rounds if the N drops too low. This is hardly optimal, but to me it seems like a better Plan B versus

hypothetically going forward with a very small participant group.

Thanks for the suggestion. As you say, it is not optimal methodologically speaking, but having too few participants is also a methodological challenge we may also face. We have included this option in the "Panel size" heading of the Methods section:

"Should enough participants drop from the study such that the total N drops below 15 or the amount of disciplines represented drops below 10, we will resume recruitment until these minimum thresholds are met. In this case we will still maintain a maximum of three participants per discipline. While this is not ideal methodologically speaking, our goal of having the input of a diverse and large enough panel is more important than having continuity across rounds. Any attrition and replacement will be thoroughly and transparently documented in the final manuscript."

3.10. Since there is a lot of flexibility in the design and decisions between rounds can be made by unforeseen motivations (p. 18), I think it would increase transparency to add brief notes on positionality, i.e. the authors' own core disciplines and perhaps some perceptions of what they personally consider important in RCR. When we then see the decision tree at Stage 2, it will be possible for readers and reviewers to reflect on those decisions and results against the stated positions. If this suggestion feels unfitting, it can naturally be rebutted.

Thank you for this suggestion – we agree! Although in the updated manuscript you will find that many research decisions to be made in the Delphi process have now been formalised and operationalised, we agree that adding positionality statements will provide transparency in what remains a complex and value-laden process. We have both included a collective positionality statement in this updated manuscript, as well as drafted an appendix containing our individual positionality statements.The in-text collective positionality statement is copied here for your convenience:

"Finally, we wish to be transparent about the contributions that we, as individuals and as a team, approach the subject of RCR. This allows the reader to evaluate our decisions with personal context in mind, given the flexibility that exists in our design, especially as the dimensions are developed between rounds, and as the framework is built. Our team comes from a background in science studies, metascience and research integrity, and we have previously published on the topics of epistemic diversity, responsibility and quality (e.g., Field & Derksen, 2021; Muller & de Rijcke, 2017; Penders, de Rijcke & Holbrook, 2020; Penders & Goven, 2010; Valkenburg, Dix, Tijdink and de Rijcke, 2021; Van Drimmelen et al. 2023). We have also previously published the scoping review on which this Delphi study directly builds. As a result, we are aware of the literature on RCR and adjacent topics. Unavoidably, our decisions will be rooted in this knowledge, from the initial choices we have made to develop this study, to the choices we will collectively make as we co-construct an RCR framework with the input of our experts. In our supplemental materials on OSF (https://osf.io/prvds), we provide individual statements about our positions in relation to the present study, structured using orienting questions proposed by Barry et al. (1999) and Olmos-Vega et al. (2023) , to further highlight our link to the research our group is conducting."

Reviewer 4

4.1. Title: Consider a wording change in the title (p1) to clarify that you are focusing upon academic practice, or research practices, or practicality, rather than applied research (which "research in practice" could be misinterpreted to be).

Thank you for the suggestion. The title has been changed to: "Mapping Cross-Disciplinary Perspectives on Responsible Conduct of Research: A Delphi Study"

4.2. The use of RCR rather than RR as the central term could perhaps be justified or discussed further – I think in this space there are lots of overlapping terms and it might be fruitful to provide a clearer definition of what the concept you refer to includes and excludes. Given this fields' propensity to allow concept (concept) creep, particularly when acknowledging the fields of ethics and integrity, it'd be fruitful to make a more decisive statement.

The point about concept creep is a good one, and certainly, we don't intend to add unnecessary new terms to the dialogue on RCR/RR(I). Part of the motivation for using RCR as opposed to RR(I) was to recognise that we are emphasizing the conduct aspect of RR(I) and leaving other issues like broader policy (involving institutions and funders, etc.) out of the discussion. The following text was added to more clearly motivate the concept we work with in the study:

"Narrowing the scope of the definition for the current study, we consider responsible conduct of research (RCR) to be a topic that requires the synthesis of many disparate aspects. While it is distinct from concepts such as research integrity (RI) or responsible research and innovation (RRI), it is no doubt closely related. We argue that RCR casts a broader net than the typical definitions of research integrity, that is, promotion of confidence and trust in research and the research process. This broader remit of RCR includes dimensions that overlap with those of RRI, such as the responsibility research has for honest and transparent dealings with citizens and society. However, while conceptualisations of RRI typically include impacts of technological innovations and research output on society, RCR concerns the subset of dimensions or responsibilities relating to the activities involved in conducting research."

4.3. "For example, reproducibility is a concept that applies to quantitative disciplines, but less so qualitative disciplines and the social sciences, and even less so in the humanities" (p2) – I would revise the sentence structure for clarity.

We agree that this sentence was quite awkward. Since we concluded that the sentence was not necessary to convey our point, it has been deleted.

4.4. "to reorient scientific research [practices] to make it [them] more effective and – crucially – more ethical and self-aware" (p3).

The sentence has been amended: "*to reorient scientific research practices to make them more effective and – crucially – more ethical and self-aware*".

4.5. "Von Schomberg points out that there is no agreed-upon definition of what RCR is; rather, it holds an invitation to discuss what RCR as a top-down signifier might in fact denote, in relation to the disciplines and research processes it engages with." (p4)– wording is a little awkward here.

Reworded as follows: *"Von Schomberg points out that there is no agreed-upon definition of what RCR is. Rather, he invites readers to consider what RCR as a top-down signifier might denote, in relation to the disciplines and research processes with which it engages."*

4.6. I read your preprint "Exploring the Dimensions of Responsible Research Systems and Cultures: A Scoping Review" with great interest and noted how little of the core conclusions (relevant to this particular manuscript) were discussed. I appreciate the need to minimise repetition across manuscripts, however I feel like the preprint provides a more comprehensive grounding for the justification of the proposed work and so could perhaps be discussed in a little greater detail. This links to the next point.

Thank you for pointing this out. The choice of how much to recapitulate from the scoping review in this protocol was indeed a reasoned one, as we hope we do not wish to make the manuscript overly long, and any readers wanting more specifics can access the scoping review manuscript openly. However we can appreciate the request for more grounding for the justification of the proposed work. We have now augmented the manuscript with several extra blocks of text throughout, including the text quoted in our answer to 4.7 below. We hope this addresses this point.

