The evolution of coordination: a phylogenetic meta-analysis and systematic review

Liam Keeble, Joel C. Wallenberg and Elizabeth E. Price

Review History

RSOS-201728.R0 (Original submission)

Review form: Reviewer 1

Do you have any ethical concerns with this paper?
No

Recommendation?
Reject

Comments to the Author(s)
Although in principle I applaud the authors’ goal of trying to bring together the results of many different cooperation studies with many different species, I am less optimistic this is going to bring forward the field of cooperation in comparative psychology for the reasons I expose below.

In particular, the study proposes the following hypotheses and questions:
Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

The delayed partner string-pull paradigm has produced positive results in a few species (but more species than the ones mentioned in this manuscript). However, this task is not a silver bullet with regards to what we can learn about animals’s skills to cooperate. As it has been argued by other researchers (e.g. Seed & Jensen, 2011) there are several simple explanations that can explain how animals learn to pass these tests. In addition, the number of trials is not the only variable that the different studies have changed. Different studies have also worked with different “delays” and different lengths of the rope, factors which can all influence whether or not and how quickly animals can succeed in this task. Therefore, I am afraid this will be a complicated analysis.

Does inhibition predict cooperative success in a delayed partner string-pull paradigm?

For the reasons explained above, there are actually not many species that have succeeded in the delay task under comparable conditions, so the N will be rather low and one would be comparing learning skills under very different conditions. In addition, the inhibition measure would be a species’ inhibition measure but not a measure obtained from the same populations that participated in the cooperation tasks. This could be acceptable if everything else was highly controlled. However, it is just another source of noise adding to the other ones, and when dealing with a rather low N.

Do social dynamics (affiliation and tolerance) predict cooperative success?

This is possibly the question I am more positive about, just because there are more studies that have looked at tolerance and affiliation. Nevertheless, I would emphasise that cooperation success, as it was measured in many of the studies, does not equate with an understanding of the contingencies and role of the partner in the cooperative endeavour. In other words, one would be looking at something about dyads’ relationships and capacity to interact manipulating together food sources but not *collaborative skills* per se.

Do brain size, social group size, or dietary factors predict cooperative understanding?

As mentioned above, success in the delay task alone does not mean cooperative understanding. The operationalization of cooperative understanding is more complicated. This study proposal mentions gaze-following too, but that also is not necessarily a good dependent measure: we know that subordinates are often nervous in the presence of dominants when manipulating food rewards, which could lead to more “monitoring” behaviour.

Other comments:

- In the pilot data presented, I noted that Chalmeau 1994 and Suchak et al. 2014 are two of the studies analysed but those studies did not use the same string-pulling task mentioned in the introduction as necessary to facilitate any conclusions about cooperation understanding. The apparatus of those studies were different to the one used in all the delay tasks, so although this may be acceptable, one needs to acknowledge that the contingencies for success with that apparatus are different ones. Individuals can pull alone without losing the opportunity to access the rewards, so waiting in that task is not necessarily needed. If the understanding measure used is “monitoring” behaviour, I am sceptical about it for the reasons mentioned above.

- How informative is the correlation - number of trials and cooperation success? Certainly, with more experience cooperative partners will become more coordinated, but (1) are there significant differences in how quickly different species reach success? (2) Are the levels of success comparable across species? i.e. some species quickly being able to wait for up to 25 sec for the partners whereas others the maximum they wait is 5sec? (3) how is success being measured in
those studies using different pulling apparatuses? Because co-acting is not necessarily cooperating.

Review form: Reviewer 2 (Ben Farrar)

Do you have any ethical concerns with this paper?
No

Recommendation?
Major revision

Comments to the Author(s)
Keeble, Wallenberg and Price propose to meta-analyses and conduct a phylogenetic analysis of data from the co-operative string-pulling literature. I think the proposal is exactly the type of approach the field needs to explore more regarding evidence synthesis, and I think it’s great that they’ve chosen to submit this as a registered report. However, as it stands, I think parts the protocol needs much greater specification, and there are several problems the researchers are likely to face when conducting this meta-analysis which might prevent it from providing strong answers to their questions. Nevertheless, I see a large amount of merit in the proposal. I think that if the authors reframed it more in terms of exploring some of the difficulties in performing comparative meta-analyses, as well as focusing on the substantive questions about co-operation, then it could make a very nice RR. I’ve framed my comments around each of the proposed analyses:

1. Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

Could the authors specify more about how they will define the number of trials when extracting data, and similarly how they will define success on the string-pulling paradigm. I can see the definition they provide in the caption to Figure 1, but this could be detailed in text.

With trial data, how will the authors deal with differing numbers of training trials and testing trials overall – i.e., if they define the number of trials as the number of testing trials, what happens if studies have markedly different numbers of training trials – or if this information is not available? Similarly, how will the authors address cases where animals are tested sequentially with different partners, potentially across multiple studies, some of which may not have been published? Overall, I think the exact extraction procedure could be specified in greater detail, and the procedure might need to be expanded such that the authors can document the full extent of heterogeneity between the studies. Understanding this heterogeneity is key to understanding what the output of the meta-analyses means – the heterogeneity might be so large that the aggregate point estimates/CIs don’t have much meaning – and I imagine most people would have high priors that that more trials leads to greater success anyway. I still think there is a lot of merit in collecting and summarizing this data, and the meta-analysis workflow will help this, but I think the main benefit of this proposal is in collecting and presenting information about individual study results and the between-study heterogeneity.

I have further questions about the models the authors are using for the meta-analysis, and how they will deal with repeated data from individuals (e.g., in the pilot report “Peron et al. 2011 Grey parrots” appears with three separate estimates – if I’ve found the paper correctly this comes from 3 parrots tested in in 3 consecutive experiments).
As it stands, I don’t think the current meta-analysis model accounts for phylogenetic relatedness – they might want to consider fitting a multi-level model which incorporates relatedness in (see e.g. Dougherty & Guillette, 2018 as an example), unless there is a reason why they would prefer to not include this information, but this should be explained.

If the present article goes through a revision, it would be nice if the pilot data, extraction protocol and code could be made available to help understand the procedures and models better.

2. Does inhibition predict cooperative success in a delayed partner string-pull paradigm?

This analysis proposes to see if inhibition scores from MacLean et al.’s study predicts success in the string-pulling. I have strong reservations about the ability of this analysis to have the statistical power to produce a meaningful analysis, most of all because the between-site replicability of the MacLean et al. data is largely untested, and where it has been tested it appears low – i.e., what have been billed as species differences in for example the cylinder task performance may not be so (see Figure 3 and the following discussion in Farrar et al., 2020). Because the replicability of inhibition data may be low (as well as for the string pulling data) – this analysis might end up predicting noise with noise. I’m not completely opposed to the analysis being performed, but I think it should be heavily caveated if so. If the authors do decide to proceed, I’d encourage them to search further for inhibition data as there are many studies that use the same tasks as MacLean et al., to ensure they are getting as much data as possible to inform the analysis.

I’m also uncertain about what the best way to interpret the results of this analysis are. On the one hand, correlation does not mean causation, and there are likely many reasons why data from the inhibition task will correlate with data from the string pulling task, without there being a causal relationship between them. On the other hand, I think again it’s incredibly likely that inhibition is causally related to some aspects of passing the string pulling task. If the authors did not find a positive result then I’d be pretty confident that this would be a false negative (i.e., some relationship between inhibition and string-pulling performance is very likely to exist), and if they do find the positive effect, I’m unsure whether the aggregated numerical estimates would mean very much, because of the heterogeneity between studies.

3. Do social dynamics (affiliation and tolerance) predict cooperative success

4. Do brain size, social group size, or dietary factors predict cooperative understanding?

I have similar reservations about these analyses to the previous two. Again, I think it’s highly likely that social dynamics will to some extent predict performance on a social co-operation task (with both causal and non-causal relationships), but it is also possible that the data quality/quantity are not there to perform meaningful quantitative analyses on these. I think the analyses will be useful, but interpreting their outputs may be difficult from a theoretical perspective.

When constructing the phylogenetic trees, how will the authors quantify the uncertainty in the output? They will be able to generate a tree that is the best fit according to some criteria, but if the overall data quality are poor, the tree might be quite likely to be inaccurate, i.e., there is no error control.

My last comments are not specific to a particular analysis:

5. The overall aim of the study
The proposal aims to make “conclusions about the evolutionary origins of cooperative problem solving” by analysing data from the co-operative string-pulling task. However, the generalizability of results from the string-pulling task to other tests (real or hypothetical) of co-operative behaviour is relatively unknown, and as such framing a meta/phylogentic analysis of string-pulling data as an analysis of co-operative behaviour in general is excessive, in my view. The data will be relevant to this question, but not strong. It could instead be framed more narrowly around the string-pulling task.

6. The search protocol and quality control

The search protocol and extraction procedure could be specified to a much a greater extent. One of the strongest outcomes of this study is a high-quality data resource on the string-pulling task. It’s important then to ensure that the search and extraction process are comprehensive and high quality. I’d recommend the authors to follow the PRISMA guidelines for reporting, and to add in several quality control stages to their extraction procedure. Regarding the search terms: searching for comparative data can be very difficult due to mass heterogeneity in how researchers report studies, and I think the Google scholar search and citation-based search will identify most studies. However, it would be good to have some verification stages here to check that key studies are not missing. It would be good if the search structure was specified more, too – for example, are the searches simply performed by inputting the four statements into google scholar as they are, or by using truncation and wildcards (e.g. searching for “co-operat*”), will they search for different combinations and spellings (cooperation as well as co-operation). It’s possible that the authors have already considered this but it would be great if the full structure of the search could be given. The authors may wish to complement their very general search of the entirety Google scholar with more targeted searches of specialist animal behaviour journals as a method of checking that all relevant publications have been identified.

Could the inclusion criteria be specified further still – will the authors include conference abstracts/thesis chapters/non-published articles? I think they definitely should, but they might want to flag these studies when examining publication bias.

In the extraction process, it is unclear what exactly is being extracted and how. It would be great to build a more comprehensive search protocol, and in particular include details on the study design (training trials, sample sizes, sequential testing etc.) and identify studies in which the participants may overlap. I think it is important to have some form of quality control in this process too, either through double extracting a decent percentage of the studies to assess inter-extractor agreement, or more preferable, to have each extraction checked by another researcher.

7. Detecting publication bias

I think its great the authors are paying attention to publication bias in the analysis, and I think the approach could be expanded to examine publication bias in some of the other data they collect too, e.g., by examining the p-value distributions of reported correlations between tolerance and success on the string pulling task. Currently, their interpretation of the funnel plot and significant Egger’s regression test is possibly excessive – a statistically significant asymmetry does not “demonstrate [as in prove] that there are several studies missing”, although it is compatible with that (see e.g. Sterne et al., 2011). There are reasons why comparative datasets from across many species may have asymmetric funnels – for example what if the smaller sample studies - with larger associated errors – are performed on species that are disproportionately likely to pass the test (e.g. a small number of elephants or chimpanzees), whereas the larger sample studies are performed on species that might be less likely to pass the test (e.g. dogs or monkeys). I don’t necessarily agree with this reasoning, and think the funnel test is good evidence of publication bias, but currently I believe its overinterpreted. I think it’s incredibly difficult to assess publication bias in comparative datasets because of the massive heterogeneity in the laboratories
and species publishing and performing the tasks. I’d encourage the authors to examine subgroups (e.g. by colour coding their funnel plots by species/groups), particularly if they get a decent amount of data on any one particular group.

Finally, the authors may wish to consider alternative methods of detecting publication bias – for example by identifying thesis chapters or conference abstracts that have not been published, or even by simply asking researchers in the field if they have performed any string-pulling tasks that have gone unpublished. Some information like this would complement the statistical analysis massively, in my opinion.

Overall, I think this will be a really valuable study. I think the first aim should be in creating a well-document data resource around the string pulling task, from which they will be able to assess how much information the meta-analyses and phylogenetic analysis can provide. There are several key problems to navigate for the analyses (heterogeneity, structure of the models, data quality) and it’s tricky to review this without having much of the information about the studies yet. This will only be available after extractions have been completed, and so I think it is really important if IPA is to be given to know the extraction process will be very high quality, and that a critical assessment of the strengths of the meta-analyses will be performed after this, or to have another review/discussion about the proposed analyses after extractions have been completed. If it turns out that the data are too low in quality (or in quantity, or homogeneity) to perform strong analyses, then this information is still very valuable to the field, and the process by which the authors will decide this will be a useful methodological advance in the field.

Sincerely,
Ben Farrar

Disclaimer: all my reviews are signed and I am happy to provide clarifications on any points that might be ambiguous to the editors/reviewers/authors (bgf22@cam.ac.uk).

Dougherty, L. R., & Guillette, L. M. (2018). Linking personality and cognition: A meta-analysis. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1756), 20170282. https://doi.org/10.1098/rstb.2017.0282

Farrar, B., Voudouris, K., & Clayton, N. (2020). Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Behavior and Cognition Research. PsyArXiv. https://doi.org/10.31234/osf.io/2vt4k

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ, 343. https://doi.org/10.1136/bmj.d4002

Decision letter (RSOS-201728.R0)

We hope you are keeping well at this difficult and unusual time. We continue to value your support of the journal in these challenging circumstances. If Royal Society Open Science can assist you at all, please don’t hesitate to let us know at the email address below.

Dear Mr Keeble,

The Editors assigned to your Stage 1 Registered Report ("The evolution of cooperative problem-solving: A phylogenetic meta-analysis") have now received comments from reviewers. We
would like you to revise your paper in accordance with the referee and editors suggestions which can be found below (not including confidential reports to the Editor). Please note this decision does not guarantee eventual acceptance.

Please submit a copy of your revised paper within six weeks (i.e. by 1st December 2020). If we do not hear from you within this time then it will be assumed that the paper has been withdrawn. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office in advance.

To revise your manuscript, log into http://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. Revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you must respond to the comments made by the referees and upload a file "Response to Referees" in "Section 2 - File Upload". Please use this to document how you have responded to the comments, and the adjustments you have made. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response.

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Kind regards,
Professor Chris Chambers
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Chris Chambers (Registered Reports Editor, Royal Society Open Science)
openscience@royalsociety.org

Associate Editor Comments to Author (Professor Chris Chambers):
Comments to the Author:
Two specialist reviewers have now assessed the manuscript, with both providing very detailed and constructive assessments of the proposal. Both reviewers are also deeply critical, noting a wide range of areas where insufficient methodological detail is provided (including code and data for the pilot analyses) and unaddressed issues that are central in meta-analysis such as methodological heterogeneity. As you will see, the reviewers (and Rev 1 especially) are also skeptical as to whether the approach, as described, is capable of answering most of the research questions.

I am convinced overall by the reviews that this submission has sufficient merit to invite a Major Revision, but I do want to stress that substantial work will be required to achieve Stage 1 IPA, and a revised submission will be returned to both of the reviewers for re-assessment.

Comments to Author:
Reviewer: 1
Comments to the Author(s)
Although in principle I applaud the authors’ goal of trying to bring together the results of many different cooperation studies with many different species, I am less optimistic this is going to bring forward the field of cooperation in comparative psychology for the reasons I expose below.
In particular, the study proposes the following hypotheses and questions:

(1) Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

The delayed partner string-pull paradigm has produced positive results in a few species (but more species than the ones mentioned in this manuscript). However, this task is not a silver bullet with regards to what we can learn about animals’ skills to cooperate. As it has been argued by other researchers (e.g. Seed & Jensen, 2011) there are several simple explanations that can explain how animals learn to pass these tests.

In addition, the number of trials is not the only variable that the different studies have changed. Different studies have also worked with different “delays” and different lengths of the rope, factors which can all influence whether or not and how quickly animals can succeed in this task. Therefore, I am afraid this will be a complicated analysis.

(2) Does inhibition predict cooperative success in a delayed partner string-pull paradigm? For the reasons explained above, there are actually not many species that have succeeded in the delay task under comparable conditions, so the N will be rather low and one would be comparing learning skills under very different conditions. In addition, the inhibition measure would be a species’ inhibition measure but not a measure obtained from the same populations that participated in the cooperation tasks. This could be acceptable if everything else was highly controlled. However, it is just another source of noise adding to the other ones, and when dealing with a rather low N.

(3) Do social dynamics (affiliation and tolerance) predict cooperative success?

This is possibly the question I am more positive about, just because there are more studies that have looked at tolerance and affiliation. Nevertheless, I would emphasise that cooperation success, as it was measured in many of the studies, does not equate with an understanding of the contingencies and role of the partner in the cooperative endeavour. In other words, one would be looking at something about dyads’ relationships and capacity to interact manipulating together food sources but not “collaborative skills” per se.

(4) Do brain size, social group size, or dietary factors predict cooperative understanding?

As mentioned above, success in the delay task alone does not mean cooperative understanding. The operationalization of cooperative understanding is more complicated. This study proposal mentions gaze-following too, but that also is not necessarily a good dependent measure: we know that subordinates are often nervous in the presence of dominants when manipulating food rewards, which could lead to more “monitoring” behaviour.

Other comments:

• In the pilot data presented, I noted that Chalmeau 1994 and Suchak et al. 2014 are two of the studies analysed but those studies did not use the same string-pulling task mentioned in the introduction as necessary to facilitate any conclusions about cooperation understanding. The apparatus of those studies were different to the one used in all the delay tasks, so although this may be acceptable, one needs to acknowledge that the contingencies for success with that apparatus are different ones. Individuals can pull alone without losing the opportunity to access the rewards, so waiting in that task is not necessarily needed. If the understanding measure used is “monitoring” behaviour, I am sceptical about it for the reasons mentioned above.