4.7. As a whole, this introduction feels a little light when attempting to convince the reader for the need of a new framework/approach and thus a slightly richer discussion of the existing frameworks and content of such might help situate this work a little more clearly. A more comprehensive mapping of existing frameworks and their scope of relevance (i.e., which fields they can successfully be applied to) might provide a more robust justification for a centralised/singular framework.

Thanks for this point. For the sake of clarity, we have reformulated this specific study to be completely stand-alone; separate from both the scoping review that precedes it and the practical framework that will follow it. As such, we have also revised large parts of the introduction.
The following passage has been added near the end of the introduction, regarding the need for our Delphi study:

"We aim for our mapping to fill the gap between the two extremes that existing conceptualisations of RCR tend to fall under: either high-level frameworks designed to be universally applicable across all disciplines (e.g., the Singapore Statement on Research Integrity, the Australian Code for Responsible Conduct of Research, or the All European Academies European Code of Conduct for Research Integrity), or prescriptive guides tailored to the practical instruction of researchers within a specific discipline or field (e.g., RCR training designed for members of a university department as part of a degree or continuing professional development, or mandated by funders such as the National Institutes

of Health, or guidance from discipline-specific learned societies such as the Society for Improvement of Psychological Science)"

4.8. The development of a single framework itself might be considered to be too big of a demand and there is a fair risk that this project can't deliver despite the teams' best efforts – to provide something which is sufficiently detailed to be practical and helpful, whilst also diverse enough to be relevant to all fields. Is a singular framework where some components are irrelevant to certain fields more preferable to various discipline-specific frameworks? I remain receptive but skeptical about the potential for the proposed work to achieve this goal and encourage the authors to reflect upon the justification presented for this goal.

We recognise that without further explanation that this may well seem to be the case. We have amended the text to add explanation in this regard– please see response 2.2, above. As that response indicates, this framework is not intended to be 'everything to everyone,' nor a practical, prescriptive guide, but rather a descriptive mapping that will help individuals and groups working on improving RCR practice to better understand the context of how RCR manifests across different fields. This should show that our goals are not quite as ambitious as first it may have appeared. Naturally, if we were going to develop a prescriptive framework that is novel and greatly improves on everything else that exists, *and* to be used widely, we would need more time and to do more work. Adding extra context relating to the broader project as we have in the introduction helps the reader understand our ambitions as they relate to other elements of our work, rather than just as stand-alone outputs.

4.9. More details on the 'previously devised reference document' (p6) might be of benefit for the reader, and could be included as part of the appendix and open materials of the study.

We apologise that this document was not part of the original submission. This was an oversight on the part of the authors. Please find the reference document, as well as information on how it was created, on the OSF page for this project at: (https://osf.io/jrf47).

4.10 The use of delphi methods are well-suited to the aims and outcomes of the project, and whilst it will be useful in equalising voice, it has been well-designed with a dissensus approach to acknowledge that consensuses/majorities are unlikely given the broad ambitions of the project.

Many thanks for this. We agree!

4.11. The criteria of "importance" (p7) for which participants rate RCR dimensions could be elaborated upon.

We agree. We have now included a block of text addressing this issue:

"*The survey will present each proposed dimension consecutively, and will ask the participants how important they consider the dimension of RCR to their specific discipline on a 7-point Likert-type scale ranging from "very unimportant" to "very important". In addition, the participants will be encouraged*

***to motivate their answers in a textbox with unlimited characters, though a motivation is not required to move to the following dimension. Note that, other than a brief explanation as part of the initial survey instructions, we will not attempt to define to the participants in any great detail what "important" means. Although in general it is advisable to be as precise as possible in elicitation of survey measurements, we believe that in this case trying to prescribe particular aspects of the concept of "importance" would counterproductively narrow our measurement, when an intuitive, broad understanding of the word may more closely capture the essence of what we wish to measure, i.e., the sense that a dimension "matters" in that discipline.*"**

4.12. It would be useful to understand how you will negotiate the jingle-jangle, the nomenclature, for a range of ideas like integrity where there is already vast proliferation of broad definitions, models and terms. There is a risk that this study can lead to a list with lots of broad and inter-related ideas with little method to differentiate between semantics and content.

It is a very good suggestion to set up some safeguards to avoid getting stuck in a semantic quagmire. We follow Karl Popper's adage that arguing over definitions is a big waste of time, but it is hugely important to clarify what you are talking about. As such, we drafted the following paragraph, to indicate that we acknowledge that the names and definitions of the dimensions we drafted are not to be taken as prescriptive definitions, but practical tools for this particular study.

"***It is important to note that the choice of dimensions, and their respective definitions, are not intended to represent an authoritative list of dimensions of RCR, nor the only way to carve up these concepts. Rather, they were designed to maximise the information gain from the Delphi process, by covering a broad range of concepts and reducing redundancy. For example, the dimension "integrity" carries many different connotations and facets, and many of these are already covered by other proposed dimensions gleaned from the research literature and interviews, such as "rigour," "transparency," and others. Therefore, we defined the dimension of "integrity" to cover a more constrained facet not already mentioned, concerning the possession of and adherence to strong moral principles. We also aimed to avoid vague or overly broad definitions, as these would not allow us to know which aspect of a multi-faceted dimension the panellists were responding to. The purpose of this study is not to come up with a consensus definition of any of these concepts; instead, the definitions are intended to make sure the concept space is covered adequately, and that participants are clear about the concepts they are rating. This caveat is included in the instructions to participants.*"**

4.13. The selective recruitment of individuals with experience of RR frameworks feels very sensible and is a core strength of the proposal given the detail provided in determining the lists and inclusion criteria. You may want to consider whether that approach may lead to more homogenous and less diverse sets of ideas based predominantly upon pre-existing models and thus may limit the contributions proffered. It may be that there could be methods used alongside the delphi to complement the process to defend against less minor editing and encourage more substantive or transformative ideas. This is a suggestion that I don't necessarily expect to see actioned, but could be considered in context of the contributions of the proposed work.

This is a valid point, and one that another reviewer commented on too. Our best course of action given the resource limitations we face is to recognise this as a possible risk to the contribution we hope to make. In the context of our broader project, this is less of a concern, but we naturally must consider this issue for the future possibility that others will want to use the framework once we have developed it. We will address this issue in the discussion/limitations section, and have added it as a footnote in the main text also:

"*We recognise that while assembling a panel of RCR experts is appropriate for the aims of this Delphi study, and for the wider aims of the project the study is part of, this approach risks leading to a somewhat homogenous set of dimensions based predominantly upon pre-existing frameworks and models. We recognise that other strategies may lead to a more substantially different or transformative framework in comparison with what already exists.*"

4.14. The practicalities of the delphi could be noted e.g., will it be facilitated through an emailed document and google forms link to provide data etc?. What does "relatively important" refer to? It might be useful to get a sense of how you will make decisions as to including/excluding suggestions at each round – how the dissensuses are maintained could be clearer as it currently sounds like any ideas not widely endorsed would be dropped.