• How informative is the correlation - number of trials and cooperation success? Certainly, with more experience cooperative partners will become more coordinated, but (1) are there significant differences in how quickly different species reach success? (2) Are the levels of success comparable across species? i.e. some species quickly being able to wait for up to 25 sec for the
partners whereas others the maximum they wait is 5sec? (3) how is success being measured in those studies using different pulling apparatuses? Because co-acting is not necessarily cooperating.

Reviewer: 2
Comments to the Author(s)
Keeble, Wallenberg and Price propose to meta-analyses and conduct a phylogenetic analysis of data from the co-operative string-pulling literature. I think the proposal is exactly the type of approach the field needs to explore more regarding evidence synthesis, and I think it’s great that they’ve chosen to submit this as a registered report. However, as it stands, I think parts the protocol needs much greater specification, and there are several problems the researchers are likely to face when conducting this meta-analysis which might prevent it from providing strong answers to their questions. Nevertheless, I see a large amount of merit in the proposal. I think that if the authors reframed it more in terms of exploring some of the difficulties in performing comparative meta-analyses, as well as focusing on the substantive questions about co-operation, then it could make a very nice RR. I’ve framed my comments around each of the proposed analyses:

1. Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

Could the authors specify more about how they will define the number of trials when extracting data, and similarly how they will define success on the string-pulling paradigm. I can see the definition they provide in the caption to Figure 1, but this could be detailed in text. With trial data, how will the authors deal with differing numbers of training trials and testing trials overall – i.e., if they define the number of trials as the number of testing trials, what happens if studies have markedly different numbers of training trials – or if this information is not available? Similarly, how will the authors address cases where animals are tested sequentially with different partners, potentially across multiple studies, some of which may not have been published? Overall, I think the exact extraction procedure could be specified in greater detail, and the procedure might need to be expanded such that the authors can document the full extent of heterogeneity between the studies. Understanding this heterogeneity is key to understanding what the output of the meta-analyses means – the heterogeneity might be so large that the aggregate point estimates/CIs don’t have much meaning – and I imagine most people would have high priors that more trials leads to greater success anyway. I still think there is a lot of merit in collecting and summarizing this data, and the meta-analysis workflow will help this, but I think the main benefit of this proposal is in collecting and presenting information about individual study results and the between-study heterogeneity.

I have further questions about the models the authors are using for the meta-analysis, and how they will deal with repeated data from individuals (e.g., in the pilot report “Peron et al. 2011 Grey parrots” appears with three separate estimates – if I’ve found the paper correctly this comes from 3 parrots tested in in 3 consecutive experiments). As it stands, I don’t think the current meta-analysis model accounts for phylogenetic relatedness – they might want to consider fitting a multi-level model which incorporates relatedness in (see e.g., Dougherty & Guillette, 2018 as an example), unless there is a reason why they would prefer to not include this information, but this should be explained.

If the present article goes through a revision, it would be nice if the pilot data, extraction protocol and code could be made available to help understand the procedures and models better.

2. Does inhibition predict cooperative success in a delayed partner string-pull paradigm?
This analysis proposes to see if inhibition scores from MacLean et al.’s study predicts success in the string-pulling. I have strong reservations about the ability of this analysis to have the statistical power to produce a meaningful analysis, most of all because the between-site replicability of the MacLean et al. data is largely untested, and where it has been tested it appears low – i.e., what have been billed as species differences in for example the cylinder task performance may not be so, (see Figure 3 and the following discussion in Farrar et al., 2020). Because the replicability of inhibition data may be low (as well as for the string pulling data) – this analysis might end up predicting noise with noise. I’m not completely opposed to the analysis being performed, but I think it should be heavily caveated if so. If the authors do decide to proceed, I’d encourage them to search further for inhibition data as there are many studies that use the same tasks as MacLean et al., to ensure they are getting as much data as possible to inform the analysis.

I’m also uncertain about what the best way to interpret the results of this analysis are. On the one hand, correlation does not mean causation, and there are likely many reasons why data from the inhibition task will correlate with data from the string pulling task, without there being a causal relationship between them. On the other hand, I think again it’s incredibly likely that inhibition is causally related to some aspects of passing the string pulling task. If the authors did not find a positive result then I’d be pretty confident that this would be a false negative (i.e., some relationship between inhibition and string-pulling performance is very likely to exist), and if they do find the positive effect, I’m unsure whether the aggregated numerical estimates would mean very much, because of the heterogeneity between studies.

3. Do social dynamics (affiliation and tolerance) predict cooperative success

4. Do brain size, social group size, or dietary factors predict cooperative understanding?

I have similar reservations about these analyses to the previous two. Again, I think it’s highly likely that social dynamics will to some extent predict performance on a social co-operation task (with both causal and non-causal relationships), but it is also possible that the data quality/quantity are not there to perform meaningful quantitative analyses on these. I think the analyses will be useful, but interpreting their outputs may be difficult from a theoretical perspective.

When constructing the phylogenetic trees, how will the authors quantify the uncertainty in the output? They will be able to generate a tree that is the best fit according to some criteria, but if the overall data quality are poor, the tree might be quite likely to be inaccurate, i.e., there is no error control.

My last comments are not specific to a particular analysis:

5. The overall aim of the study

The proposal aims to make “conclusions about the evolutionary origins of cooperative problem solving” by analysing data from the co-operative string-pulling task. However, the generalizability of results from the string-pulling task to other tests (real or hypothetical) of co-operative behaviour is relatively unknown, and as such framing a meta/phylogentic analysis of string-pulling data as an analysis of co-operative behaviour in general is excessive, in my view. The data will be relevant to this question, but not strong. It could instead be framed more narrowly around the string-pulling task.

6. The search protocol and quality control
The search protocol and extraction procedure could be specified to a much greater extent. One of the strongest outcomes of this study is a high-quality data resource on the string-pulling task. It’s important then to ensure that the search and extraction process are comprehensive and high quality. I’d recommend the authors to follow the PRISMA guidelines for reporting, and to add in several quality control stages to their extraction procedure. Regarding the search terms: searching for comparative data can be very difficult due to mass heterogeneity in how researchers report studies, and I think the Google scholar search and citation-based search will identify most studies. However, it would be good to have some verification stages here to check that key studies are not missing. It would be good if the search structure was specified more, too – for example, are the searches simply performed by inputting the four statements into google scholar as they are, or by using truncation and wildcards (e.g. searching for “co-operat*”), will they search for different combinations and spellings (cooperation as well as co-operation). It’s possible that the authors have already considered this but it would be great if the full structure of the search could be given. The authors may wish to complement their very general search of the entirety Google scholar with more targeted searches of specialist animal behaviour journals as a method of checking that all relevant publications have been identified.

Could the inclusion criteria be specified further still – will the authors include conference abstracts/thesis chapters/non-published articles? I think they definitely should, but they might want to flag these studies when examining publication bias.

In the extraction process, it is unclear what exactly is being extracted and how. It would be great to build a more comprehensive search protocol, and in particular include details on the study design (training trials, sample sizes, sequential testing etc.) and identify studies in which the participants may overlap. I think it is important to have some form of quality control in this process too, either through double extracting a decent percentage of the studies to assess inter-extractor agreement, or more preferable, to have each extraction checked by another researcher.

7. Detecting publication bias

I think its great the authors are paying attention to publication bias in the analysis, and I think the approach could be expanded to examine publication bias in some of the other data they collect too, e.g., by examining the p-value distributions of reported correlations between tolerance and success on the string pulling task. Currently, their interpretation of the funnel plot and significant Egger’s regression test is possibly excessive – a statistically significant asymmetry does not “demonstrate [as in prove] that there are several studies missing”, although it is compatible with that (see e.g. Sterne et al., 2011). There are reasons why comparative datasets from across many species may have asymmetric funnels – for example what if the smaller sample studies - with larger associated errors – are performed on species that are disproportionately likely to pass the test (e.g. a small number of elephants or chimpanzees), whereas the larger sample studies are performed on species that might be less likely to pass the test (e.g. dogs or monkeys). I don’t necessarily agree with this reasoning, and think the funnel test is good evidence of publication bias, but currently I believe its overinterpreted. I think it’s incredibly difficult to assess publication bias in comparative datasets because of the massive heterogeneity in the laboratories and species publishing and performing the tasks. I’d encourage the authors to examine subgroups (e.g. by colour coding their funnel plots by species/groups), particularly if they get a decent amount of data on any one particular group.

Finally, the authors may wish to consider alternative methods of detecting publication bias – for example by identifying thesis chapters or conference abstracts that have not been published, or even by simply asking researchers in the field if they have performed any string-pulling tasks that have gone unpublished. Some information like this would complement the statistical analysis massively, in my opinion.
Overall, I think this will be a really valuable study. I think the first aim should be in creating a well-document data resource around the string pulling task, from which they will be able to assess how much information the meta-analyses and phylogenetic analysis can provide. There are several key problems to navigate for the analyses (heterogeneity, structure of the models, data quality) and it’s tricky to review this without having much of the information about the studies yet. This will only be available after extractions have been completed, and so I think it is really important if IPA is to be given to know the extraction process will be very high quality, and that a critical assessment of the strengths of the meta-analyses will be performed after this, or to have another review/discussion about the proposed analyses after extractions have been completed. If it turns out that the data are too low in quality (or in quantity, or homogeneity) to perform strong analyses, then this information is still very valuable to the field, and the process by which the authors will decide this will be a useful methodological advance in the field.

Sincerely,
Ben Farrar

Disclaimer: all my reviews are signed and I am happy to provide clarifications on any points that might be ambiguous to the editors/reviewers/authors (bgf22@cam.ac.uk).

Dougherty, L. R., & Guillette, L. M. (2018). Linking personality and cognition: A meta-analysis. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1756), 20170282. https://doi.org/10.1098/rstb.2017.0282

Farrar, B., Voudouris, K., & Clayton, N. (2020). Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Behavior and Cognition Research. PsyArXiv. https://doi.org/10.31234/osf.io/2vt4k

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ, 343. https://doi.org/10.1136/bmj.d4002

Author's Response to Decision Letter for (RSOS-201728.R0)

See Appendix A.

RSOS-201728.R1 (Revision)

Review form: Reviewer 1

Do you have any ethical concerns with this paper?
No

Recommendation?
Reject
Comments to the Author(s)

Although this revised version addresses some of the issues I raised in the previous round, I still remain unconvinced that a phylogenetic meta-analysis in this field of research can currently bring the field forward. The reasons are, as I commented before, the following:

- There are very few species which have passed the delay task (i.e. chimpanzees, elephants, maybe dogs and dolphins), so with such scarce evidence, I am not sure what is the point of a phylogenetic analysis. In addition, it is now suggested to constrain more which studies enter the analysis to make the comparison more valid (due to the many different testing and experimental protocols), but how many studies would we be left with? if currently only 2-3 species have been considered to pass the task?

- In this revised version Cooperation has been substituted with "coordination", but evidence for coordination can only come from studies showing active efforts to synchronise actions in time and space. Therefore, many studies using other cooperation apparatuses are less indicative of coordination (Suchak et al. Mendres and De Waal, 2003.) because subjects can pull indiscriminately until the apparatus moves (co-occurring when another subject has joined). However, if I understand correctly, these different studies would also enter the analysis. I don't think by changing the word to "coordination", authors can circumvent the difficulty of making conclusions about the mechanisms that bring success about in these co-acting tasks.

- The question about the quality of the relationship between subjects and success in a cooperation task is interesting, but it has already been shown experimentally in several species (e.g. Capuchin monkeys, De Waal & Davis, 2003; Tonquean macaques, Petit et al. 1992; Chimpanzees, Melis et al. 2006; Rooks, Seed et al. 2008). These previous studies have shown the effect of the relationship between partners on cooperative problem solving behaviour *within* a species and *across* species.

One last point is that there is a probably not a *single* method or experimental protocol that can best investigate animals' capacity to solve problems cooperatively with others (and coordinate actions). The type of tasks that can be used with primates are maybe not equally suited for birds or dolphins and the other way around. Therefore, the delay task with the string-pulling paradigm should not be seen as the "silver bullet" to understanding Cooperation/Coordination, but instead as one task that together with others may give us some insight into animals' capacity to coordinate actions with each other.

What we urgently need are multiple paradigms and measures for each species and/or the exact same experiment across different species. Instead, what we currently have is one unique method implemented in myriad ways.

Review form: Reviewer 2 (Ben Farrar)

Do you have any ethical concerns with this paper?
No

Recommendation?
Major revision

Comments to the Author(s)
The revised proposal has addressed many of the previous comments, increased the rigour of the proposal and added in several additional analyses. This will be a complicated project, and there is
still work to be done to ensure it would be effective should IPA be given, but I still think it is a useful project.

1. Structure

The proposal would be easier to follow if the introduction was split in two, with a first section outlining the original research goals (meta-analysing string pulling data), and then a section outlining the challenges of doing so, rather than having the challenges spread throughout the introduction. This may be a personal preference, but I think it would improve the article’s usability as a resource for meta-analysis in the area, too.

2. Quality control and assessment

The authors have increased the amount of quality control and assessment stages in the revision.

a) Searches
Although the searches are not fully comprehensive (e.g., animal cognition journals such as The Journal of Comparative Psychology, International Journal of Comparative Psychology, Behavioural Processes, Learning and Behavior are not included in the journal search), there is enough redundancy across the three searches to ensure that the vast majority of relevant articles will be identified.

b) Extraction
The extraction process, where the articles are extracted twice by the same individual and then another coding 10% seems fine, although if there are inconsistencies across this 10% more papers will need double coding.

c) Quality Checklist
The authors have introduced a quality checklist – a modified NICE quality appraisal checklist. I applaud the use of such a checklist, and think this is an interesting project in itself, so should be retained. However, using the checklist as a basis for inclusion/exclusion in the meta-analysis might be a problem, as I’m unsure how the authors will be able to judge some of the questions, and it is likely that most studies would end up being excluded if the checklist was followed strictly.

For example, for question 1b:

Do the selected participants or areas represent the eligible population or area? (e.g. were both sexes represented fairly equally? Were a variety of ages represented fairly equally?)

How will the authors know what the eligible populations are without contacting every study author to see which animals were available? Often in captivity the ages of the animals are similar, so many studies will not represent a variety of ages fairly.

For question 2b:

How well were likely confounding factors identified and controlled?

This would require the authors to pretty much peer-review each paper, unless the authors wanted to decide on a few key confounds for the string pulling task and check these in each paper.
Performing the checklist will be an interesting exercise, possibly to quantify just how far these studies are from the strong epidemiological studies that the quality appraisal form was designed to identify. However, I doubt it can be used effectively as a basis of inclusion/exclusion unless it heavily modified.

d) The reliability of the phylogenetic tree

AIC will be used to compare competing trees, but this will only compare the fit of the trees relative to each other, and not absolutely, i.e., the tree with the lowest AIC will not necessarily be a good tree, it will just be the best fitting of all the ones compared on the current dataset. I think this needs to be made clearer in the manuscript, particularly given the noise and limited amount of data that will be informing the trees.

4) Meta-analysis/statistical analyses

The meta-analytic model has been updated to be multi-level, which is good as it can attempt to account for phylogenetic relatedness. However, I don’t think the current approach, which includes species and genus only (random = ~1 | Paper/Species/Genus), will capture most of this phylogenetic information – it would only be useful where many species come from the same genus.

A couple of examples of how to include phylogeny with the rma.mv function can be found in Cinar et al., 2020 (Model 9) and Davies et al., 2020. Both of these provide good descriptions of how they generate the phylogenetic tree (Davies et al. in particular for when distantly related taxa are used), and provide example code in the supporting information. The approach is to include a single random effect of phylogeny, based on a correlation matrix derived from a phylogenetic tree. Both Cinar et al. and Davies et al. use the same APE package to generate these as the current authors are using, so their code should be adaptable.

The authors have a large number of analyses planned, and they could consider how they will control for multiple testing/whether they think it is an issue.

5) Data and code

Thank-you for sharing the data and code. I managed to reproduce the pilot analysis fully and the data file was readable. When the full dataset is generated, it would be useful to have a data-dictionary to accompany the file explain precisely what each column is and how it was coded – as I expect this dataset might be of interest to many people.

Overall, the new proposal is improved, and I still think it is a feasible and worthwhile project. I share Reviewer 1’s concerns about the ability of the project to provide strong answers to the original questions of interest. But even if it fails to answer these questions effectively, then highlighting the barriers around synthesizing evidence like would be a very useful paper. The feasibility of these analyses will depend on the quantity and quality of the overall dataset, which we will only know when it has been extracted. Therefore, it would be useful to conduct a feasibility assessment of the different analyses after the data have been extracted but before conducting them. If IPA is given to the project, I expect a longer Stage 2 review may be necessary than with other registered reports. I still think the project is very interesting and will be of interest those interested in string-pulling tasks but also those interested in evidence synthesis in the field.

Ben

Minor point
1. Code reproducibility

Code, line 355 there is a comma after the first value in

\>

ref = c("HM015213",)

This should be removed to read

ref = c("HM015213")

I was unable to run the code of the phylogenetic analysis – but I assume this just hasn’t been completed yet.

References:

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2020). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. EcoEvoRxiv. https://doi.org/10.32942/osf.io/su4zv

Davies, A. D., Lewis, Z., & Dougherty, L. R. (2020). A meta-analysis of factors influencing the strength of mate-choice copying in animals. Behavioral Ecology, 31(6), 1279–1290. https://doi.org/10.1093/beheco/araa064

Decision letter (RSOS-201728.R1)

The editorial office reopened on 4 January 2021. We are working hard to catch up after the festive break. If you need advice or an extension to a deadline, please do not hesitate to let us know – we will continue to be as flexible as possible to accommodate the changing COVID situation. We wish you a happy New Year, and hope 2021 proves to be a better year for everyone.

Dear Mr Keeble,

The Editors assigned to your Stage 1 Registered Report ("The evolution of coordination: A phylogenetic meta-analysis and systematic review") have now received comments from reviewers. We would like you to revise your paper in accordance with the referee and editors suggestions which can be found below (not including confidential reports to the Editor). Please note this decision does not guarantee eventual acceptance.

Please submit a copy of your revised paper within three weeks (i.e. by the 27-Jan-2021). If we do not hear from you within this time then it will be assumed that the paper has been withdrawn. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office in advance. We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Editors, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers.

To revise your manuscript, log into http://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. Revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you must respond to the comments made by the referees and upload a file "Response to Referees". Please use this to document how you have
responded to the comments, and the adjustments you have made. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response.

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Kind regards,
Royal Society Open Science Editorial Office
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Chris Chambers (Registered Reports Editor, Royal Society Open Science)
openscience@royalsociety.org

Associate Editor Comments to Author (Professor Chris Chambers):
Associate Editor: 1
Comments to the Author:
The revised manuscript was returned to the two original reviewers. Both reviewers offer a range of comments on the updated proposal. Reviewer 2 is broadly positive but also notes several methodological issues requiring further clarification. As in the first round, Reviewer 1 is again the more critical of the reviewers, questioning the overarching rationale for the research and the heterogeneity of the studies included in the meta-analysis. These are major issues that will need to be thoroughly addressed to achieve IPA.

Given the progress made in improving the design, and the enthusiasm of Reviewer 2, I want to give the authors a final opportunity to address these concerns before issuing a final Stage 1 editorial decision.

Comments to Author:
Reviewer: 1
Comments to the Author(s)
Although this revised version addresses some of the issues I raised in the previous round, I still remain unconvinced that a phylogenetic meta-analysis in this field of research can currently bring the field forward. The reasons are, as I commented before, the following:

- There are very few species which have passed the delay task (i.e. chimpanzees, elephants, maybe dogs and dolphins), so with such scarce evidence, I am not sure what is the point of a phylogenetic analysis. In addition, it is now suggested to constrain more which studies enter the analysis to make the comparison more valid (due to the many different testing and experimental protocols), but how many studies would we be left with? if currently only 2-3 species have been considered to pass the task?

- In this revised version Cooperation has been substituted with "coordination", but evidence for coordination can only come from studies showing active efforts to synchronise actions in time and space. Therefore, many studies using other cooperation apparatuses are less indicative of coordination (Suchak et al. Mendres and De Waal, 2003..) because subjects can pull indiscriminately until the apparatus moves (co-occurring when another subject has joined). However, if I understand correctly, these different studies would also enter the analysis. I don't think by changing the word to "coordination", authors can circumvent the difficulty of making conclusions about the mechanisms that bring success about in these co-acting tasks.
- The question about the quality of the relationship between subjects and success in a cooperation task is interesting, but it has already been shown experimentally in several species (e.g. Capuchin monkeys, De Waal & Davis, 2003; Tonquean macaques, Petit et al. 1992; Chimpanzees, Melis et al. 2006; Rooks, Seed et al. 2008). These previous studies have shown the effect of the relationship between partners on cooperative problem solving behaviour *within* a species and *across* species.

One last point is that there is a probably not a *single* method or experimental protocol that can best investigate animals' capacity to solve problems cooperatively with others (and coordinate actions). The type of tasks that can be used with primates are maybe not equally suited for birds or dolphins and the other way around. Therefore, the delay task with the string-pulling paradigm should not be seen as the "silver bullet" to understanding Cooperation/Coordination, but instead as one task that together with others may give us some insight into animals' capacity to coordinate actions with each other.

What we urgently need are multiple paradigms and measures for each species and/or the exact same experiment across different species. Instead, what we currently have is one unique method implemented in myriad ways.

Reviewer: 2
Comments to the Author(s)
The revised proposal has addressed many of the previous comments, increased the rigour of the proposal and added in several additional analyses. This will be a complicated project, and there is still work to be done to ensure it would be effective should IPA be given, but I still think it is a useful project.

1. Structure

The proposal would be easier to follow if the introduction was split in two, with a first section outlining the original research goals (meta-analysing string pulling data), and then a section outlining the challenges of doing so, rather than having the challenges spread throughout the introduction. This may be a personal preference, but I think it would improve the article’s usability as a resource for meta-analysis in the area, too.

2. Quality control and assessment

The authors have increased the amount of quality control and assessment stages in the revision.

a) Searches
Although the searches are not fully comprehensive (e.g., animal cognition journals such as The Journal of Comparative Psychology, International Journal of Comparative Psychology, Behavioural Processes, Learning and Behavior are not included in the journal search), there is enough redundancy across the three searches to ensure that the vast majority of relevant articles will be identified.

b) Extraction
The extraction process, where the articles are extracted twice by the same individual and then another coding 10% seems fine, although if there are inconsistencies across this 10% more papers will need double coding.

c) Quality Checklist
The authors have introduced a quality checklist – a modified NICE quality appraisal checklist. I applaud the use of such a checklist, and think this is an interesting project in itself, so should be retained. However, using the checklist as a basis for inclusion/exclusion in the meta-analysis might be a problem, as I’m unsure how the authors will be able to judge some of the questions, and it is likely that most studies would end up being excluded if the checklist was followed strictly.

For example, for question 1b:

Do the selected participants or areas represent the eligible population or area? (e.g. were both sexes represented fairly equally? Were a variety of ages represented fairly equally?)

How will the authors know what the eligible populations are without contacting every study author to see which animals were available? Often in captivity the ages of the animals are similar, so many studies will not represent a variety of ages fairly.

For question 2b:

How well were likely confounding factors identified and controlled?

This would require the authors to pretty much peer-review each paper, unless the authors wanted to decide on a few key confounds for the string pulling task and check these in each paper.

Performing the checklist will be an interesting exercise, possibly to quantify just how far these studies are from the strong epidemiological studies that the quality appraisal form was designed to identify. However, I doubt it can be used effectively as a basis of inclusion/exclusion unless it heavily modified.

d) The reliability of the phylogenetic tree

AIC will be used to compare competing trees, but this will only compare the fit of the trees relative to each other, and not absolutely, i.e., the tree with the lowest AIC will not necessarily be a good tree, it will just be the best fitting of all the ones compared on the current dataset. I think this needs to be made clearer in the manuscript, particularly given the noise and limited amount of data that will be informing the trees.

4) Meta-analysis/statistical analyses

The meta-analytic model has been updated to be multi-level, which is good as it can attempt to account for phylogenetic relatedness. However, I don’t think the current approach, which includes species and genus only (random = ~1|Paper/Species/Genus), will capture most of this phylogenetic information – it would only be useful where many species come from the same genus.

A couple of examples of how to include phylogeny with the rma.mv function can be found in Cinar et al., 2020 (Model 9) and Davies et al., 2020. Both of these provide good descriptions of how they generate the phylogenetic tree (Davies et al. in particular for when distantly related taxa are used), and provide example code in the supporting information. The approach is to include a single random effect of phylogeny, based on a correlation matrix derived from a phylogenetic tree. Both Cinar et al. and Davies et al. use the same APE package to generate these as the current authors are using, so their code should be adaptable.
The authors have a large number of analyses planned, and they could consider how they will control for multiple testing/whether they think it is an issue.

5) Data and code

Thank you for sharing the data and code. I managed to reproduce the pilot analysis fully and the data file was readable. When the full dataset is generated, it would be useful to have a data-dictionary to accompany the file explain precisely what each column is and how it was coded – as I expect this dataset might be of interest to many people.

Overall, the new proposal is improved, and I still think it is a feasible and worthwhile project. I share Reviewer 1’s concerns about the ability of the project to provide strong answers to the original questions of interest. But even if it fails to answer these questions effectively, then highlighting the barriers around synthesizing evidence like would be a very useful paper. The feasibility of these analyses will depend on the quantity and quality of the overall dataset, which we will only know when it has been extracted. Therefore, it would be useful to conduct a feasibility assessment of the different analyses after the data have been extracted but before conducting them. If IPA is given to the project, I expect a longer Stage 2 review may be necessary than with other registered reports. I still think the project is very interesting and will be of interest those interested in string-pulling tasks but also those interested in evidence synthesis in the field.

Ben

Minor point

1. Code reproducibility
   Code, line 355 there is a comma after the first value in
   > ref = c("HM015213",)
   This should be removed to read
   ref = c("HM015213")
   I was unable to run the code of the phylogenetic analysis – but I assume this just hasn’t been completed yet.

References:

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2020). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. EcoEvoRxiv. https://doi.org/10.32942/osf.io/su4zv

Davies, A. D., Lewis, Z., & Dougherty, L. R. (2020). A meta-analysis of factors influencing the strength of mate-choice copying in animals. Behavioral Ecology, 31(6), 1279–1290. https://doi.org/10.1093/beheco/araa064

Author’s Response to Decision Letter for (RSOS-201728.R1)

See Appendix B.
Do you have any ethical concerns with this paper?
No

Recommendation?
Accept in principle

Comments to the Author(s)
The revised version has addressed all my concerns with the project. I have re-checked the code and the random effects structure now seems appropriate. I missed the Cinar et al. 2020 reference off my last review - here it is in case it’s useful for the authors:

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2020, November 23). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. https://doi.org/10.32942/osf.io/su4zv

I think the project is well-planned and feasible and will be a useful study for the reasons I outlined in previous reviews. I also agree with nearly all of Reviewer 1’s previous comments about the limitations of the existing string-pull data, and the probable ineffectiveness of various methods of combining them. However, I feel this is exactly the type of project that can effectively highlight these problems – problems that generalise across much comparative research. In addition, the systematic review aspect and curated dataset of this current study should be the next best method for comparing these studies and assessing what the data from the string-pull task mean, even if its marred with uncertainty.

Ben

A minor point:

The introduction starts with: “Three major issues in the field of comparative cognition are: 1) that researchers generalise findings from very small samples to a much larger population… the present study will attempt to demonstrate how these problems can be met with modern systematic review, and meta-analytic…”

This risks overstating the problem of small samples and generalizability somewhat – it is a large problem in some areas of compcog, but others seem to do very well with small samples, e.g. animal learning, animal psychophysics. I also think it oversells how well systematic review and meta-analysis can meet this problems -&gt; perhaps “how these problems can be assessed with systematic review and meta-analysis” might be a better phrasing.

Decision letter (RSOS-201728.R2)

We hope you are keeping well at this difficult and unusual time. We continue to value your support of the journal in these challenging circumstances. If Royal Society Open Science can assist you at all, please don’t hesitate to let us know at the email address below.
Dear Mr Keeble,

On behalf of the Editors, I am pleased to inform you that your Manuscript RSOS-201728.R2 entitled "The evolution of coordination: A phylogenetic meta-analysis and systematic review" has been accepted in principle for publication in Royal Society Open Science subject to minor revision in accordance with the referee and editor suggestions. Please find their comments at the end of this email.

The reviewers and handling editors have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the comments and revise your manuscript.

Please you submit the revised version of your manuscript within 7 days (i.e. by the 11-Feb-2021). If you do not think you will be able to meet this date please let me know immediately.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions". Under "Actions," click on "Create a Revision." You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referees and you should upload a file "Response to Referees". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referees.

Full author guidelines can be found here https://royalsocietypublishing.org/rsos/registered-reports.

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Kind regards,
Royal Society Open Science Editorial Office
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Chris Chambers (Subject Editor, Royal Society Open Science)
openscience@royalsociety.org

Associate Editor Comments to Author (Professor Chris Chambers):
Associate Editor: 1
Comments to the Author:
One of the previous reviewers (Reviewer 2) was available to assess the Stage 1 submission. Based on this reviewer's assessment, and my own reading of the revised manuscript, I think the proposal is now more solid and we can soon move forward with Stage 1 in-principle acceptance (IPA). However, before doing so, please attend to the final point by Reviewer 2 concerning claims about the potential implications of the work. I agree with the reviewer on this point, and although this minor change in phrasing could be addressed at Stage 2, it makes sense to get the Introduction into as final a state as possible to minimise any need for later revisions.
Once the authors have made this change, IPA will be awarded without further in-depth Stage 1 review.

Reviewer comments to Author:
Reviewer: 2
Comments to the Author(s)
The revised version has addressed all my concerns with the project. I have re-checked the code and the random effects structure now seems appropriate. I missed the Cinar et al. 2020 reference off my last review - here it is in case it’s useful for the authors:

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2020, November 23). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. https://doi.org/10.32942/osf.io/su4zv

I think the project is well-planned and feasible and will be a useful study for the reasons I outlined in previous reviews. I also agree with nearly all of Reviewer 1’s previous comments about the limitations of the existing string-pull data, and the probable ineffectiveness of various methods of combining them. However, I feel this is exactly the type of project that can effectively highlight these problems – problems that generalise across much comparative research. In addition, the systematic review aspect and curated dataset of this current study should be the next best method for comparing these studies and assessing what the data from the string-pull task mean, even if its marred with uncertainty.

Ben

A minor point:

The introduction starts with: “Three major issues in the field of comparative cognition are: 1) that researchers generalise findings from very small samples to a much larger population… the present study will attempt to demonstrate how these problems can be met with modern systematic review, and meta-analytic…”

This risks overstating the problem of small samples and generalizability somewhat – it is a large problem in some areas of compcog, but others seem to do very well with small samples, e.g. animal learning, animal psychophysics. I also think it oversells how well systematic review and meta-analysis can meet this problems -> perhaps “how these problems can be assessed with systematic review and meta-analysis” might be a better phrasing.

Author's Response to Decision Letter for (RSOS-201728.R2)
See Appendix C.

Decision letter (RSOS-201728.R3)

We hope you are keeping well at this difficult and unusual time. We continue to value your support of the journal in these challenging circumstances. If Royal Society Open Science can assist you at all, please don't hesitate to let us know at the email address below.

Dear Mr Keeble
On behalf of the Editor, I am pleased to inform you that your Manuscript RSOS-201728.R3 entitled "The evolution of coordination: A phylogenetic meta-analysis and systematic review" has been accepted in principle for publication in Royal Society Open Science.

You may now progress to Stage 2 and complete the study as approved. Before commencing data collection we ask that you:

1) Update the journal office as to the anticipated completion date of your study.

2) Register your approved protocol on the Open Science Framework (https://osf.io/rr) or other recognised repository, either publicly or privately under embargo until submission of the Stage 2 manuscript. Please note that a time-stamped, independent registration of the protocol is mandatory under journal policy, and manuscripts that do not conform to this requirement cannot be considered at Stage 2. The protocol should be registered unchanged from its current approved state, with the time-stamp preceding implementation of the approved study design. We strongly recommend using the dedicated Stage 1 RR registration portal at https://osf.io/rr

Following completion of your study, we invite you to resubmit your paper for peer review as a Stage 2 Registered Report. Please note that your manuscript can still be rejected for publication at Stage 2 if the Editors consider any of the following conditions to be met:

- The results were unable to test the authors’ proposed hypotheses by failing to meet the approved outcome-neutral criteria.
- The authors altered the Introduction, rationale, or hypotheses, as approved in the Stage 1 submission.
- The authors failed to adhere closely to the registered study procedures. Please note that any deviations from the approved procedures must be communicated to the editor immediately for approval, and prior to the completion of data collection. Failure to do so can result in revocation of in-principle acceptance and rejection at Stage 2 (see complete guidelines for further information).
- Any post-hoc (unregistered) analyses were either unjustified, insufficiently caveated, or overly dominant in shaping the authors’ conclusions.
- The authors’ conclusions were not justified given the data obtained.

We encourage you to read the complete guidelines for authors concerning Stage 2 submissions at https://royalsocietypublishing.org/rsos/registered-reports#ReviewerGuideRegRep. Please especially note the requirements for data sharing, reporting the URL of the independently registered protocol, and that withdrawing your manuscript will result in publication of a Withdrawn Registration.

Once again, thank you for submitting your manuscript to Royal Society Open Science and we look forward to receiving your Stage 2 submission. If you have any questions at all, please do not hesitate to get in touch. We look forward to hearing from you shortly with the anticipated submission date for your stage two manuscript.

Kind regards,
Royal Society Open Science Editorial Office
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Chris Chambers (Registered Reports Editor, Royal Society Open Science)
openscience@royalsociety.org
Author's Response to Decision Letter for (RSOS-201728.R3)

See Appendices D-F.

RSOS-201728.R4

Review form: Reviewer 2 (Ben Farrar)

Is the manuscript scientifically sound in its present form?
No

Are the interpretations and conclusions justified by the results?
No

Is the language acceptable?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Major revision

Comments to the Author(s)
The authors have completed the study in-line with the Stage 1 submission, with their noted change to the search after finding Animal Behaviour did not accept wild cards in the search string. This could have been overcome by using a search engine (e.g., Scopus, Web of Science) to perform the searches but the authors might not have had access.