Thanks for bringing this up. We identified two distinct points from this comment. The first relates to the logistics of the Delphi process, and the second on the decision-making in assessing stability and consensus between rounds.

As for the logistics: Delphi participants will complete each round of the Delphi process via an online survey hosted on Qualtrics (https://www.qualtrics.com/uk/). We added the following sentence to clarify this in the manuscript:

"*Each phase or round of this Delphi will be hosted asynchronously in an online environment that panelists can access through a link provided in an email. When panelists follow this link, they will arrive at a Qualtrics survey containing the Delphi questionnaire.*"

As for the second point; we have thoroughly redrafted and augmented the section on the Delphi process to explain the decision-making in much greater detail (see sections titled "Delphi procedure" and "Feedback reports and analysis plans.") You will find that many research decisions to be made in the Delphi process have already been formalised and operationalised, including the decision on "dropping" a dimension from the Delphi. This is now only possible when a dimension has reached stability, in which case it will be considered "set", and reported as such. For specific information on this procedure, see our response to the following point.

4.15. On p16 by "no change" do you mean 'minimal change'? You could provide some scope of what this might be to be more precise.

As mentioned in our response to point 4.14 above, we have now formalised and operationalised when we consider the responses to a particular dimension to have no or minimal changes between

rounds, i.e. when they are "stable". We have copied a relevant paragraph from the section titled "Stability" from the manuscript below:

"*We use a simple metric for this: for each dimension, we will take the absolute value of the change in ratings for each participant. If the mean of these absolute-value change scores is less than the equivalent of 1 point on our rating scale (i.e., 16.66% of the total breadth of the rating scale), we will consider the dimension stable. This tracks closely with the recommended cutoff for stability of 15% difference recommended by Scheibe et al. (2002), which is based on an empirical estimation of the random change between rounds. However, we will also temper this quantitative stability judgement with qualitative analysis: if the qualitative data contain novel arguments for the importance or unimportance of a dimension that we have reason to believe may sway the panel substantially in the following round, we will not consider the dimension to be stable."*

4.16. This protocol is quite brief and could include a more detailed note of what dimensions of the research process will be made openly available. For example, will iterative versions of the reference documents, participants' data and ratings, decision-making log, etc. be made fully available?

Copied from our response to 3.8 above:
Thanks for pointing out that this information was missing. Indeed we plan to share all data and research materials. We have included the following paragraph:

"*We believe in the importance of data sharing, both from the perspective of accountability, as well as the potential re-use of our data. As such, we will share all data and analysis, guided by the TOP Guidelines. We will do this by making the feedback reports openly available in a suitable repository. These feedback reports will include the pseudonymised 'raw' data, both quantitative and qualitative, along with our analyses of this data, subject to any redactions by study participants for their privacy. Final versions of all study materials will be uploaded to this page before the study starts.*"

In sum, the project and manuscript as a whole are well-constructed and provide a clear account of a delphi study that has the potential to form an RCR framework of benefit to our scientific community. There is no doubt that there is much work necessary in this space and that this proposed study has potential value to contribute to a number of developments. However, I hope my feedback encourages the author team to reconsider how to manage the broad terms they discuss (and might negotiate with participants), the justification for a new framework, the need/contributions of a single framework, the potential for the project to do little more than merge pre-existing models (this may itself be a valuable contribution but seems different to the intentions outlined here), and to encourage a little more detail on the procedural and practicality of the project for further transparency. I do hope my thoughts are of value to the research team, I wish them all the very best in conducting work in this important space, and I look forward to reading the next version of this work whether that be as reviewer or reader!

Thank you for your feedback as well as your positive evaluation of our protocol!

Reviewer 5

Thanks for your thorough review. We appreciate the time it must have taken you to identify and describe potential improvements for our study and are very grateful for it, as it certainly made our protocol much better.

5.1. Can the authors please provide a citation to and/or summary of the methods of their "interviews with a range of RCR scholars and practitioners"?

Copied from our response to comment 3.2. above:
Thank you for pointing this out. Indeed, such detail should be made clear for readers. We have added the following paragraph to the manuscript, briefly outlining the data collection and analysis steps that were taken to come to the initial reference document. As pointed out in the text, we have also added an appendix (https://osf.io/jrf47) that describes the analysis process in greater detail.

"*Before drafting the reference document, i.e., a proposed list of RCR dimensions, the authors conducted a scoping review of the existing RCR literature (Field et al., 2024) and interviews with RCR scholars and practitioners (see the interview guide at https://osf.io/xv98y). The interviews consisted of 10 one-on-one interviews and one focus group which included two moderators and eight participants. The articles included in the scoping review and interview transcripts were subject to a thematic analysis, conducted by SMF, which involved coding topically salient sections of text and combining related codes into themes (see the appropriate appendix for a detailed description of this process at https://osf.io/jrf47). This analysis generated a series of overarching themes which reflected salient dimensions of RCR from the literature on the topic. These dimensions are the core of the initial dimension list, to which dimensions taken from existing, older RCR frameworks were added (i.e., the Singapore Statement on Research Integrity, the Australian Code for Responsible Conduct of Research, and the All European Academies European Code of Conduct for Research Integrity).*"

5.2. I find it difficult to appraise and approve the Stage 1 protocol without seeing the initial reference document underlying the document (or, at the very least, a list of the dimensions/items to be rated). The reader does not know what this document looks like, how many items it has that will go into the Delphi, whether these items are potentially double-barreled, etc. Stage 1 review seems premature without seeing the document/items to be rated in the Delphi process.

Copied from response to reviewer comment 1.5.: We apologise that these materials were not part of the original submission where they should have been. This was an oversight on the part of the first author. Please find the reference document on the OSF page for this project at: https://osf.io/jrf47.