Overall I believe the authors have performed a useful study and generated a useful dataset. However I believe the inferences the authors make, and general structure of the results and discussion could be much improved to ensure the value of the study is communicated effectively and that the interpretation is justified.

Overall Structure
I thought the overall results and discussion section could be made easier to follow, especially because it jumps between methodological challenges, quality control and analyses. Could subheadings be used corresponding to each of the objectives set out in the introduction to make the structure of the results clearer to the reader?

Quality control
The authors provided a well presented dataset of both the raw data and the quality control, but I was surprised not to the the reasons for the decisions in the quality control document not also
included. While the Stage 1 submission didn't explicitly state that the reasons and relevant text for the decisions to the quality control questions would be recorded, I thought this would have been performed. For example, the authors state that populations were rarely representative, but understanding how this decision was made is important for the quality of the dataset (i.e., have a column in the datafile with the same description from the full text extract and a comment on why the decision has been made). If these data weren't extracted, I think it would be really valuable for the authors to extract them and add them to the dataset for each question. Commenting on why the decision has been made for internal validity seems particularly important (and what criteria were screened for this), especially as the authors make the claim that "studies of animal coordination are largely internally valid". This is not necessarily true, rather what the authors did find was that one coder assessing a small number of string pulling tasks did not find reasons to doubt their internal validity.

Conclusions

Another conclusion the authors made surprised me. About publication bias, the authors said their analysis suggested little publication bias in studies assessing animal co-ordination. This surprised me as I know of unpublished studies in this area, and that the statistical power of their analyses to detect publication bias is very low. If their overall sample size is 17 included studies, what number of unpublished similar studies would mean that there is little publication bias? Even 4 or 5 unpublished studies would be a lot here.. If they wanted to make the conclusion about little publication bias then I think power analyses to see the degree of e.g. asymmetry their analyses could detect would be necessary. Alternatively I see no issue in highlighting how difficult it can be to assess publication in very heterogenous datasets with small sample sizes (this is the conclusion I would draw, with a strong prior that some studies will be missing).

Discussion

The authors discussion focuses on co-ordination tasks and what their data can tell us about string-pulling. As the other reviewer at Stage 1 predicted, this might not be so much. I think a section of the discussion should be devoted to the challenges the authors faced in collecting and synthesising data in this project, as I feel this is one of the major contributions it makes.

Minor Comments

Reasons for exclusion that are given in the text (Lines 442 - 450) also be presented in the PRISMA diagram

Could the raw agreement also be included in Table 2, i.e., 5/5 statements extract the same, or 3/5? I think this would be easier for people to understand given the small number of double coded variables.

Could the quality control statements also be written in Table 1

Decision letter (RSOS-201728.R4)

We hope you are keeping well at this difficult and unusual time. We continue to value your support of the journal in these challenging circumstances. If Royal Society Open Science can assist you at all, please don't hesitate to let us know at the email address below.

Dear Mr Keeble:
On behalf of the Editor, I am pleased to inform you that your Stage 2 Registered Report RSOS-201728.R4 titled "The evolution of coordination: A phylogenetic meta-analysis and systematic review" has been deemed suitable for publication in Royal Society Open Science subject to minor revision in accordance with the referee suggestions. Please find the referees' comments at the end of this email.

The reviewers and Subject Editor have recommended publication, but also suggest some minor revisions to your manuscript. We invite you to respond to the comments and revise your manuscript. Below the referees' and Editors' comments (where applicable) we provide additional requirements. Final acceptance of your manuscript is dependent on these requirements being met. We provide guidance below to help you prepare your revision.

Please submit your revised manuscript and required files (see below) no later than 21 days from today's (ie 21-Dec-2021) date. Note: the ScholarOne system will 'lock' if submission of the revision is attempted after the deadline. If you do not think you will be able to meet this deadline please contact the editorial office immediately.

Please note article processing charges apply to papers accepted for publication in Royal Society Open Science (https://royalsocietypublishing.org/rsos/charges). Charges will also apply to papers transferred to the journal from other Royal Society Publishing journals, as well as papers submitted as part of our collaboration with the Royal Society of Chemistry (https://royalsocietypublishing.org/rsos/chemistry). Fee waivers are available but must be requested when you submit your revision (https://royalsocietypublishing.org/rsos/waivers).

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Kind regards,
Royal Society Open Science Editorial Office
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Chris Chambers
(Registered Reports Editor, Royal Society Open Science)
openscience@royalsociety.org

Associate Editor Comments to Author (Professor Chris Chambers):
Associate Editor: 1
Comments to the Author:
One of the two original Stage 1 reviewers kindly returned to evaluate the Stage 2 submission, and I have decided that this reviewer's evaluation together with my own reading is sufficient for us to proceed with an interim editorial decision. As you will see, the reviewer is broadly positive about the completed manuscript, while also offering a range of helpful suggestions for revision, including structural changes to the reporting of results, documentation of the data (particularly in relation to the quality assessment), and more in depth consideration of interpretative limitations in the Discussion (as well as methodological lessons learned).

I concur with this assessment and overall feel that you have completed a very careful and thorough investigation. Provided you are able to respond comprehensively to the points raised by the reviewer, final acceptance should be forthcoming without requiring further in-depth review.
Comments to Author:
Reviewer: 2

Comments to the Author(s)
The authors have completed the study in-line with the Stage 1 submission, with their noted change to the search after finding Animal Behaviour did not accept wild cards in the search string. This could have been overcome by using a search engine (e.g., Scopus, Web of Science) to perform the searches but the authors might not have had access.

Overall I believe the authors have performed a useful study and generated a useful dataset. However I believe the inferences the authors make, and general structure of the results and discussion could be much improved to ensure the value of the study is communicated effectively and that the interpretation is justified.

Overall Structure

I thought the overall results and discussion section could be made easier to follow, especially because it jumps between methodological challenges, quality control and analyses. Could subheadings be used corresponding to each of the objectives set out in the introduction to make the structure of the results clearer to the reader?

Quality control

The authors provided a well presented dataset of both the raw data and the quality control, but I was surprised not to the the reasons for the decisions in the quality control document not also included. While the Stage 1 submission didn’t explicitly state that the reasons and relevant text for the decisions to the quality control questions would be recorded, I thought this would have been performed. For example, the authors state that populations were rarely representative, but understanding how this decision was made is important for the quality of the dataset (i.e., have a column in the datafile with the same description from the full text extract and a comment on why the decision has been made). If these data weren’t extracted, I think it would be really valuable for the authors to extract them and add them to the dataset for each question. Commenting on why the decision has been made for internal validity seems particularly important (and what criteria were screened for this), especially as the authors make the claim that "studies of animal coordination are largely internally valid". This is not necessarily true, rather what the authors did find was that one coder assessing a small number of string pulling tasks did not find reasons to doubt their internal validity.

Conclusions

Another conclusion the authors made surprised me. About publication bias, the authors said their analysis suggested little publication bias in studies assessing animal co-ordination. This surprised me as I know of unpublished studies in this area, and that the statistical power of their analyses to detect publication bias is very low. If their overall sample size is 17 included studies, what number of unpublished similar studies would mean that there is little publication bias? Even 4 or 5 unpublished studies would be a lot here. If they wanted to make the conclusion about little publication bias then I think power analyses to see the degree of e.g. asymmetry their analyses could detect would be necessary. Alternatively I see no issue in highlighting how difficult it can be to assess publication in very heterogenous datasets with small sample sizes (this is the conclusion I would draw, with a strong prior that some studies will be missing).

Discussion
The authors' discussion focuses on co-ordination tasks and what their data can tell us about string-pulling. As the other reviewer at Stage 1 predicted, this might not be so much. I think a section of the discussion should be devoted to the challenges the authors faced in collecting and synthesising data in this project, as I feel this is one of the major contributions it makes.

Minor Comments
Reasons for exclusion that are given in the text (Lines 442 - 450) also be presented in the PRISMA diagram.

Could the raw agreement also be included in Table 2, i.e., 5/5 statements extract the same, or 3/5? I think this would be easier for people to understand given the small number of double coded variables.

Could the quality control statements also be written in Table 1?

===PREPARING YOUR MANUSCRIPT===

Your revised paper should include the changes requested by the referees and Editors of your manuscript.

You should provide two versions of this manuscript and both versions must be provided in an editable format:
- one version should clearly identify all the changes that have been made (for instance, in coloured highlight, in bold text, or tracked changes);
- a 'clean' version of the new manuscript that incorporates the changes made, but does not highlight them. This version will be used for typesetting.

Please ensure that any equations included in the paper are editable text and not embedded images.

Please ensure that you include an acknowledgements' section before your reference list/bibliography. This should acknowledge anyone who assisted with your work, but does not qualify as an author per the guidelines at https://royalsociety.org/journals/ethics-policies/openness/.

While not essential, it will speed up the preparation of your manuscript proof if you format your references/bibliography in Vancouver style (please see https://royalsociety.org/journals/authors/author-guidelines/#formatting). You should include DOIs for as many of the references as possible.

If you have been asked to revise the written English in your submission as a condition of publication, you must do so, and you are expected to provide evidence that you have received language editing support. The journal would prefer that you use a professional language editing service and provide a certificate of editing, but a signed letter from a colleague who is a proficient user of English is acceptable. Note the journal has arranged a number of discounts for authors using professional language editing services (https://royalsociety.org/journals/authors/benefits/language-editing/).

===PREPARING YOUR REVISION IN SCHOLARONE===

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre - this may be accessed by clicking on "Author" in the dark toolbar at the top of the
 Attach your point-by-point response to referees and Editors at the 'View and respond to decision letter' step. This document should be uploaded in an editable file type (.doc or .docx are preferred). This is essential, and your manuscript will be returned to you if you do not provide it.

Please ensure that you include a summary of your paper at the 'Type, Title, & Abstract' step. This should be no more than 100 words to explain to a non-scientific audience the key findings of your research. This will be included in a weekly highlights email circulated by the Royal Society press office to national UK, international, and scientific news outlets to promote your work. An effective summary can substantially increase the readership of your paper.

At the 'File upload' step you should include the following files:
-- Your revised manuscript in editable file format (.doc, .docx, or .tex preferred). You should upload two versions:
  1) One version identifying all the changes that have been made (for instance, in coloured highlight, in bold text, or tracked changes);
  2) A 'clean' version of the new manuscript that incorporates the changes made, but does not highlight them.
-- An individual file of each figure (EPS or print-quality PDF preferred [either format should be produced directly from original creation package], or original software format).
-- An editable file of each table (.doc, .docx, .xls, .xlsx, or .csv).
-- An editable file of all figure and table captions.
Note: you may upload the figure, table, and caption files in a single Zip folder.
-- Any electronic supplementary material (ESM).
-- If you are requesting a discretionary waiver for the article processing charge, the waiver form must be included at this step.
-- If you are providing image files for potential cover images, please upload these at this step, and inform the editorial office you have done so. You must hold the copyright to any image provided.
-- A copy of your point-by-point response to referees and Editors. This will expedite the preparation of your proof.

At the 'Details & comments' step, you should review and respond to the queries on the electronic submission form. In particular, we would ask that you do the following:
-- Ensure that your data access statement meets the requirements at https://royalsociety.org/journals/authors/author-guidelines/#data. You should ensure that you cite the dataset in your reference list. If you have deposited data etc in the Dryad repository, please only include the 'For publication' link at this stage. You should remove the 'For review' link.
-- If you are requesting an article processing charge waiver, you must select the relevant waiver option (if requesting a discretionary waiver, the form should have been uploaded, see 'File upload' above).
-- If you have uploaded any electronic supplementary (ESM) files, please ensure you follow the guidance at https://royalsociety.org/journals/authors/author-guidelines/#supplementary-material to include a suitable title and informative caption. An example of appropriate titling and captioning may be found at https://figshare.com/articles/Table_S2_from_Is_there_a_trade-off_between_peak_performance_and_performance_breadth_across_temperatures_for_aerobic_scope_in_teleost_fishes_/3843624.

At the 'Review & submit' step, you must view the PDF proof of the manuscript before you will be able to submit the revision. Note: if any parts of the electronic submission form have not been
completed, these will be noted by red message boxes - you will need to resolve these errors before you can submit the revision.

**Author's Response to Decision Letter for (RSOS-201728.R4)**

See Appendix G.

**Decision letter (RSOS-201728.R5)**

We hope you are keeping well at this difficult and unusual time. We continue to value your support of the journal in these challenging circumstances. If Royal Society Open Science can assist you at all, please don't hesitate to let us know at the email address below.

**Dear Mr Keeble:**

It is a pleasure to accept your Stage 2 Registered Report entitled “The evolution of coordination: A phylogenetic meta-analysis and systematic review” in its current form for publication in Royal Society Open Science.

Please remember to make any data sets or code libraries 'live' prior to publication, and update any links as needed when you receive a proof to check - for instance, from a private 'for review' URL to a publicly accessible 'for publication' URL. It is good practice to also add data sets, code and other digital materials to your reference list.

Our payments team will be in touch shortly if you are required to pay a fee for the publication of the paper (if you have any queries regarding fees, please see https://royalsocietypublishing.org/rsos/charges or contact authorfees@royalsociety.org).

The proof of your paper will be available for review using the Royal Society online proofing system and you will receive details of how to access this in the near future from our production office (openscience_proofs@royalsociety.org). We aim to maintain rapid times to publication after acceptance of your manuscript and we would ask you to please contact both the production office and editorial office if you are likely to be away from e-mail contact to minimise delays to publication. If you are going to be away, please nominate a co-author (if available) to manage the proofing process, and ensure they are copied into your email to the journal.

Please see the Royal Society Publishing guidance on how you may share your accepted author manuscript at https://royalsociety.org/journals/ethics-policies/media-embargo/. After publication, some additional ways to effectively promote your article can also be found here https://royalsociety.org/blog/2020/07/promoting-your-latest-paper-and-tracking-your-results/.

Thank you for your fine contribution. On behalf of the Editors of Royal Society Open Science, we look forward to your continued contributions to the journal.

**Kind regards,**
Royal Society Open Science Editorial Office
Royal Society Open Science
Appendix A

We would first like to thank the editor for their recognition that this work has the potential for publication in Royal Society Open Science, and also the reviewers for their insightful comments regarding how this study can be improved. Given that such detailed comments were provided, and because we wanted to make sure we are replying to them in full, comments will be reproduced below with replies to comments indicated by bold font. Altered text in the manuscript has been highlighted, and reference to lines is made in this document where necessary.

1 Replies to reviewer 1

Although in principle I applaud the authors’ goal of trying to bring together the results of many different cooperation studies with many different species, I am less optimistic this is going to bring forward the field of cooperation in comparative psychology for the reasons I expose below.

In particular, the study proposes the following hypotheses and questions:

1. Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

The delayed partner string-pull paradigm has produced positive results in a few species (but more species than the ones mentioned in this manuscript). However, this task is not a silver bullet with regards to what we can learn about animals’s skills to cooperate. As it has been argued by other researchers (e.g. Seed & Jensen, 2011) there are several simple explanations that can explain how animals learn to pass these tests.

In light of this comment we have made the implications of this study clearer in the introduction, and changed terminology where appropriate from ‘cooperative success’ to ‘coordination success’, to avoid misunderstanding. We acknowledge more forcefully that there are several competing explanations for success in the string-pull task, and acknowledge that these may limit the conclusions of our study.
In addition, the number of trials is not the only variable that the different studies have changed. Different studies have also worked with different “delays” and different lengths of the rope, factors which can all influence whether or not and how quickly animals can succeed in this task. Therefore, I am afraid this will be a complicated analysis.

We acknowledge the concern of the reviewer that studies vary significantly in their design. Because of this concern, much detail has been added to the search protocol, study inclusion criteria, quality assessment and data extraction (sections 1.1, 1.2, 1.3, 1.4), and we hope that, because of this, the studies included in our analyses will be more specific to a task design. This will limit our conclusions, but hopefully make the conclusions we can draw more reliable. Furthermore, the aims of the study outlined in the introduction have been reframed so that one of our primary aims for the present study is to highlight the problems that variation in methods present for conducting the comparative meta-analyses necessary to make strong evolutionary conclusions, and thus provide an illustration of the concern the reviewer rightfully raises (lines 76-90, and section 1.3).

(2) Does inhibition predict cooperative success in a delayed partner string-pull paradigm? For the reasons explained above, there are actually not many species that have succeeded in the delay task under comparable conditions, so the N will be rather low and one would be comparing learning skills under very different conditions. In addition, the inhibition measure would be a species’ inhibition measure but not a measure obtained from the same populations that participated in the cooperation tasks. This could be acceptable if everything else was highly controlled. However, it is just another source of noise adding to the other ones, and when dealing with a rather low N.

This analysis has been removed from the study due to concerns expressed by both reviewers. We acknowledge that the data used in the analysis is likely to unreliable to provide any reliable conclu-
sions. Furthermore, as reviewer 1 points out, inhibition measures and measures of success in a coordination task would have been from different populations. The best way to assess this question would likely be to conduct further coordination experiments and including a measurement of individual inhibition, before comparing the relationship between individual inhibition and success in the coordination task across species. At the moment, the data is not reliable enough to assess this question [2].

(3) Do social dynamics (affiliation and tolerance) predict cooperative success? This is possibly the question I am more positive about, just because there are more studies that have looked at tolerance and affiliation. Nevertheless, I would emphasise that cooperation success, as it was measured in many of the studies, does not equate with an understanding of the contingencies and role of the partner in the cooperative endeavour. In other words, one would be looking at something about dyads’ relationships and capacity to interact manipulating together food sources but not *collaborative skills* per se.