5.3. I applaud the thought that has already gone into the eligibility criteria for experts. However, without further detail, I am not sure that I could replicate all of the eligibility criteria for expert panelists aside from publications (at least one manuscript, and a sufficiently senior/leading role on the manuscript). To be an expert, do the frameworks/codes need to be of a sufficient quality or influence, and the role of the person in developing the framework/code to be of a sufficient significance? What kind of role in training/community activities does one need to have, and how many trainings/activities completed? What

does “support” of researchers entail operationally to be eligible? A table of operationalized inclusion version exclusion criteria would help make these recruitment decisions more interpretable and replicable.

Thank you for pointing this out - this could have been much clearer indeed! We have now added a paragraph outlining the inclusion criteria in much greater detail. For convenience, we pasted the paragraph below:

“*For our recruitment strategy, we have operationalised the above concepts into the following inclusion criteria: a participant must have (co)authored at least one peer reviewed article (in articles of more than two co-authors, their position in the author list must indicate leadership in the project in terms of its content; i.e., being in first or second author position, or being corresponding author) including the following keywords: “RRI”, “responsible research and innovation” “research integrity” or “responsible research” AND/OR include one or more of these keywords in their personal institutional webpage AND/OR have taught RCR/RRI to researchers, AND/OR have been involved in a project focusing on RCR/RRI (such as the European Commission’s NewHoRRIzon, MoRRI or SUPER MoRRI projects: https://newhorrizon.eu; https://super-morri.eu/morri-2014-2018/), AND/OR, finally, have been part of a RCR/RRI network or working group (such as the RRING network: https://rring.eu, or the UKRI: https://www.ukri.org).”*

While these operationalisations do not ensure that all participants on the list will be RCR experts, we consider them a valid proxy for the purpose of this study.”

5.4. It is not clear how the authors operationalize “a very diverse and somewhat large participant sample, in terms of disciplinary, geographical, and institutional contexts”. For example, they later write “Once we have reached the target sample and are content that the sample is as diverse as possible (we will be monitoring this aspect as we approach candidate panelists), we will proceed with the Delphi panel.” What criteria will be used to be content that the sample is sufficiently diverse in terms of discipline, geography, and institution: at least one person from each of the disciplines and geographies listed? What about institutions?

We identified two separate points in this comment. First, regarding how we define diversity, and second, about assessing sufficient diversity.

Regarding the first point, we acknowledge that we have not been as clear as we could have been regarding our meaning of diversity. We were referring to diversity in terms of disciplinary (our first concern) and regional diversity (within the limits of the UK and EU primarily). All mentions in the protocol text to other dimensions of diversity have been removed to reflect this. In addition, we have added the following paragraph to explicitly explain this position:

“We also note that despite our goal of developing output that is diverse in terms of the scientific disciplines that are represented in it, this output will represent only a limited selection of countries, regions and cultures. While the broader project within which this study is situated concerns RCR in the UK and regions of Europe, and a Euro-centric approach is appropriate to those ends, we emphasise

that our findings will be produced with the input of a largely Western participant sample. We discuss the impact of this on our findings further in the limitations section in the discussion."

Regarding the second point, in order to avoid disciplinary bias in our sample, where a disproportionate amount of experts have a background in a specific discipline, we decided to include a minimum amount of disciplines present in our sample, as well as a maximum amount of participants from any one discipline. We have added the following paragraph to the manuscript to explain this strategy:

"To avoid disciplinary bias in our sample, i.e., where a disproportionate amount of experts would have a background in a specific discipline, we also decided to include a minimum amount of disciplines present in our sample before we start the Delphi study. As such, the starting panel must represent a minimum of 15 disciplines, with no more than 3 participants from one single discipline."

5.5. It is not clear why the authors chose 40 panelists as their targeted sample size for the first round, given the literature that they cite recommends either 20-30 participants (Melander) or 10-50 participants (Turoff). It is also not clear why the authors chose 20 panelists as the targeted sample size for the final round.

Thanks for pointing out that this information was missing. We have revised the "Panel Size" section to reflect reviewer comments and provide better justifications, noting that there will necessarily be some arbitrariness:

"*Recommendations and empirical studies on dissensus Delphi methods vary on desirable sample size. For instance, 20-30 participants seems to be a sufficient panel based on Melander (2018; note that this is more than typical consensus Delphi panels tend to require), while Turoff (2002) suggests that between 10 and 50 panellists is sufficient for a dissensus Delphi. Choosing a minimum number necessarily contains an arbitrary factor, as well as a pragmatic one. Considering our goal of disciplinary diversity in the expert panel, we elected to use a minimum on the higher end of the average that these two sources suggest. As such, we have set the minimum panel size at the start of the process to be 30 panellists. To avoid disciplinary bias in our sample, i.e., where a disproportionate amount of experts would have a background in a specific discipline, we also decided to include a minimum amount of disciplines present in our sample before we start the Delphi study. As such, the starting panel must represent a minimum of 15 disciplines, with no more than 3 participants from one single discipline."*

5.6. The authors write that they have a list of "approximately 95 individuals". What does "approximately" mean here: can an exact number be provided instead?

Our apologies for this confusion. We used the word "approximately", as we were still refining the list at the time of the first submission. We have removed this mention from the draft, as we included it originally to give evidence of feasibility that there were enough experts to recruit, but we consider now that this information does not necessarily belong in the protocol. Instead, we added more detail about our recruitment criteria (as noted in point 5.3).

5.7. What “scholarly literature” did the authors use to identify potential panelists: the studies included in their scoping review?

We agree that this was not reported in sufficient detail in the first draft. We added a sentence to explain: “The recruitment strategy for this list will be largely centered on the articles included in the scoping review that preceded the Delphi study (Field et al., 2024).”

5.8. Given the number of universities in the world and their often-limited search functionality, it would be helpful for the authors to explain how they found the “online researcher profiles” from which they identified eligible participants.

We agree, this could also be clearer. The section on “Recruitment Strategy” now outlines in more detail how participants will be identified:

“The recruitment strategy for this list will be largely centered on the articles included in the scoping review that preceded the Delphi study (Field et al., 2024). First and second authors of included articles will be searched whether they satisfied the inclusion criteria. Where a specific discipline is missing potential participants, we will carry out google searches with combinations of the following keywords: “RCR”, “RRI”, “responsible research and innovation” “research integrity” and “responsible research” and the names of the specific UoA’s, until we have a list of eligible participants per discipline. Should more recruitment be necessary, we will then actively recruit people through our own networks, including those identified in a previous, more general call for participants for the wider project through professional contacts of the authors, social media, and local university networks.”
(Note that we rewrite the section in future tense, as we recognise we will need to revise and add to our list.)