(4) Do brain size, social group size, or dietary factors predict cooperative understanding? As mentioned above, success in the delay task alone does not mean cooperative understanding. The operationalization of cooperative understanding is more complicated. This study proposal mentions gaze-following too, but that also is not necessarily a good dependent measure: we know that subordinates are often nervous in the presence of dominants when manipulating food rewards, which could lead to more “monitoring” behaviour.

The aims of the final two analyses have been reframed in terms of ‘ability to coordinate in a delay string-pull task’ as opposed to ‘understanding of cooperative problem’, given that the latter terminology may be misleading (see lines 95-100, and terminology changes throughout). As reviewer 1 has previously pointed out, there are competing explanations for success in the string-pull task, and given that subjects could still be successful in this task despite not fully ‘understanding’ the task, whatever that may mean, we agree that
framing our questions and measure in terms of ‘understanding’ was inappropriate. We hope that changing our terminology will, in this case, reframe the analysis in terms of the relationship between social dynamics and success in a task, or the relationship between certain evolutionary predictors and success in a task, without making any claims regarding whether or not animals are ‘understanding’ the task and all its intricacies. Gaze following has also been removed as a measure of ‘understanding’, as this could be a measure of many different things.

Other comments:

- In the pilot data presented, I noted that Chalmeau 1994 and Suchak et al. 2014 are two of the studies analysed but those studies did not use the same string-pulling task mentioned in the introduction as necessary to facilitate any conclusions about cooperation understanding. The apparatus of those studies were different to the one used in all the delay tasks, so although this may be acceptable, one needs to acknowledge that the contingencies for success with that apparatus are different ones. Individuals can pull alone without losing the opportunity to access the rewards, so waiting in that task is not necessarily needed. If the understanding measure used is “monitoring” behaviour, I am sceptical about it for the reasons mentioned above.

The authors hope that a more developed search protocol (section 1.1), and a more thorough inclusion protocol (sections 1.2 and 1.3) based on the PRISMA guidelines and diagram [3] will remove studies like those mentioned by reviewer 1 that are not as relevant to the questions or measures outlined in the present study proposal, but were included in pilot data due to a less refined protocol for inclusion. Furthermore, as mentioned above, measures of ‘understanding’ have been reframed as measures of ‘success in a coordination task’, to avoid misunderstanding.
• How informative is the correlation - number of trials and cooperation success? Certainly, with more experience cooperative partners will become more coordinated, but (1) are there significant differences in how quickly different species reach success? (2) Are the levels of success comparable across species? i.e. some species quickly being able to wait for up to 25 sec for the partners whereas others the maximum they wait is 5sec? (3) how is success being measured in those studies using different pulling apparatuses? Because co-acting is not necessarily cooperating.

The data extraction protocol has been amended so that variables measuring trials to success, delay time used in delay-release task, and whether the task used is delay- or simultaneous- release task are recorded from studies where possible (see lines 227-235). However, data may not be available from studies to conduct any statistical analysis on such data. As a result, a new aim of the present study will be to provide a comprehensive dataset that allow researchers to access and compare these aspects of string-pull studies across species (see lines 76-90). We hope that this will help researchers design future studies that are directly comparable to existing results, as well as illustrate the difficulties encountered when methods are used with slight variations.

2 Replies to reviewer 2

Comments to the Author(s) Keeble, Wallenberg and Price propose to meta-analyses and conduct a phylogenetic analysis of data from the co-operative string-pulling literature. I think the proposal is exactly the type of approach the field needs to explore more regarding evidence synthesis, and I think it’s great that they’ve chosen to submit this as a registered report. However, as it stands, I think parts the protocol needs much greater specification, and there are several problems the researchers are likely to face when conducting this
meta-analysis which might prevent it from providing strong answers to their questions. Nevertheless, I see a large amount of merit in the proposal. I think that if the authors reframed it more in terms of exploring some of the difficulties in performing comparative meta-analyses, as well as focusing on the substantive questions about co-operation, then it could make a very nice RR. I’ve framed my comments around each of the proposed analyses:

We thank reviewer 2 for their encouraging comments. We also acknowledge, and are grateful for, the insight that this registered report provides an opportunity for us to explore the difficulties that the application of meta-analysis to the field presents, and an opportunity to create a valuable data resource for other interested researchers. As such we have attempted to reframe this registered report as a valuable exploration of the difficulties of research synthesis in comparative cognition, while simultaneously allowing us to collate multiple aspects of studies that will hopefully illuminate gaps in the literature and help direct future research, and assess their rigour using systematic review.

1. Does number of trials predict cooperative success in a delayed partner string-pull paradigm?

Could the authors specify more about how they will define the number of trials when extracting data, and similarly how they will define success on the string-pulling paradigm. I can see the definition they provide in the caption to Figure 1, but this could be detailed in text.

With trial data, how will the authors deal with differing numbers of training trials and testing trials overall – i.e., if they define the number of trials as the number of testing trials, what happens if studies have markedly different numbers of training trials – or if this information is not available? Similarly, how will the authors address cases where animals are tested sequentially with different partners, potentially across multiple studies, some of which may not have been published?

Differences between testing and training trials will be assessed
using quality assessment during systematic review (see section 1.3 and 1.4). For meta analysis, estimates assessing the relationships between number of trials and coordination success will only be included if that measure was obtained using overall number of trials (both training and test) a subject received (see lines 161-164).

Overall, I think the exact extraction procedure could be specified in greater detail, and the procedure might need to be expanded such that the authors can document the full extent of heterogeneity between the studies. Understanding this heterogeneity is key to understanding what the output of the meta-analyses means – the heterogeneity might be so large that the aggregate point estimates/CIs don’t have much meaning – and I imagine most people would have high priors that that more trials leads to greater success anyway. I still think there is a lot of merit in collecting and summarizing this data, and the meta-analysis workflow will help this, but I think the main benefit of this proposal is in collecting and presenting information about individual study results and the between-study heterogeneity.

We hope that, by including a quality assessment stage (section 1.3) in this systematic review, we can include a critical summary of individual study results as a major part of the results section of this review. Furthermore, we have elaborated on analysing not only all studies together, but also include subgroup analyses of species, in an attempt to summarise not only findings and heterogeneity across and between all studies all studies, but also analyse findings and heterogeneity between and within species (see lines 283-302).

I have further questions about the models the authors are using for the meta-analysis, and how they will deal with repeated data from individuals (e.g., in the pilot report “Peron et al. 2011 Grey parrots” appears with three separate estimates – if I’ve found the paper correctly this comes from 3 parrots tested in 3 consecutive experiments).

Data from studies conducting analysis at individual and group levels will be analysed both separately and together, and where indi-
viduals have been tested repeatedly, their estimates will be averaged across those tests (See lines 275-281).

As it stands, I don’t think the current meta-analysis model accounts for phylogenetic relatedness – they might want to consider fitting a multi-level model which incorporates relatedness in (see e.g. Dougherty & Guillette, 2018 as an example), unless there is a reason why they would prefer to not include this information, but this should be explained.

If the present article goes through a revision, it would be nice if the pilot data, extraction protocol and code could be made available to help understand the procedures and models better.

Pilot data and code for both pilot analyses and proposed final analysis have been included for review. In light of reviewer 2’s comments, models have been made multi-level to account for phylogenetic relatedness and also data from studies that appear repeatedly in the dataset (see lines 266-275).

2. Does inhibition predict cooperative success in a delayed partner string-pull paradigm?

This analysis proposes to see if inhibition scores from MacLean et al.’s study predicts success in the string-pulling. I have strong reservations about the ability of this analysis to have the statistical power to produce a meaningful analysis, most of all because the between-site replicability of the MacLean et al. data is largely untested, and where it has been tested it appears low – i.e., what have been billed as species differences in for example the cylinder task performance may not be so, (see Figure 3 and the following discussion in Farrar et al., 2020). Because the replicability of inhibition data may be low (as well as for the string pulling data) – this analysis might end up predicting noise with noise. I’m not completely opposed to the analysis being performed, but I think it should be heavily caveated if so. If the authors do decide to proceed, I’d encourage them to search further for inhibition data as there are many studies that use the same tasks as MacLean et al., to ensure they are getting as much data as possible to inform the analysis.
I’m also uncertain about what the best way to interpret the results of this analysis are. On the one hand, correlation does not mean causation, and there are likely many reasons why data from the inhibition task will correlate with data from the string pulling task, without there being a causal relationship between them. On the other hand, I think again it’s incredibly likely that inhibition is causally related to some aspects of passing the string pulling task. If the authors did not find a positive result then I’d be pretty confident that this would be a false negative (i.e., some relationship between inhibition and string-pulling performance is very likely to exist), and if they do find the positive effect, I’m unsure whether the aggregated numerical estimates would mean very much, because of the heterogeneity between studies.

This analysis has been removed from the present study. Both reviewers rightly point out that using scores at the species level is inappropriate given that there is likely a lot of variation between individuals, and that the potential unreliability of the measures from the MacLean et al. [4] study means the analysis may be, at this point in time, largely redundant and too difficult to interpret.

3. Do social dynamics (affiliation and tolerance) predict cooperative success

4. Do brain size, social group size, or dietary factors predict cooperative understanding?

I have similar reservations about these analyses to the previous two. Again, I think it’s highly likely that social dynamics will to some extent predict performance on a social co-operation task (with both causal and non-causal relationships), but it is also possible that the data quality/quantity are not there to perform meaningful quantitative analyses on these. I think the analyses will be useful, but interpreting their outputs may be difficult from a theoretical perspective.

When constructing the phylogenetic trees, how will the authors quantify the uncertainty in the output? They will be able to generate a tree that is the best fit according to some criteria, but if the overall data quality are poor, the tree might be quite likely to be inaccurate, i.e., there is no error control.
We recognise that the analyses proposed in this paper are likely to be difficult to interpret based on concerns of both reviewers. One primary cause of this is likely high variability in study designs. However, we hope that the now extended quality assessment protocol (section 1.3) will allow us to discuss where issues arise with comparability across studies and demonstrate the difficulties of research synthesis in comparative cognition, as well as help us include only the most comparable estimates in meta analysis. To assess the reliability of phylogenetic trees, the r package phangorn \cite{5} will be used. Trees will be compared against each other, and the tree with the lowest AIC will be applied in phylogenetic generalised least squares models (see lines 246-262).

My last comments are not specific to a particular analysis:

5. The overall aim of the study

The proposal aims to make “conclusions about the evolutionary origins of cooperative problem solving” by analysing data from the co-operative string-pulling task. However, the generalizability of results from the string-pulling task to other tests (real or hypothetical) of co-operative behaviour is relatively unknown, and as such framing a meta/phylogenetic analysis of string-pulling data as an analysis of co-operative behaviour in general is excessive, in my view. The data will be relevant to this question, but not strong. It could instead be framed more narrowly around the string-pulling task.

Both reviewers rightly acknowledge that our claims about the conclusions that could be drawn from the proposed analysis are excessive. In response, we have narrowed our proposed aims and potential conclusions to success in a coordination task, and removed claims previously made that our analysis can tell us much about ‘understanding and problem-solving’ capabilities. Especially given that success in a coordination task may be a result of factors other than social cognition and/or ‘understanding’ \cite{1}, and given that ‘understanding’ and ‘problem solving’ are ill defined concepts in themselves. As such, we
have replaced these phrases with ‘coordination success’ and ‘coordination ability’ where relevant.

6. The search protocol and quality control

The search protocol and extraction procedure could be specified to a much greater extent. One of the strongest outcomes of this study is a high-quality data resource on the string-pulling task. It’s important then to ensure that the search and extraction process are comprehensive and high quality. I’d recommend the authors to follow the PRISMA guidelines for reporting, and to add in several quality control stages to their extraction procedure. Regarding the search terms: searching for comparative data can be very difficult due to mass heterogeneity in how researchers report studies, and I think the Google scholar search and citation-based search will identify most studies. However, it would be good to have some verification stages here to check that key studies are not missing. It would be good if the search structure was specified more, too – for example, are the searches simply performed by inputting the four statements into google scholar as they are, or by using truncation and wildcards (e.g. searching for “co-operat*”), will they search for different combinations and spellings (cooperation as well as co-operation). It’s possible that the authors have already considered this but it would be great if the full structure of the search could be given. The authors may wish to complement their very general search of the entirety Google scholar with more targeted searches of specialist animal behaviour journals as a method of checking that all relevant publications have been identified.

The search protocol has been specified in greater detail, including truncated terms and a repetition of the search protocol conducted in specialist animal behaviour/cognition journals. Furthermore, a quality assessment stage has been included to address the problem of differences in methods and methodological rigour across studies (see sections 1.1 and 1.3).

Could the inclusion criteria be specified further still – will the authors include conference abstracts/thesis chapters/non-published articles? I think they definitely should, but they might want to flag these studies when examining
publication bias.

More detail has been added to the inclusion criteria (sections 1.1, 1.2). The study will include theses and abstracts, and will also approach principal investigators of existing studies to attempt to ascertain unpublished data and results.

In the extraction process, it is unclear what exactly is being extracted and how. It would be great to build a more comprehensive search protocol, and in particular include details on the study design (training trials, sample sizes, sequential testing etc.) and identify studies in which the participants may overlap. I think it is important to have some form of quality control in this process too, either through double extracting a decent percentage of the studies to assess inter-extractor agreement, or more preferable, to have each extraction checked by another researcher.

Data extraction will be conducted twice by the primary author of the study, with a week separating the two extraction phases. 10% of extractions will be repeated by another author on the paper to assess inter-extractor reliability. Details on study design have been included in the data extraction protocol (see lines 228-235), in an attempt to facilitate the aim of creating a useful data source for other researchers on the subject of coordination tasks.

7. Detecting publication bias

I think it’s great the authors are paying attention to publication bias in the analysis, and I think the approach could be expanded to examine publication bias in some of the other data they collect too, e.g., by examining the p-value distributions of reported correlations between tolerance and success on the string pulling task. Currently, their interpretation of the funnel plot and significant Egger’s regression test is possibly excessive – a statistically significant asymmetry does not “demonstrate [as in prove] that there are several studies missing”, although it is compatible with that (see e.g. Sterne et al., 2011). There are reasons why comparative datasets from across many species may have asymmetric funnels – for example what if the smaller sample studies - with larger associ-
ated errors – are performed on species that are disproportionately likely to pass the test (e.g. a small number of elephants or chimpanzees), whereas the larger sample studies are performed on species that might be less likely to pass the test (e.g. dogs or monkeys). I don’t necessarily agree with this reasoning, and think the funnel test is good evidence of publication bias, but currently I believe its overinterpreted. I think it’s incredibly difficult to assess publication bias in comparative datasets because of the massive heterogeneity in the laboratories and species publishing and performing the tasks. I’d encourage the authors to examine subgroups (e.g. by colour coding their funnel plots by species/groups), particularly if they get a decent amount of data on any one particular group. Finally, the authors may wish to consider alternative methods of detecting publication bias – for example by identifying thesis chapters or conference abstracts that have not been published, or even by simply asking researchers in the field if they have performed any string-pulling tasks that have gone unpublished. Some information like this would complement the statistical analysis massively, in my opinion.

We would like the to thank reviewer 2 for acknowledging and encouraging our attempt to assess publication bias in this area of research. We have added further tests of publication bias (Orwin’s fail safe N, an assessment of p-value distribution, cumulative meta-analysis ordered by sample size, and a meta-regression model assessing differences in findings between published and unpublished data if possible). Furthermore, these tests will be repeated with species subgroups should sufficient data be available to do so. See lines 312-334.

Overall, I think this will be a really valuable study. I think the first aim should be in creating a well-document data resource around the string pulling task, from which they will be able to assess how much information the meta-analyses and phylogenetic analysis can provide. There are several key problems to navigate for the analyses (heterogeneity, structure of the models, data quality) and it’s tricky to review this without having much of the information about
the studies yet. This will only be available after extractions have been completed, and so I think it is really important if IPA is to be given to know the extraction process will be very high quality, and that a critical assessment of the strengths of the meta-analyses will be performed after this, or to have another review/discussion about the proposed analyses after extractions have been completed. If it turns out that the data are too low in quality (or in quantity, or homogeneity) to perform strong analyses, then this information is still very valuable to the field, and the process by which the authors will decide this will be a useful methodological advance in the field.

We thank reviewer 2 for their insightful suggestions. We have attempted to reframe this report so that the collection and quality assessment of coordination studies, and data from these studies, are a primary aim. We hope that the remaining analyses will serve as a demonstration of the difficulties that research synthesis of comparative data presents, but also provide a useful demonstration of the opportunities available to comparative researchers should more reliable data be available/collected.

References

[1] A. M. Seed, K. Jensen, Large-scale cooperation, Nature 472 (7344) (2011) 424–425.

[2] B. Farrar, K. Voudouris, N. Clayton, Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research.

[3] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, et al., Preferred reporting items for systematic reviews and meta-analyses: the prisma statement, PLoS med 6 (7) (2009) e1000097.

[4] E. L. MacLean, B. Hare, C. L. Nunn, E. Addessi, F. Amici, R. C. Anderson, F. Aureli, J. M. Baker, A. E. Bania, A. M. Barnard, et al., The evolution
of self-control, Proceedings of the National Academy of Sciences 111 (20) (2014) E2140–E2148.