5.9. The above two recruitment strategies (scholarly literature, online researcher profiles) can help identify experts based on the eligibility criteria related to publications and frameworks/codes. How will the authors find those who are not researchers, but rather only meet the training, community activity, or support inclusion criteria?

This is somewhat addressed in the previous point: we have already reached some potential participants involved only in training or community projects through our previous recruitment calls and our professional networks. However, we target primarily researchers (past or currently active), some of which may be involved in other roles such as training and support. We acknowledge this is an imperfect compromise, and will discuss this in the Limitations section of the discussion.

5.10. Are experts from North America explicitly excluded? Or is their omission a consequence of the methods used to create the sampling frame? Whichever the reason, please clarify and provide a rationale for why this is not an issue.

We apologise if this was not clear from the first draft of our manuscript. In the second draft of this manuscript we have explained and justified our geographic focus in greater detail (see also the answer to 5.4 above). Note that this does not exclude panellists from North America.

5.11. The authors write "We will work to ensure that as much of that diversity as possible filters into the final sample." How? Incentives? Follow-up emails? Obtaining explicit commitment upfront? Offering co-authorship?

We have deleted this sentence from the manuscript, as we will not be able to "ensure" any level of diversity apart from what logically follows from our recruitment strategy and inclusion criteria. The makeup of the panel and its potential implications for our findings will be discussed in the discussion section of our report as per usual.

5.12. The authors write "We will initially approach 68 (i.e., two persons for each of the 34 UoA) and will continue to contact possible candidates until we receive consent to participate from 40 people." Do the authors already know which 68 people from the list of ~95 they plan to approach first? If so, what is the breakdown of how they were identified (i.e., how many are researchers with publications, framework/code authors, trainers/activists, or support staff). Given that the preliminary work has focused heavily on researchers and the published literature, I am wondering whether this panel is already skewing toward researchers with publications at the outset.

We have redrafted our section on "Panel Size" to now reflect that we will simply identify and approach three participants per UoA, rather than prioritise amongst a larger list:

"We will initially approach three persons from each UoA, aiming for well above our minimum sample size. We will start the Delphi process when either 1. at least two persons from each UoA have agreed to take part, or 2. after three weeks of recruitment have elapsed, as long as our minimum panel size and disciplinary diversity requirements are met. If this is not the case we will continue recruiting until the minimum numbers are met."

Unfortunately, we do not believe it will be feasible to explicitly note which participants were identified through publications versus other methods (although we commend the intention of this idea!), as this was not noted in the original sourcing of the list, and at any rate, many participants identified through one route will have applicable inclusion criteria from other routes as well, making this distinction ultimately not very helpful.

Indeed this means that we may be skewing our participation toward researchers with publications. We will discuss this in the Discussion section, as it is an important point to note.

5.13. What concretely do the authors mean by "actively recruit people through our own networks"?

See our answer to 5.8, which outlines the detail we added to our recruitment strategy. The sentence in question now reads: ***Should more recruitment be necessary, we will then actively recruit people through our own networks, including those identified in a previous, more general call for participants***

for the wider project through professional contacts of the authors, social media, and local university networks."

5.14. Related to the request for more operationalized eligibility criteria: how will the authors vet the eligibility of the people offered through snowball sampling?

In the same way as people were deemed eligible in the initial recruitment stage. This is now explicitly explained in the manuscript with the following sentence:
"Potential participants pointed out by declining participants will be vetted as to whether they meet the inclusion criteria before they are invited to participate."

5.15. Do the authors have a copy of the recruitment email that they can share?

A final version of the recruitment e-mail is not yet available, but will be uploaded to the project's OSF page before the recruitment starts, as per our data and material sharing policy. We will likely craft individualised versions of recruitment emails, as this can aid uptake. We will explain that we are running a Delphi study to investigate RCR across disciplines, and that we have contacted them as they have expertise in the field. We will invite them to participate both for the good of the community and the early career researchers involved in this project (as well as any intrinsic interest).

5.16. The authors write "If one month has elapsed and we have not successfully recruited more than 30 individuals, we will go ahead with the Delphi." Why 30 instead of 40 as stated earlier in the protocol? Clarity and coherence in these cutoffs are important to allow Stage 2 reviewers to assess to what degree the panel achieved what it set out to achieve.

We agree that the clarity and coherence in these cutoffs were insufficient in the original manuscript. We have revamped the "Panel Size" and "Recruitment Strategy" sections accordingly, including the below text:
"...we have set the minimum panel size at the start of the process to be 30 panellists. To avoid disciplinary bias in our sample, i.e., where a disproportionate amount of experts would have a background in a specific discipline, we also decided to include a minimum amount of disciplines present in our sample before we start the Delphi study. As such, the starting panel must represent a minimum of 15 disciplines, with no more than 3 participants from one single discipline.

We will initially approach three persons from each UoA, aiming for well above our minimum sample size. We will start the Delphi process when either 1. at least two persons from each UoA have agreed to take part, or 2. after three weeks of recruitment have elapsed, as long as our minimum panel size and disciplinary diversity requirements are met. If this is not the case we will continue recruiting until the minimum numbers are met."

5.17. The authors write "Dissensus is operationalized as people adding new elements to the framework that they consider important yet are 'missing' from their ideal conceptualization of RCR." This definition of dissensus differs from how the term is typically used in the Delphi literature (i.e., variation/dispersion in ratings/rankings). Consensus-oriented Delphi processes commonly allow participants to identify

missing items (e.g., this is standard practice in reporting guideline development). In addition, the authors later write "Subsequent rounds will focus on validating items in an increasingly refined reference list (see the following subsection Developing the Framework for details on this), as participants (hopefully) converge on the most important items, modifying their previous responses based on others' ratings and feedback." This desire for participants to "hopefully converge" is the goal of a consensus-oriented Delphi and antithetical to a dissensus Delphi. As such, I think the use of "dissensus" is inappropriate to describe this proposed Delphi study.

Thank you for pointing this out. This comment shows that our original manuscript did indeed contain inconsistencies. We have rectified all instances pointed to in this comment. In the following paragraph, we outline our main aim with this Delphi study, specifically regarding the role of consensus/dissensus:

"*While most Delphi approaches are consensus-based, meaning that they aim to converge on a selection of important elements of a reference document (Diamond et al., 2014; von der Gracht, 2012), our approach aims to map and refine the existing breadth of perspective on various dimensions of RCR. It is important to note that though this does constitute a departure from the typically practiced Delphi, this actually is a reversion to how the Delphi process was originally intended. Multiple authors have commented that the value of a Delphi study lies exactly in mapping the distributions of opinions instead of generating consensus (Scheibe, 2002; Linstone and Turoff, 2011).*"

5.18. Do the authors have a copy of the demographics survey/questionnaire that they can share?

Our questionnaires are shared in the "Materials" folder in our online OSF project: https://osf.io/7n8bh/. We note in the manuscript that:

"Prior to Phase 1 of the Delphi, we also ask for two participant demographics, namely their geographical region, and years of expertise in their discipline; this is merely to understand the overall makeup of our panel, and will be reported separately to any analysis of the main data."

5.19. Do the authors have a copy of the Phase 1 survey/questionnaire instrument that they can share? I find it difficult to appraise the quality of proposed Phase 1 methods without explicitly seeing the instrument questions (and the Initial Reference Document).

Yes. An exported pdf version of the questionnaires (hosted on Qualtrics) of both phase 1 and 2, as well as the initial reference document, are available on our OSF page: https://osf.io/7n8bh/

5.20. The authors write "We will ask them to answer in relation to their primary field of expertise (that is, the one we recruited them for)." Will the panelists be told the field to which the authors assigned them? This may differ from the field with which panelists self-identify.

Thanks for pointing this out. During the recruitment process we will indicate the discipline we had categorised them as being part of and will also ask them to confirm that they do indeed have

expertise in this area. We have included the following passage to this in the manuscript to clarify this:

"*Each questionnaire will start by asking which discipline a participant identifies with for the purposes of the Delphi, which they will have confirmed with the researchers beforehand during recruitment.*"

5.21. The authors write "The authors will pool the information derived from the first round, construct a feedback report for the participants, and revise the reference document in preparation for Phase 2. The feedback report will include, for instance, the calculated median and IQRs per item, a depiction of the distribution of the responses per item, and a report of what items will be excluded based on low importance ratings. The revised reference document will reflect the participant's suggested dimensions." However, the authors only reported qualitative questions via open-text boxes in Phase 1 (i.e., missing dimensions, additional insights). What closed-item questions are there in Phase 1 that can yield a median/IQR? What is the rating scale? And how many items/dimensions will be rated in Phase 1?

Thanks for pointing out this inconsistency. Indeed, in phase 1, we will not be conducting any measurements that can yield medians or IQRs. In the new manuscript version we make it clear that in this phase we will strictly be asking the panellists to suggest further dimensions for the list. In order to clarify this process further, we added a visualisation of the Delphi process. We have pasted the paragraph that explains this procedure and the corresponding visualisation below:

"*The process will start with an initial modification phase, Phase 1 (the green box in Figure 1), in which participants can suggest additions to the proposed list of dimensions. Here the goal is to broaden the scope of the initial list of dimensions, capturing the various disciplinary perspectives of the panel. After the research team incorporates these suggestions and updates the dimension list, the second phase, Phase 2 (the red box in Figure 1) will involve multiple rounds in which participants rate the importance of these dimensions to RCR in their discipline. In this phase, the goal is to probe which items are more broadly appreciated by the sample (i.e., which might be universally valuable in RCR practice), versus which might be more discipline-specific. In this way, the list of RCR dimensions can be 'weighted' by importance across disciplines.*"

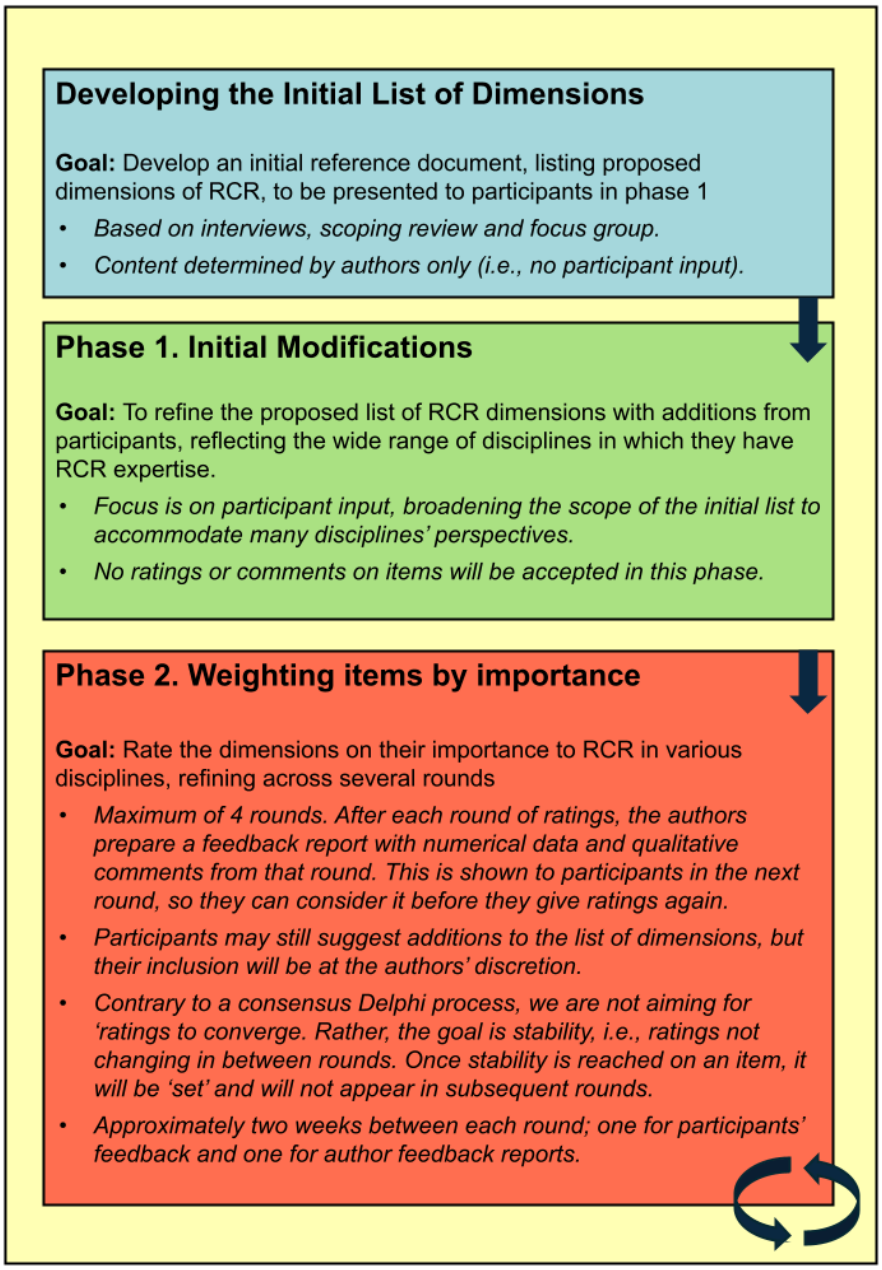

Figure 1. Flowchart showing the phases featuring in this modified Delphi protocol.