[5] K. P. Schliep, phangorn: phylogenetic analysis in r, Bioinformatics 27 (4) (2011) 592–593.
Appendix B

Comments from the editor and reviewers have been reproduced here so that we can attempt to address them all in a thorough manner. Our responses are indicated in bold.

1 Associate Editor Comments to Author (Professor Chris Chambers)

The revised manuscript was returned to the two original reviewers. Both reviewers offer a range of comments on the updated proposal. Reviewer 2 is broadly positive but also notes several methodological issues requiring further clarification. As in the first round, Reviewer 1 is again the more critical of the reviewers, questioning the overarching rationale for the research and the heterogeneity of the studies included in the meta-analysis. These are major issues that will need to be thoroughly addressed to achieve IPA.

Given the progress made in improving the design, and the enthusiasm of Reviewer 2, I want to give the authors a final opportunity to address these concerns before issuing a final Stage 1 editorial decision.

We are grateful to the editor for their further consideration of the present registered report, and will aim to update the manuscript in light of further comments from reviewers. We would also like to express our thanks to both reviewers for their very useful comments and suggestions, and we hope we have addressed them appropriately.

2 Reviewer: 1

Although this revised version addresses some of the issues I raised in the previous round, I still remain unconvinced that a phylogenetic meta-analysis in this field of research can currently bring the field forward. The reasons are, as I commented before, the following:

- There are very few species which have passed the delay task (i.e. chim-
panzees, elephants, maybe dogs and dolphins), so with such scarce evidence, I am not sure what is the point of a phylogenetic analysis. In addition, it is now suggested to constrain more which studies enter the analysis to make the comparison more valid (due to the many different testing and experimental protocols), but how many studies would we be left with? If currently only 2-3 species have been considered to pass the task?

• In light of comments from both reviewer 1 and reviewer 2, we have decided to no longer require studies to adhere to certain quality assessment criteria for their inclusion in statistical analyses. As both reviewer 1 and 2 have noted, doing so would likely have resulted in too few studies being included in such an analysis. Furthermore, we hope that our revised introduction and rationale demonstrate that the systematic review (especially the quality assessment) aspect of the present study and the actual collation of results from animal coordination studies is a primary aim (see lines 78-156). There may indeed be too few studies and findings to conduct a thoroughly reliable phylogenetic analysis after data synthesis (in which case, such an analysis will not be conducted), but we hope that the data synthesis itself, and the assessment of the quality of the studies that can be included, will be a useful resource for the field in directing researchers towards conducting studies that are necessary for asking phylogenetic questions. The feasibility of phylogenetic analyses will be assessed post data collection, and will not be conducted if certain constraints are not met (see lines 405-409 and 417-420). However, we believe that starting to synthesise data towards this end is still a valuable pursuit for the field.

• In this revised version Cooperation has been substituted with "coordination", but evidence for coordination can only come from studies showing
active efforts to synchronise actions in time and space. Therefore, many studies using other cooperation apparatuses are less indicative of coordination (Suchak et al. Mendres and De Waal, 2003..) because subjects can pull indiscriminately until the apparatus moves (co-occurring when another subject has joined). However, if I understand correctly, these different studies would also enter the analysis. I don’t think by changing the word to "coordination", authors can circumvent the difficulty of making conclusions about the mechanisms that bring success about in these co-acting tasks.

- We hope that a clarification of rationale in the introduction demonstrates that making conclusions about mechanisms is not a primary goal of the present research. Rather, the primary goal is to assess the plausibility and validity of evidence synthesis in the field of comparative cognition/psychology (see line 121-156). Furthermore, inclusion criteria has been amended so that only studies using apparatus where the task is failed if only one individual pulls are included, which we hope addresses a large source of heterogeneity between studies that may be included in analysis (see lines 201-204). We hope this addresses reviewer 1’s comment that our study would not have extracted data from coordination paradigms only.

- The question about the quality of the relationship between subjects and success in a cooperation task is interesting, but it has already been shown experimentally in several species (e.g. Capuchin monkeys, De Waal & Davis, 2003; Tonquean macaques, Petit et al. 1992; Chimpanzees, Melis et al. 2006; Rooks, Seed et al. 2008). These previous studies have shown the effect of the relationship between partners on cooperative problem solving behaviour *within* a species and *across* species.

The introduction has been amended to acknowledge this point from Reviewer 1 (see lines 36-45). However, we also note that
such results are yet to be standardised and compared, which would allow more reliable comparisons of effect sizes across species and can most effectively be achieved using meta analytic methods (see lines 42-45).

One last point is that there is a probably not a *single* method or experimental protocol that can best investigate animals’ capacity to solve problems cooperatively with others (and coordinate actions). The type of tasks that can be used with primates are maybe not equally suited for birds or dolphins and the other way around. Therefore, the delay task with the string-pulling paradigm should not be seen as the ”silver bullet” to understanding Cooperation/Coordination, but instead as one task that together with others may give us some insight into animals’ capacity to coordinate actions with each other.

What we urgently need are multiple paradigms and measures for each species and/or the exact same experiment across different species. Instead, what we currently have is one unique method implemented in myriad ways.

We agree with reviewer 1’s point that no single task can provide definitive answers in any sub-field of the evolution of cognition. However, we believe that certain tasks do provide a useful unit of comparison in a field that must use comparisons when answering its central questions. This is especially so when potentially comparable variations on a task have been used (see lines 198-201). The comparability of such variations must and can be assessed using systematic review. Thus we argue that our study could demonstrate a useful approach towards gathering evidence for answering central questions in the field of comparative cognition, whilst also acknowledging that future evidence synthesis may need to be conducted along dimensions different to those chosen by us for the present study (see lines 86-106).
3 Reviewer: 2

The revised proposal has addressed many of the previous comments, increased the rigour of the proposal and added in several additional analyses. This will be a complicated project, and there is still work to be done to ensure it would be effective should IPA be given, but I still think it is a useful project.

3.1 Structure

The proposal would be easier to follow if the introduction was split in two, with a first section outlining the original research goals (meta-analysing string pulling data), and then a section outlining the challenges of doing so, rather than having the challenges spread throughout the introduction. This may be a personal preference, but I think it would improve the article’s usability as a resource for meta-analysis in the area, too.

We agree with reviewer 2 that a restructuring of the introduction is necessary following the changes made in the first round of peer review, and we hope that we have clarified the rationale of the study by doing so (see lines 1-156).

3.2 Quality control and assessment

The authors have increased the amount of quality control and assessment stages in the revision.

a) Searches Although the searches are not fully comprehensive (e.g., animal cognition journals such as The Journal of Comparative Psychology, International Journal of Comparative Psychology, Behavioural Processes, Learning and Behavior are not included in the journal search), there is enough redundancy across the three searches to ensure that the vast majority of relevant articles will be identified.

b) Extraction The extraction process, where the articles are extracted twice by the same individual and then another coding 10% seems fine, although if
there are inconsistencies across this 10% more papers will need double coding.

**More papers will be double coded if inter-extracter reliability on the first 10% is low (see lines 292-293).**

c) Quality Checklist The authors have introduced a quality checklist – a modified NICE quality appraisal checklist. I applaud the use of such a checklist, and think this is an interesting project in itself, so should be retained. However, using the checklist as a basis for inclusion/exclusion in the meta-analysis might be a problem, as I’m unsure how the authors will be able to judge some of the questions, and it is likely that most studies would end up being excluded if the checklist was followed strictly.

For example, for question 1b:

Do the selected participants or areas represent the eligible population or area? (e.g. were both sexes represented fairly equally? Were a variety of ages represented fairly equally?)

How will the authors know what the eligible populations are without contacting every study author to see which animals were available? Often in captivity the ages of the animals are similar, so many studies will not represent a variety of ages fairly.

For question 2b:

How well were likely confounding factors identified and controlled?

This would require the authors to pretty much peer-review each paper, unless the authors wanted to decide on a few key confounds for the string pulling task and check these in each paper.

Performing the checklist will be an interesting exercise, possibly to quantify just how far these studies are from the strong epidemiological studies that the quality appraisal form was designed to identify. However, I doubt it can be used effectively as a basis of inclusion/exclusion unless it heavily modified.

The checklist, although still an integral part of the overall study, has been abandoned as a method for the exclusion of studies in the proposed statistical analyses (see lines 280-286).
3.3 The reliability of the phylogenetic tree

AIC will be used to compare competing trees, but this will only compare the fit of the trees relative to each other, and not absolutely, i.e., the tree with the lowest AIC will not necessarily be a good tree, it will just be the best fitting of all the ones compared on the current dataset. I think this needs to be made clearer in the manuscript, particularly given the noise and limited amount of data that will be informing the trees.

We hope that this has now been made clearer in the manuscript (see lines 327-329).

3.4 Meta-analysis/statistical analyses

The meta-analytic model has been updated to be multi-level, which is good as it can attempt to account for phylogenetic relatedness. However, I don’t think the current approach, which includes species and genus only (random = 1—Paper/Species/Genus), will capture most of this phylogenetic information – it would only be useful where many species come from the same genus.

A couple of examples of how to include phylogeny with the rma.mv function can be found in Cinar et al., 2020 (Model 9) and Davies et al., 2020. Both of these provide good descriptions of how they generate the phylogenetic tree (Davies et al. in particular for when distantly related taxa are used), and provide example code in the supporting information. The approach is to include a single random effect of phylogeny, based on a correlation matrix derived from a phylogenetic tree. Both Cinar et al. and Davies et al. use the same APE package to generate these as the current authors are using, so their code should be adaptable.

Meta analysis models have been amended to include matrix structures based on phylogenetic trees following Davies et al. [1] (see lines 344, 382, and RegCode.R file)

The authors have a large number of analyses planned, and they could consider how they will control for multiple testing/whether they think it is an issue.
Multiple testing adjustments have been added into the statistical analysis to control for multiple testing (see lines 421-425, and RegCode.R).

3.5 Data and code

Thank-you for sharing the data and code. I managed to reproduce the pilot analysis fully and the data file was readable. When the full dataset is generated, it would be useful to have a data-dictionary to accompany the file explain precisely what each column is and how it was coded – as I expect this dataset might be of interest to many people.

We thank reviewer 2 for the suggestion of a data dictionary, and will produce one once the full dataset is generated.

Overall, the new proposal is improved, and I still think it is a feasible and worthwhile project. I share Reviewer 1’s concerns about the ability of the project to provide strong answers to the original questions of interest. But even if it fails to answer these questions effectively, then highlighting the barriers around synthesizing evidence like would be a very useful paper. The feasibility of these analyses will depend on the quantity and quality of the overall dataset, which we will only know when it has been extracted. Therefore, it would be useful to conduct a feasibility assessment of the different analyses after the data have been extracted but before conducting them. If IPA is given to the project, I expect a longer Stage 2 review may be necessary than with other registered reports. I still think the project is very interesting and will be of interest those interested in string-pulling tasks but also those interested in evidence synthesis in the field.

We thank reviewer 2 for their interest in the present proposed project. We agree that the feasibility of statistical analyses will depend upon dataset quality, and we have added criteria for conducting certain analyses after dataset generation (see lines 405-409 and 417-420). We acknowledge that phylogenetic analyses may be difficult to
conduct given available information, and will assess the feasibility of these analyses post data collection. Furthermore, we hope that the larger role of quality assessment and data generation in our primary aims for the study will provide a useful resource for researchers in comparative cognition, regardless of the feasibility of these analyses at present. We hope we have represented this aim clearly in the manuscript (see lines 1-156).

3.6 Minor point

1. Code reproducibility Code, line 355 there is a comma after the first value in ref = c("HM015213"). This should be removed to read ref = c("HM015213") I was unable to run the code of the phylogenetic analysis – but I assume this just hasn’t been completed yet.

We offer our apologies. The code for phylogenetic analysis is currently incomplete as information regarding the generated dataset will be necessary before some details can be added. It is currently serving as a template for those later details. This should have been made clear prior to submission and a note has been added to the code file.

References

[1] A. D. Davies, Z. Lewis, L. R. Dougherty, A meta-analysis of factors influencing the strength of mate-choice copying in animals, Behavioral Ecology 31 (6) (2020) 1279–1290.
Appendix C

Once again we thank the editor and reviewers for their helpful comments and, this time, in-principle acceptance of our work.

We have addressed reviewer 2’s comment by removing ‘that researchers generalise findings from very small samples to a much larger population...’ as a ‘major issue’ in the introduction (see lines 2-3), as reviewer 2 rightly points out that many areas of the field of Comparative Cognition do very well with small samples [1].

Finally, ‘met’ has been changed to ‘assessed’ on line 6 so as not to overstate the utility of systematic review and meta analytic methods in meeting the challenges stated in the opening of the introduction.

References

[1] P. L. Smith, D. R. Little, Small is beautiful: In defense of the small-n design, Psychonomic bulletin & review 25 (6) (2018) 2083–2101.
Appendix D

Dear Professor Chris Chambers,

We thank you once again for the careful consideration of our work by both editors at Royal Society Open Science and external reviewers. We hope that we present here a piece of high quality research that has been of higher quality because of the registered report process.

Both the link to the pre-registration and final dataset can be found at the top of page 18 in the stage 2 manuscript. Both are pre-registration and data are held on the Open Science Framework.

No data other than pilot data was collected or analysed prior to the date of in principle acceptance.

We would be grateful for any feedback on the paper, and hope that it meets the rigorous scientific standards of your journal.

Sincerely,

Liam Keeble
Appendix E

The evolution of coordination: A phylogenetic meta-analysis and systematic review

Liam Keeble¹, Joel C. Wallenberg² and Elizabeth E. Price³

¹Henry Wellcome Building, Medical School, Newcastle Upon Tyne, NE2 4HH, United Kingdom
²Percy Building, School of English Literature, Language and Linguistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom
³School of Psychology, Newcastle University

Keywords

Cooperation, Problem-solving, Meta-analysis, Coordination, Cognition

Corresponding author

Liam Keeble: L.Keeble1@newcastle.ac.uk
Abstract

To solve many cooperative problems, humans must have evolved the ability to solve physical problems in their environment by coordinating their actions. There have been many studies conducted across multiple different species regarding coordinating abilities. These studies aim to provide data which will help illuminate the evolutionary origins of cooperative problem solving and coordination. However, it is impossible to make firm conclusions about the evolutionary origins of coordinating abilities without a thorough comparative analysis of the existing data. Furthermore, there may be certain aspects of the literature that make it very difficult to confidently address evolutionary and meta-analytic questions. This study aimed to rectify this by using meta-analysis, phylogenetic analysis, and systematic review to analyse the data already obtained across multiple studies, and to assess the reliability of this data. We found that many studies did not provide the information necessary for meta analysis, or were not comparable enough to other studies to be included in analyses, meaning meta analyses were underpowered or could not be conducted due to low samples of both studies and different species. Overall, we found that many studies reported small positive effects across studies, but the standard errors of these effects frequently traversed zero.
1 Introduction

Two major issues in the field of comparative cognition are: 1) that methods can vary across studies measuring the same variables with different species [1], and 2) that this makes it difficult to conduct thorough, quantitative comparisons between species in many cognitive tasks. The present study will attempt to demonstrate how these problems can be assessed with modern systematic review, and meta-analytic and phylogenetic methods, while also demonstrating potential issues and downfalls of these methods. We will attempt this by pooling results from studies assessing the abilities of different species to coordinate their actions, and assessing these studies in their reliability and rigour.

1.1 The string pull task

The ability to coordinate actions has been essential to the daily activity and survival of human beings throughout their evolutionary history [2, 3, 4, 5]. These abilities have also been discovered to be a part of many an animal’s behavioural repertoire [6].

The cooperative string-pull task has been used frequently across a variety of different species to test what factors influence an animal’s ability to solve complex coordination problems. To be successful in this task, two individuals must pull on two ends of a string simultaneously to retrieve rewards from a platform. If only one of the individuals pulls on the string without an other, the string will come loose and the task will be impossible to solve [7, 8]. Variations on this task have since been designed to test different species of animal in their abilities to coordinate their behaviour. For example, the simultaneous button pushing task used with dolphins [9].

One important iteration of this task is the delay test, where a partner is held back from the apparatus for a period of time. To be successful, participants must wait for their partner to arrive at the apparatus before pulling. The species that have been found to be successful in this iteration are chimpanzees [7], Asian elephants [10] and kea [11] among others. Unsuccessful species tested
using this paradigm include rooks [12], ravens [13], African grey parrots [14] and
domestic dogs [15]. In the simultaneous release iteration of the task (where both
participants are introduced to the apparatus at the same time), tolerance of a
conspecific being in close proximity is found as a good predictor of cooperative
success [7, 16, 13].

However, some researchers are skeptical as to whether tolerance is the best
predictor of success in this task [17]. Some have found that relationship quality
or affiliation between the participants is also a good predictor of success both
within and across species [18, 19, 20, 21, 12, 7], and other social factors likely
do play a fundamental role. But, researchers are yet to assess how these social
factors interact, and also whether there are any other social or non-social factors
also effecting coordination success. Such findings are yet to be synthesised and
compared in a standardised fashion. A meta analytic approach can standardise
measures of success both within and across species, and compare these measures
at the appropriate level and with the appropriate measures [22].