5.22. The authors write “In Phase 2, Round 1 the participants are … asked to rate how important each dimension is for their sense of RCR on a 9-point scale (where 9 corresponds to Highly Important and 1 to Unimportant).” This scale is not symmetrical around the middle-point 5. Best practice is to use the same stem for “1” and “9” (e.g., Highly Unimportant to Highly Important). Based on the scale provided, I assume participants will interpret 5 as “neither unimportant or important”, so “important” would fall somewhere between 5 and 9 (probably 7). This is problematic, as the symmetrical opposite of “important” is “unimportant”, but “unimportant” is 1 while “important” is 7.

Thanks for pointing this out. We have adjusted our rating scale to be symmetrical around the middle point, and adjusted it to 7 points, after weighing the diminishing returns of measurement precision in larger scales versus the extra cognitive burden to participants. Please find the corresponding paragraph from the protocol pasted below:

“*The survey will present each proposed dimension consecutively, and will ask the participants how important they consider the dimension of RCR to their specific discipline on a 7-point rating scale ranging from “very unimportant” to “very important”.*”

5.23. The authors write “In the feedback report and the revised reference document, we will also include information about what items were added and which were dropped.” What are the operational criteria for dropping an item?

This comment points out a crucial shortcoming in our original protocol, in that we did not define and operationalise the concept of stability for our Delphi process. We have given much thought on this topic since the last submission of this protocol, and arrived at the following passage:

“***An important aspect of the Delphi process is the concept of stability, or when we consider the answers to be similar enough between two or more subsequent rounds that we can consider the answer “settled” or “definite”. Since panel responses can vary greatly between rounds - as per the explicit aim of the Delphi process - it is important to assess whether the panel’s response on any given dimension is still developing or whether it can be considered settled. In fact, different authors have argued that assessing the level of consensus in a Delphi study is meaningless without having assessed stability of responses, since the response may not be an accurate reflection of the conclusive judgment of the panel (Dajani et al., 1979, Scheibe et al., 2002). To reduce participant burden, where stability is reached, the item will be considered ‘set’, and not feature in subsequent rounds.***

We use a simple metric for stability: for each dimension, we will take the absolute value of the change in ratings for each participant. If the mean of these absolute-value change scores is less than the equivalent of 1 point on our rating scale (i.e., 16.66% of the total breadth of the rating scale), we will consider the dimension stable. This tracks closely with the recommended cutoff for stability of 15% difference recommended by Scheibe et al. (2002), which is based on an empirical estimation of the random change between rounds. However, we will also temper this quantitative stability judgement with qualitative analysis: if the qualitative data contain novel arguments for the importance or unimportance of a dimension that we have reason to believe may sway the panel substantially in the following round, we will not consider the dimension to be stable.”

5.24. The authors write “Subsequent rounds will focus on validating items in an increasingly refined reference list (see the following subsection Developing the Framework for details on this), as participants (hopefully) converge on the most important items, modifying their previous responses based on others’ ratings and feedback.” What do the authors mean by “validating” items? That the panelists will reach consensus or stability in responses?

We have now rewritten the whole “Delphi procedure” section to better communicate our intentions, and no longer use the terminology “validating.” Indeed, in this original passage we meant that we would aim to reach stability, which we now further elaborate in the section “Stability.” Please find the relevant passages explaining this pasted at point 5.23 above.

5.25. The authors write “We will conclude the process after a maximum of 4 Delphi rounds. Melander’s review suggests between 2 and 3 rounds is the average for a consensus Delphi, therefore, since we are including an initial dissensus round, we will conduct a maximum of 4 rounds in Phase 2 (for a possible maximum of 5 rounds including the one round in Phase 1, where participants suggest dimensions).” I had to re-read this section a few times, as it reads first as if there will be 4 rounds overall, then only 4 rounds in Phase 2, and therefore 5 rounds overall. Perhaps this could be solved by changing the first sentence to say a maximum of 5 overall rounds, including the initial round in Phase 1?

Thank you for this suggestion. This certainly could have been clearer. We have changed the relevant section as follows:

“*We will conclude Phase 2 after a maximum of four Delphi rounds. Melander’s review suggests between two and three rounds is the average for a consensus Delphi; however, because we expect a particular diversity of disciplines and perspectives in our Delphi, we will allow up to four rounds if needed.*”

5.26. The authors write “We will conclude the process earlier if no change is observed in the IQR and Median calculations for all items between two given subsequent rounds, or if the author team agrees that little enough change (i.e., so little change as to render the difference conceptually meaningless) has occurred between two rounds.” This concept is referred to as “stability” in the Delphi literature; please provide an operationalized analytical measure for stability (https://www.sciencedirect.com/science/article/pii/S0040162512001023).

As mentioned already in our response to point 5.23., we have explained this process in much greater detail in the current version of the manuscript. Thanks for the suggested literature, which was a great help in our revision.

5.27. I find the authors’ operational definitions for consensus problematic for two inter-related reasons. Firstly, they don’t capture all possibilities (e.g., what happens to items with a Median > 7 but IQR > 2, a Median = 5 but IQR < 2, or a Median = 2 but IQR < 5?). Secondly, by varying the IQR thresholds across the three categorical levels of importance, the definitions conflate agreement/consensus with the panel decision if consensus is reached. One example of resolving this issue: the IQR of 2 could be used as the cut-off for agreement/disagreement. In this case, any item with an IQR of 2 or less would reach

“consensus” in the panel, and then the tertile in which the median falls would determine the decision (7-9 is high importance, 4-6 is moderate importance, 1-3 is low importance). Any item with an IQR > 2 would mean that the panel did not reach consensus.