However, social factors may not play as fundamental a role in predicting
success in the cooperative string-pull as many researchers believe. Success could
in fact be dictated by physical problem solving cognition or inhibition alone, and
subjects could be using environmental cues to solve the problem of the task [24].
If data from less social species is missing from the comparative literature, then
it is impossible to assess the extent to which social factors, as opposed to non-
social factors, are responsible for success in the ability to coordinate actions.
It is therefore important to assess bias towards testing more social species in
the studies using this task. This would provide an empirical demonstration
that researchers should be testing a wider range of species from a variety of
different social environments, and a stronger foundation from which to direct
future research.

Recently, a review of this literature has suggested that there is a bias towards
testing social animals using this paradigm [24], and the lack of data concerning
the performance of less social species makes it difficult to ascertain the effects
that living in large groups and having tolerant relationships has on the evolution
of abilities to coordinate actions. This may also lead to (or, alternatively, be the result of) a publication bias towards successful species. Similar biases may also be found in the way different species are tested, the amount of trials in which they are trained on the task, and the number of trials in which they are tested on the task [24]. There is little consensus as to the evolutionary consequences of certain factors (social group size, diet, brain size) on the evolution of the ability to solve coordination problems. This is largely due to there being little quantitative comparison between species. Without such data, it is impossible to draw any conclusions regarding the selective pressures that contribute to the evolution of coordinating abilities.

As such, a systematic review and quantitative meta-analysis are necessary to standardise and rigorously estimate the effects of different methodological approaches across species, to review the methodological rigour and differences across studies, and to attempt to demonstrate how analyses may be conducted if reliable data is present.

1.2 Challenges of analysis

There are several issues in the comparative cognition literature that make drawing reliable conclusions from comparative and/or meta-analytic studies potentially difficult [1]. Differences in the level of analysis (group success or individual success) can mean studies are difficult to compare directly, and can vary the conclusions researchers draw from their results; different sample sizes can make results more or less reliable; and the methods (e.g. number of training and test trials, and apparatus) used by different researchers can make certain subjects, and thus species, more or less likely to be successful.

There is a lot of heterogeneity in the use of the cooperative string-pull task when testing the coordination abilities of different species [24], meaning results are difficult to compare, and studies are rarely of the same quality. Systematic review and, more specifically, quality assessment are useful tools for assessing the quality of studies, where quality is frequently defined as how well a study’s
methods are matched to its goals [25]. Although quality assessment tools are most frequently used to assess whether results from studies testing for the same effect can be synthesised, they also provide a method for assessing whether studies are comparable, especially where the overall goals of a field, and thus the goals of individual studies in that field, are to make comparative assessments possible. These methods are therefore useful for assuring the comparability of results in the field of comparative cognition/psychology. Although comparisons between species in a single task will not provide a definitive answer to the evolution of coordination, tasks do provide a useful unit for comparison. Many questions in the field will eventually need to compare the performance of species in the same (or a very similar) task, using the same (or a very similar) design, and using a standardised measure of success in that task to attain answers, and future meta-analyses can synthesise evidence along different dimensions to bolster bodies of evidence further. But, doing so requires that individual studies ensure high levels of comparability in order to achieve high quality. However, the quality of existing literature has yet to be extensively assessed in these terms.

Due to large heterogeneity between study designs and a lack of comparable data, meta and phylogenetic analysis is, at present, going to be difficult for coordination tasks. Thus, it is essential that the sources of such heterogeneity and the effects this has on conducting meta and phylogenetic analyses need to be assessed. Quality assessment can identify sources of heterogeneity [25], and piloting meta analytic and phylogenetic analysis on available data can identify the effects such heterogeneity can have on analysis. Conducting such analyses rarely leads to completely firm conclusions. In fact, their primary usefulness is frequently derived from their ability to identify gaps and sources of bias in existing literature [26]. Without identifying these gaps and sources of bias, it remains difficult to identify what studies may need to be conducted in the future to make reliable comparative analyses possible, where existing studies may need to be modified and repeated, and what sources of heterogeneity in design need attention when designing those new and repeated studies.

The most productive method for combating these issues is to begin con-
structing a dataset, and piloting possible analyses on existing data, whilst also
assessing the quality of that existing data. This will allow gaps in the literature,
unreliable studies, and plausible analyses to be identified. Thus, the aims of the
present study are as follows:

- To assess the quality of studies using animal coordination tasks.
- To assess the comparability of these same studies.
- To generate a comprehensive dataset of results from these studies.
- To identify any form of bias in these studies.
- And to determine the plausibility of the application of meta analysis and
  phylogenetic analysis to data from these studies.

The present study will assess the quality of, and pool the results from, studies
using coordination tasks. Tests will be conducted to estimate the effects of
publication bias and methodological differences (e.g. level of analysis, sample
size, length of rope, and number of training and test trials) across studies, and
to address some of the following prescient questions regarding differences in
coordination success across species using meta and phylogenetic analysis:

- Does number of trials predict success in a simultaneous-release coordina-
tion task?
- Do social dynamics (affiliation and tolerance) predict success in a simultaneous-
  release coordination task?
- Do brain size, social group size, or dietary factors predict success in a
delayed-release coordination task?

It is unlikely that analyses for the final question will be possible given avail-
able data. However, data will still be collected towards this question in the
hope that this will identify where future studies can be conducted to make such
analyses plausible for the field. Despite limited available data, some data will
allow the precision of such analyses to be assessed given available data.
Such a dataset of studies will be highly beneficial to the field, making it easier for researchers to direct their efforts towards gaps in the literature, and hopefully making it possible to assess evolutionary questions with reliable data in the future [26]. As such, the present study will extract data for some statistical analyses, but will also be used as an opportunity to extract some data that will not be used for analysis, but will help researchers to identify gaps in the literature and design studies that will be comparable to existing results for future comparative analysis.

2 Methods

2.1 Search protocol

All searches will be conducted using Google Scholar. Papers will be extracted from the first fifty pages returned from each search with the following truncated terms:

- Cooperat* string-pull
- Cooperat* problem solv*
- Cooperat* partner choice
- Tolera* and cooperat*
- Coordinat* task
- Tolerat* and coordinat*

Each of these phrases will also be used again, but with the instances of ‘cooperat*’ spelt ‘co-operat*’, and ‘coordinat*’ spelt ‘co-ordinat*’, so that the search protocol accounts for the alternative spellings of the terms. The search using the phrase ‘Cooperat* string-pull’ will be repeated once again but with ‘string-pull’ written as ‘string pull’, in case the term is used without the hyphen in any of the literature.
This search protocol will be repeated in the databases of the journals Animal Behaviour, Animal Cognition, and Animal Behaviour and Cognition in case the searches in Google Scholar do not return all relevant results.

All studies which cite Melis et al. [7], Hirata and Fuwa [8] or Crawford [27] according to Google Scholar will also be included.

After these searches are complete, and following PRISMA guidelines [28], all duplicates of studies will be removed, and then abstracts will be screened for studies that are not relevant to the questions of the present review. These non-relevant studies will also be removed from the dataset. The remaining studies will be read, and those that do not meet the inclusion criteria will also be removed.

Finally, the best method of assessing publication bias is to compare unpublished results with published results [29]. The principal investigators (traditionally the final author) of all studies found using the above search protocol will be contacted in an attempt to gather unpublished results for comparison.

2.2 Criteria for inclusion

All forms of publication that searches return that fit the following criteria will be included in the analysis (including theses and conference abstracts). However, sources that are not published in academic journals will be excluded from most analyses of publication bias.

To be included in the present meta-analysis, studies must use an apparatus that requires two individuals to coordinate their behaviour in order to be successful, and must also assure that individuals cannot receive a reward without coordinating their behaviour (e.g. participate in an individual task for reward). This is a necessary generalisation for the comparative element of the study [30], given that a string-pull apparatus is difficult to replicate with certain species that may survive in water and/or lack the anatomical necessities to operate such an apparatus (e.g. dolphins). The apparatus used in a study must also have a mechanism whereby if one individual attempts the task without another then
the task is failed, or researchers must have deemed one individual attempting
the apparatus without an other as failure.

Studies must also include at least one of the following measures: a measure
of the relationship between tolerance of individuals in a dyad and their success
in a simultaneous-release coordination task; a measure of success in a delay-
release coordination task; a measure of the relationship between relationship
quality/affiliation of individuals in a dyad and success in a simultaneous-release
coordination task; or a correlation measure of number of trials and success in a
simultaneous-release coordination task.

When testing for a relationship between social dynamics and coordination
success, tolerance and affiliation between a dyad can be measured either using
observational or experimental methods, but cooperative success must be tested
using a simultaneous release string-pull task, where both individuals are intro-
duced to the apparatus at the same time. Tolerance is frequently measured
by assessing how willing two individuals are to share food/eat in close prox-
imity [7, 12, 13, 31]. Affiliation is frequently assessed using observational data
(e.g. how frequently two individuals are near each other, or how frequently two
individuals exhibit affiliative behaviours towards each other) [18, 19, 20, 21].

Studies providing measures of the relationship between trials and coordina-
tion success must have tested subjects in a simultaneous-release string-pull task,
and the number of trials used must include both number of test and training
trials received by an individual.

2.3 Quality assessment

Quality assessment of included studies will take place before data extraction. A
quality assessment tool has been created that takes items from the checklist in
appendix G of the Methods for the development of NICE public health guidance
[32]. This checklist forms the basis for assessing the quality of non-randomized
studies that report correlations and associations. It has been modified slightly
to make it more relevant to the field and studies at hand. We hope that com-
parative researchers will use and adapt this tool for their own purposes in future systematic reviews. The questions included are:

1. **Population**
   
   (a) Is the source population well described?
   
   (b) Do the selected participants or areas represent the eligible population or area? (e.g. were both sexes represented fairly equally? Were a variety of ages represented fairly equally?)

2. **Methods**
   
   (a) Was the selection of explanatory variables based on a sound theoretical basis?
   
   (b) How well were likely confounding factors identified and controlled?
   
   (c) Was the apparatus used appropriate for the species tested?*

3. **Outcomes**
   
   (a) Were the outcome measures and procedures reliable? (subjective/objective measures, inter-rater reliability)
   
   (b) Were all the important outcomes assessed?

4. **Analyses**
   
   (a) Was the study sufficiently powered to detect an intervention effect (if one exists)? (Is power assessed/reported?)
   
   (b) Were multiple explanatory variables considered in the analyses?
   
   (c) Were the analytical methods appropriate?
   
   (d) Was the precision of association given or calculable? Is association meaningful? (Were estimates, standard error, p-values, confidence intervals reported?)

5. **Summary**
(a) Are the study results internally valid (i.e. unbiased)?

(b) Are the findings generalisable to the source population (i.e. externally valid)?

Answers to the aforementioned questions can be coded on the following scale, taken from appendix G of the *Methods for the development of NICE public health guidance* [32]:

- ++: ‘...for that particular aspect of study design, the study has been designed or conducted in such a way as to minimise the risk of bias.’

- +: ‘...either the answer to the checklist question is not clear from the way the study is reported, or that the study may not have addressed all potential sources of bias for that particular aspect of study design.’

- -: ‘...for those aspects of the study design in which significant sources of bias may persist.’

- Not Reported (NR): ‘...for those aspects in which the study under review fails to report how they have (or might have) been considered.’

- Not applicable (NA): ‘...for those study design aspects that are not applicable given the study design under review (for example, allocation concealment would not be applicable for case–control studies).’

Items followed by a ‘*’ are items that are not originally from the *Methods for the development of NICE public health guidance* [32], but have been added by the authors of the present study.

Entries in the checklist will also include study ID, a description of the study design, and the name of the quality assessor.

Studies will not be excluded from the meta analysis if they are found to have sources of bias during quality assessment. Doing so would likely make meta analysis impossible due to lack of data. However, leaving these studies in the meta analysis despite them being judged relatively unreliable will also
serve our purpose of illustrating the bias and heterogeneity such studies can introduce into a meta analysis, and so they will be flagged in the visualisations and discussions of results.

2.4 Data extraction

Data extraction will be performed by the lead author. Another author will extract data from 10% of the total studies included in the present study, and inter-extractor agreement will be assessed. Data extraction will also be conducted by the lead author a second time one week after initial extraction. If inter-extractor agreement is low (Cohen’s Kappa <0.8) for the first 10%, 30% of all papers will be double coded [33].

The data extracted will include all the relevant measures that allowed a study to be included in the present analysis, alongside the average number of test and training trials per individual, whether or not the study employed repeated measures, the level of analysis (group/individual), the length of the rope used in the task, the sample size, the delay time used in delay-release coordination tasks, the average number of training trials until success, whether the string-pull used is delay-release or simultaneous-release, a description of the apparatus (e.g. string-pull, button-push), whether subjects in a study have participated in a string-pull task in another study, and the conclusion drawn in a delay-release string-pull (were species successful or not?).

If studies provide a measure of success, or a measure of relationship regarding the above variables, but not as a correlation coefficient (where relationships between variables are concerned), or as an odds ratio (where success in the delay task is concerned), then these measures will be converted using functions from the compute.es package [34]. Provided, that is, that the relevant information for doing so is made available by authors. If the primary investigators of studies who are contacted provide unpublished data where it is possible to attain effects for the above measures from that data, then the appropriate analysis will be carried out to do so.
2.5 Generating the phylogeny

The ape R package will be used to analyse phylogenetic trees \[^{35}\] constructed using the phangorn R package \[^{36}\]. Genetic data for mtDNA of species will be downloaded from GenBank \[^{37}\] using the ‘read.GenBank’ ape function. Muscle software will be used to align genetic sequences \[^{38}\] via the phyloch R package \[^{39}\]. A rooted phylogenetic tree will be constructed (‘upgma’ phangorn function), and two unrooted phylogenetic trees will be constructed using two different neighbour joining (using the ‘NJ’ and ‘BIONJ’ functions respectively) algorithms \[^{40}\] for comparison.

The phangorn R package will also be used to assess the reliability of trees. Maximum likelihood methods will be used to assess tree reliability, and trees will then be compared using AIC \[^{36, 41, 42}\]. The tree with the lowest AIC will be used to generate lambda correlation structures \[^{43, 44}\] to be used in generalised least squares models for testing evolutionary predictors of success in a coordination task (see ‘Statistical analyses’ subsection). It is important to note that this is not the best fitting tree with reality, but the tree that fits the data best of those trees compared. Two different trees will be used in final analyses: one which includes all species in the meta-analysis, and one which includes all primates included in the meta-analysis. Since a variety of species of primate are tested using similar paradigms, the data available likely allows for more reliable quantitative comparisons at the within-family level.

Data for brain size, social group size and dietary breadth will be taken from several sources \[^{45, 46, 47, 48}\].

2.6 Statistical analysis

Multilevel random-effects meta-analysis models will be conducted using the metafor R package \[^{49}\]. The ‘rma.mv’ function will be used to conduct the meta-analysis of results assessing whether the number of trials correlated with coordination success across studies, whether tolerance correlated with coordination success across studies, whether affiliation correlated with coordination success
across studies, and whether there was a similar effect of success in a delay-release
coordination task across studies. Following Dougherty and Guillette [50] and
Davies et al. [22], study, species, and phylogeny will be added to these models
as random factors.

One potential issue will be with handling data from studies where analy-
thesis has been conducted at different levels (individual and group). The above
analysis will be repeated but with data from only studies analysing at the in-
dividual level, and with data from only the studies analysing at the group level
respectively. Where the same individuals have been tested across studies at the
individual level, the average of their effect size will be taken and included in the
dataset as opposed to including two measures for a single individual.

Study heterogeneity ($\tau$, $Q$) can be used to assess if studies share a common
effect size. Usually, random-effects models are able to assess a true effect across
studies, even when we assume the effects of different studies may vary. But,
when we also expect there to be variation in the true effects of different groups,
a subgroup analysis is necessary. In evolutionary meta-analyses, where different
species are tested in one task, it might be expected that there will be a different
true effect size for each species, as we likely expect some species to be more
successful than others in a specific task or paradigm, and this is valuable in-
formation for making evolutionary comparisons. Given that different estimates
may be expected for studies of different species, true effect size for an individ-
ual species will be difficult to estimate by only analysing all studies together.
Therefore, species subgroups will be analysed (when the amount of studies per
subgroup exceeds 2) for effect size wherever possible alongside larger analyses,
and species will be tested as a moderator variable in all meta-analyses where
evolutionary factors are not analysed as moderator variables. This will hopefully
make effect size for different species easier to estimate, and will allow analyses
to be conducted regarding the differing levels of success of species. However,
it is important to note that the number of studies included in these analyses
are likely to be small, and results from analyses should be assessed with this in
mind. Despite this, we hope this analysis will serve as a useful demonstration
of more powerful analyses that could be conducted in the future.

Publication bias will be analysed in multiple different ways, and assessments for publication bias will be conducted for all studies testing the success of subjects in a delay-release coordination task, and for subgroups of studies testing the success of a single species in a delay-release coordination task if sufficient data is available. P-value distributions will be assessed for all studies testing coordination success in a delay-release string-pull task. Differences between unpublished and published results will be assessed using a random effects meta-analysis model with whether the publication was published or not used as a moderator variable, and study, species, and phylogeny included as random factors [29, 50, 22]. Funnel plots and trim and fill methods will be used to assess whether there are any potential studies missing from the dataset (using the ‘funnel’ and ‘trimfill’ metafor functions). The ‘regtest’ function from the metafor package will be used to test for publication biases regarding cooperative understanding with a mixed-effects meta-regression model for plot asymmetry [49]. Orwin’s fail-safe N (calculated using the metafor ‘fsn’ function) will be used to assess the amount of potential studies that would have to exist in the literature, and to have been missed by the present meta-analysis, in order for an effect that does exist to be non-existent in the population [51]. A cumulative meta-analysis (calculated using the metafor ‘cumul’ function) will be conducted using a fixed-effects model ranked by sample size from largest samples to small. This will assess whether studies with smaller samples - which may mean the effect detected is less accurate - are skewing the results of larger sample studies - which should be more precise [29].