Thanks for pointing this out, as well as providing a suggestion for the revision. In the current revised protocol you will find that we followed up on your suggestion of separating IQR and median, conceptually (reporting the IQR as an indication of agreement/consensus, and the median as an indication of the panel decision), as they were indeed conflated in the first version of the protocol. Please find the relevant text from the manuscript pasted below:

“Our results will be structured as follows: first, as an overview of the main findings, we will provide a table containing all final dimensions from Phase 2 of the Delphi process, noting which dimensions reached stability, and outlining the final variances and interquartile ranges, as a proxy for consensus, and the median rated importance, as an indication of importance.”

5.28. Related but distinct from the above point: the thresholds for low/moderate/high make sense to me (tertile in which the median falls), though I think a justification/citation for the chosen IQR threshold would be helpful.

Though we have kindly taken up on your suggestion for the thresholds to pre-specify some simple labels for measures of importance, we have decided to leave the analysis of the distributions of the answers (e.g. consensus) completely descriptive. First of all, because we consider an emphasis on the binary consensus-nonconsensus divide unsuitable for our goal of mapping the disciplinary differences in perspectives on RCR. While we would also like to be able to summarise how ‘universal’ a dimension’s rating is across the panel (i.e., a measure of level of agreement/‘consensus’), this is much more difficult to operationalise. We will default to planning only an exploratory and descriptive analysis of universality, for reasons explained below.

We are interested in (and expect we might realistically find) at least three categories of response distributions: either strong agreement around a single point (i.e., ‘consensus’), a relatively flat distribution across all points, or a multi-modal distribution with two or more distinct peaks. We decided that setting an IQR threshold for ‘consensus’ would not suffice, as we expect the data to be non-normally distributed; no matter how low the threshold (i.e., how narrow the interquartile range), there would still be the danger of falsely categorising a distribution with considerable peaks of outliers as consensus (e.g., 52% of responses rated 4, but 24% of responses rated 1 and the other 24% of responses rated 7). This could be solved by using a stricter rule that if the entire range (100%) of responses span a range of 2 points or less on our 1-7 scale, we would consider this ‘consensus’; however, we are concerned that a dichotomous threshold of ‘consensus’ versus ‘non-consensus’ still does not address our ability to categorise the shape of ‘non-consensus’ distributions. We explored some statistical indicators for these three categories of response distributions (e.g. cluster analysis), but found these to be overly complex for our low sample and research goals.

5.29. The authors write “We re-emphasize at this juncture the exploratory nature of this Delphi study. We will need to see the distributions of each item’s data before determining for certain whether these quantitative categories (i.e., the median and IQR threshold ranges defined earlier) are valid and applicable. If they are not, we will redefine our categories and transparently report the change and its motivation.” I understand that the authors are not testing a hypothesis, though I do not believe that this Delphi process is “exploratory” in that the authors will not merely report descriptive statistics of what they find, but rather will use the results instrumentally to create their framework using thresholds of significance. Consequently, I am alarmed by their statement here as it reads as data mining for consensus (https://www.sciencedirect.com/science/article/pii/S0895435617300161). The authors essentially appear to be saying that they will choose the thresholds for consensus based on the pattern of results obtained, or “consensus hacking” (akin to p-hacking but for consensus rather than statistical significance). If the authors want to make claims about consensus decisions, the thresholds for these consensus decisions need to be pre-specified and the authors open to the decisions yielded by the panel (e.g., the panel finds that all items are important, or perhaps no items are important). If the authors are truly conducting an exploratory study, then they should simply report the results from the panel using simple descriptive statistics and avoid defining consensus post hoc based on the nature of the panel results.

Thanks for pointing this out, as well as providing literature with further explanation of these processes as it proved valuable in our revision. We have removed the quoted section from the manuscript, as we agree with the reviewer that it was problematic. We have reformulated the expected outputs of the study to reflect our aims. As noted in 5.28 above, since we no longer are using the Delphi outputs to prescriptively create a framework, and do not expect to be able to pre-specify criteria for labeling the different types of response distributions we expect (as ‘universal,’ ‘bimodal,’ ‘flat,’ etc. except by visual interpretation, we have left this analysis completely exploratory. We will simply present these distributions per dimension, as well as our explicitly descriptive interpretations of them. Please see our completely overhauled “Expected results” section at the end of our revised manuscript for more details.

5.30. The title in the submitted document (“Mapping the Scope of Responsible Research in Practice: A MAD (Modified ReActive Dissensus) Delphi Study”) does not match the title in the submission system (“Capturing Perspectives on Responsible Research Practice: A Delphi Study”).

Thanks for drawing our attention to this. In response to the suggestion in 4.1, we have now changed the title further to “Mapping Cross-Disciplinary Perspectives on Responsible Conduct of Research: A Delphi Study” in the manuscript - hopefully this can be changed on the submission system as well.

5.31. There is an (unintended?) implication that the social sciences are not quantitative: “For example, reproducibility is a concept that applies to quantitative disciplines, but less so qualitative disciplines and the social sciences, and even less so in the humanities.” I am a social scientist that primarily conducts quantitative research involving meta-analyses of randomized trials conducted by other social scientists; reproducibility is quite applicable to this area.

We agree that this sentence did not do justice to the diversity of methodologies within the social sciences. The specific sentence has been deleted.

5.32. The abstract should provide a summary of the methods of the proposed Delphi process.

We have added a very brief description of the process. The relevant part (second half) of the abstract now reads:

"Relying on the expertise and knowledge of a carefully selected multidisciplinary panel of RCR scholars and practitioners, this Delphi study aims to expand the current (underspecified) frameworks of RCR to develop a more diverse and comprehensive concept of what constitutes RCR across disciplines, along with a mapping that captures this updated understanding. The Delphi process will begin with participants refining a provisional list of dimensions of RCR collated from previous literature and interviews, then will proceed with several rounds of rating the importance of each dimension to particular disciplines. [After completion of the study, we will report the details of participant numbers, rounds of Delphi, and a summary of results here.]

5.33. While I agree with the underlying sentiment, I personally found the introductory paragraphs a bit grandiose/hyperbolic. For example, I understand how scientific research is irrelevant when it "becomes misaligned with the needs and expectations of society", but I'm not clear how it is "putting the lives of people and the environment at risk." Fraudulent research on vaccines, sure: but irrelevant research on an unimportant topic?

We recognise that there are different preferences in terms of writing style, and that ours may not be in line with yours. We have chosen to leave the referenced text as is, but duly note that the style may not please everyone. We will keep this in mind when drafting the Discussion section!