For an evolutionary analysis of good predictors of coordination success, several moderator variables will be tested in separate models using the ‘rma.mv’ function from the metafor package [49]. These variables will be endocranial volume, population group size and dietary breadth. Each will be used as an individual predictor of success in a delay-release coordination task. Species and study will be included in models as a random factor, given that some species and studies will appear repeatedly in the dataset. Again, this analysis will
be repeated but with results from studies analysed at only the individual and
group level respectively. This evolutionary analysis will only be conducted if it is
deemed feasible after data collection, where feasible means there are more than
10 different species that can be included in analysis. Where conducted, these
tests will primarily be used to assess the precision (the size of the confidence
interval for estimates) of such models based on available data.

Generalised least squares models for phylogenetic analysis will be constructed
using the ‘gls’ function from the nlme r package [52]. Brownian motion models
of evolution generated from the phylogenetic trees will be fitted to these mod-
els to account for phylogenetic relationships between the species tested. This
will be achieved using the ‘corPagel’ function from the ape r package [35]. The
models will use endocranial volume, various measures of population group size
and various measures of dietary breadth as individual predictors of success in
a delay-release coordination task. Again, if this analysis is not deemed feasible
(as defined above) post data collection, then it will not be conducted; and where
conducted, it will be primarily used to address the precision (as defined above)
of such models with limited data.

P-values will be adjusted using the Holm method [53] where tests are: re-
peated in meta-analytic models and phylogenetic models; conducted to assess
differences between multiple species; conducted on all data and then repeated
with subsets of results from studies analysing at the individual and the group
levels; and where models are repeated to test publication bias.

The ‘forest’ function from the metafor package will be used to produce forest
plots, and the ‘funnel’ function will be used to produce funnel plots [49].

3 Results

The pre-registered protocol was carried out with only one change to the search
strategy as described in the following paragraph. The only changes made to
analysis code was to stop the code for the analyses that could not be conducted
- based on details outlined in the pre-registered methods - from running. The
pre-registration can be found at: https://osf.io/v6qb7. The final data and data analysis script can be found at: https://osf.io/hr6ma/.

463 studies cited Melis et al. [7], 196 studies cited Hirata and Fuwa [8], and 354 studies cited Crawford [27]. The search proposed could not be conducted in the journal Animal Behaviour due to the database not accepting the search format (wildcards). Instead, this search was conducted in the journals Learning and Behaviour, and Behavioural Ecology and Sociobiology. 949 studies were extracted in searches in Behavioural Ecology and Sociobiology. 616 studies were extracted in searches conducted in the Animal Cognition journal database. 193 studies were extracted from the database of the journal Learning and Behaviour.

As shown in figure 1, of the 42 full-text articles assessed for eligibility and removed, 17 were removed because they did not report any of the necessary measures, and 17 were removed because participants could still succeed in a task without coordination. 2 were removed because the results were repeated in other publications, 3 because they employed the wrong iteration of the task for the result reported (i.e. they employed a delay task but an estimate of the relationship between affiliation and success, or a simultaneous release iteration and an estimate of coordination success, or used neither a simultaneous or delay release iteration), and 3 more were removed because they were unaccessible.

Quality assessment finds that while most studies describe the source population well (1a in table 1 - numbers and letters in brackets refer to this same table in the present paragraph), studies are frequently not representative of the population at large (1b). The selection of explanatory variables were usually mostly based on recent theory (2a), and confounding factors tended to be well controlled in studies (2b). Apparatus were generally well adapted to the species tested (2c). Studies were generally split in terms of their tests of rater reliability, with half of the studies reporting these tests and half of the the studies not
Records identified through database search (n=7001)

Additional records identified through journal databases (n=1758)

Records after duplicates removed (n=6649)

Records screened (n=6649)

Records excluded (n=6590)

Full-text articles assessed for eligibility (n=59)

Full-text articles excluded (n=42)

Studies included in qualitative and quantitative synthesis (n=17)

Figure 1: A prisma diagram of the search protocol
reporting them (3a). Studies generally reported all important outcomes, and all important outcomes were generally assessed (3b). Studies rarely reported the power of their study designs (4a), but they generally considered multiple explanatory variables in analyses (4b) and analytical methods were usually appropriate (4c). Studies mostly reported the precision of effect sizes, despite the quality of reports being variable between studies (4d). In summary, quality assessment found that study results were internally valid (5a), but often were not very generalisable to larger populations (5b).

Table 1: A table showing how many articles were coded in each category for each quality assessment question.

| Question | ++ | + | - | NR |
|----------|----|---|---|----|
| Population | 10 | 5 | 2 | 0 |
| 1b | 1 | 4 | 11 | 1 |
| Methods | 0 | 16 | 1 | 0 |
| 2b | 1 | 13 | 2 | 1 |
| 2c | 15 | 2 | 0 | 0 |
| Outcomes | 8 | 1 | 0 | 8 |
| 3b | 7 | 10 | 0 | 0 |
| Analyses | 0 | 0 | 2 | 15 |
| 4b | 0 | 14 | 3 | 0 |
| 4c | 4 | 11 | 2 | 0 |
| 4d | 3 | 6 | 8 | 0 |
| Summary | 1 | 16 | 0 | 0 |
| 5b | 0 | 2 | 15 | 0 |
In some cases, where studies used several training conditions, the number of training trials was recorded as the number of trials in which an individual could interact with the string-pull apparatus prior to the trial used to collect data on the result of interest for the meta-analysis. In some cases, some rope lengths and delay times were made incremental in studies, and were recorded in the present dataset as such. Risk ratios were calculated from comparisons of success rates to an imaginary population of the same size with a success rate of 50% to provide effect sizes for success in the delay task that are appropriate for meta analysis. An external coder (RW) coded 5 papers from the dataset and reliability analyses were conducted. Reliability between coders was mostly high, as illustrated in table 2. However, the sample was frequently not large enough to provide highly reliable estimates of inter-rater similarity.

50% of studies analysed at the dyadic level, and 50% at the group level. At the group level: 94% were repeated measures; 53% of studies employed both simultaneous string pull and delay string pull, and 29% of studies employed just simultaneous, with 18% employing only the delay iteration of the task; 58% of studies claimed their subjects were successful in a task, 23% claimed no success, and the rest were ambiguous or irrelevant; 94% were journal articles, and 94% were published. Of studies that drew a conclusion on the success of their study subjects, 72% claimed their subjects were successful. The dataset included eight studies of birds (7 different species: kea, ravens, rooks, grey parrots, orange-winged parrots, blue-throated macaws and peach-fronted parakeets) and nine studies of mammals (5 different species: chimpanzees, dolphins, dogs, marmosets and elephants).

16 authors were contacted via email in an attempt to ascertain the existence of any unpublished datasets that may be relevant to our study. 14 authors were initially contacted, and two of those authors recommended we contact another researcher, which we did. If authors did not respond to an initial email, a second email was sent a month after the first. Of the 16 authors contacted, 4 responded stating that they did not have any unpublished material.

Comparisons between phylogenetic trees found that the tree with the lowest
Table 2: Table of results from reliability analysis. Percentages refer to the percentage of responses the same between the two coders, and $\kappa$ values are weighted $\kappa$ values.

| Variable                        | Result                               |
|---------------------------------|--------------------------------------|
| Experience estimate             | ICC=1, $p<0.001$                     |
| P values of delay estimates     | ICC=0, $p=0.49$                      |
| Success in delay task estimates | ICC=1, $p<0.001$                     |
| No. of training trials         | ICC=0.69, $p=0.047$                  |
| No. of test trials             | ICC=1, $p<0.001$                     |
| Delay task rope length         | 100%, $\kappa=NA$                    |
| Sample size                    | ICC=0.71, $p=0.038$                  |
| Delay time                     | ICC=0.92, $p=0.0017$                 |
| Success rate                   | ICC=1, $p<0.001$                     |
| Trial length                   | ICC=0.8, $p=0.017$                   |
| Simultaneous and/or delay task | 100%, $\kappa=0.64$                  |
| Previous participation         | 100%, $\kappa=0$                     |
| Conclusion code                | 100%, $\kappa=1$                     |
| Publication type               | 100%, $\kappa=1$                     |
| Publication status             | 100%, $\kappa=0.56$                  |
| Analysis level                 | 80%, $\kappa=0$                      |
| Repeated measures              | 100%, $\kappa=0$                     |

AICc was the tree constructed using neighbour joining tree estimation. The difference between the AICc of this tree and the next best (constructed using the unweighted paired group method) was 25.6.

As shown in figure 2, the relationship over all extracted results between experience with the task and success in the task was 0.4 ($se=0.2$, $z=2$, adjusted $p=0.28$). There was significant heterogeneity ($Q=40.59$, df=8, adjusted $p=0.01$), but a lot of this heterogeneity could be accounted for by including species in the model as a moderator ($Q=39.95$, df=5, adjusted $p=0.01$). When
species was included, heterogeneity in effects was reduced \((Q=0.63, \text{df}=3, \text{adjusted } p=1)\).

When data from different dyads (as opposed to different groups) participating was synthesised, the relationship between experience and success was high \((\text{est.}=0.96, \text{se}=0.09, z=10.47, \text{adjusted } p=0.01)\), and heterogeneity in effects was low \((Q=0.5, \text{df}=2, \text{adjusted } p=1)\). When analysing results of studies that analysed data at the level of the whole group of subjects, the relationship between experience with the task and cooperative success was 0.29 \((\text{s.e.}=0.17, z=1.78, \text{adjusted } p=0.48)\), heterogeneity was estimated at \(Q=9.24 (\text{df}=5, \text{adjusted } p=0.5)\). Species did not have a significant effect as a moderator in this analysis \((Q=9.24, \text{adjusted } p=0.5, \text{df}=5)\). But, when species was included as a moderator variable, heterogeneity between study effect sizes was very low \((Q=0, \text{adjusted } p=1)\).

There was substantial heterogeneity between levels of success in studies using the delay iteration of a coordination task \((\tau=0.2, Q=13.4, \text{df}=6, \text{adjusted } p=0.12)\). The model found an estimate of success of 0.38 \((\text{se}=0.11, z=3.47, \text{adjusted } p=0.009)\), as shown in figure 3. Figure 4 illustrates the change in effect size as the effect sizes of studies using smaller and smaller samples are added to the same meta analysis. A trim-and-fill method applied to this model finds the same estimate of 0.38 \((\text{se}=0.11, z=3.47, \text{adjusted } p=0.009)\). This model also estimated a similar level of heterogeneity when accounting for unpublished results \((\tau=0.2, Q=13.4, \text{df}=6, \text{adjusted } p=0.12)\). However, a regression test for funnel plot asymmetry finds a \(z\) value of -3.55 \((\text{adjust } p=0.009)\), and finds a limit estimate of 1.47 \((\text{ci: 0.88-2.06})\). Similarly, a rank correlation test finds a \(\tau\) value of -0.98 \((\text{adjusted } p=0.009)\), and 0 missing studies. A fail-safe \(N\) analysis estimates that no extra studies, with an average effect size of 0.28 and a target effect size of 0.14, would need to be added to the analysis before the effect size shrunk to a trivial value.

No unpublished data were found when authors of studies were contacted via email. Only one unpublished study was identified during searches, meaning there is only one unpublished study in the analysed dataset. A model assessing
Figure 2: A forest plot of estimates for the relationship between experience with a coordination task and success in that task.

| Species                          | Estimate  | 95% CI     |
|----------------------------------|-----------|------------|
| Nestor notabilis, 2016           | -0.11     | [-0.42, 0.20] |
| Ani glaucogularis, 2020          | 0.22      | [-0.22, 0.66] |
| Psittacus erithacus, 2011         | 0.84      | [0.16, 1.52]  |
| Psittacus erithacus, 2011         | 0.69      | [-0.35, 1.73]  |
| Psittacus erithacus, 2011         | 0.99      | [0.79, 1.19]  |
| Psittacus erithacus, 2011         | 0.85      | [0.33, 1.37]  |
| Corvus corax, 2015                | 0.18      | [-0.60, 0.96]  |
| Pan troglodytes, 2007             | 0.60      | [-0.65, 1.85]  |
| Calithrix jaccus, 2021            | 0.55      | [0.01, 1.11]  |

RE Model: 0.40 [0.01, 0.79]

The difference between published and unpublished studies finds a significant difference between the two (estimate=-8.86, se=2.68, z=-3.31, adjusted $p=0.009$).

There is significant heterogeneity between both the moderator ($Q=10.96$, df=1, adjusted $p=0.009$) and studies in general ($Q=13.4$, df=6, adjusted $p=0.12$).

The other models proposed in the pre-registration could not be conducted due to the samples of studies for those analyses being too low. These analyses were assessments of the relationships between cooperative success and tolerance, and cooperative success and affiliation, and meta analyses of studies analysing at the dyadic level. Furthermore, phylogenetic analyses could not be conducted since the number of different species in analyses never exceeded the required 10 different species. Analyses of individual species could not be conducted due to
Discussion

The present study finds that, whilst studies of animal coordination are largely internally valid, reports of statistical power and inter-rater reliability are few or variable, and many samples used are not representative of animal populations. Unsurprisingly, where statistical analyses could be conducted, it was found that species differences accounted for a lot of variation in results. Only two of our research questions could be assessed: The relationship between number of trials and success in a coordination task; and the presence of publication bias in the
Figure 4: Cumulative estimates of the success of different species in a delay iteration of coordination tasks, added by largest to smallest sample sizes.

The success of most species included in the analysis was correlated with their experience with the coordination task. Statistical tests for publication bias find little evidence of such, despite the one unpublished study that could be included in analyses providing different findings to the published studies. But, these tests may be misleading due to the confirmatory nature of typical assessments of animal cognition.

Only 17 studies could be included in quality assessment due to many studies not meeting the inclusion criteria. Even then, only some of these studies could be included in certain analyses. The maximum number of studies that could be included in any one analysis was 6. This illustrates the disparities in methods, goals and reported outcomes of different studies of animal coordination. Al-
though it may have been possible to widen inclusion criteria, and thus include
more studies in some analyses, this would have limited the specificity of the
questions that could have been asked. Ideally, future studies should use more
standardised study designs when aiming to assess comparative, evolutionary
questions, making specific meta analyses more plausible [54, 46]. Tinbergen’s
four questions can be used to distinguish when studies should be designed similar
to others, or can be more flexible in their design [55]. When studies are asking
evolutionary questions (as opposed to mechanistic, functional, or developmental
questions), these can only ultimately be answered by quantitative comparisons.
Although qualitative judgements can be useful, quantification of species perfor-
mance allows a much clearer measure for comparison. Thus comparability in
study design and reported results is of the utmost importance when aiming to
answer such questions.

Quality assessment finds that, while most studies of animal coordination
retain high validity, it is difficult to ensure that studies are externally valid. Ex-
ternal validity is a known problem in comparative cognition generally [30], and
is likely due to the populations of interest being difficult to access. Species are
frequently represented in the research literature by small groups of the same,
fairly idiosyncratic [50] individuals, and the same is true of the papers assessed
in the present study. Furthermore, studies rarely reported statistical power.
This may be due to low sample sizes frequently being a practical necessity in
the field. However, studies with low sample sizes can still be high powered if
appropriate study designs and statistical procedures are used [56]. However,
if samples remain unrepresentative, this may make assessing the larger evolu-
tionary questions exceedingly difficult. Given that the individual rather than
the group may be the most reliable unit of interest in assessments of cognition
[50, 30], comparisons at the species level may be unreliable due to potential
individual differences in the population that are not present in the sample.
Conducting analyses at both levels will likely produce the most informative and
useful results [30] for assessing all levels of evolutionary questions [55].

Inter rater reliability was high for most coded variables. This is reassuring,
given that recent trials of meta research in comparative cognition have been less
successful when it comes to coding papers in the field [1]. That the number of
training trials had a slightly lower ICC value than most other variables is telling.
This is a difficult variable to code for. Training trials varied a lot between both
individual subjects within studies and also between studies. Furthermore, since
statistical tests of success in a delay task are frequently conducted using various
different methods both within and between papers, this may account for the low
reliability of p values from those statistical tests. However, given the low sample
size in reliability analyses, reliability of these analyses themselves is likely low.

One result found in the present study was a relationship between experi-
ence with a coordination task and success in that task. This result was fairly
consistent across most species included in the analysis, yet there was also hetero-
genecty across species, as expected. This suggests that all species included in the
analysis had mechanisms available to them to decipher successful behavioural
strategies for the task, but that some species were more efficient learners than
others. However, exactly what was being learned/adapted to cannot be con-
cluded. The high level of success and low level of heterogeneity in the analysis
when including only analyses at the dyadic level is likely due to these results
coming from a single species (Psittacus erithacus). This species showed a high
relationship between experience and success. However, in the studies testing
this species, subjects were given high amounts of training, and were trained
until they had completed the task successfully in 90% of trials. The range of
total training trials per subject across studies included in the present study was
1-120, and some could not be specified due to variable numbers of training trials
or unspecified total numbers of training trials. This illustrates an obvious bias
across species arising from differences in study design. It represents a significant
barrier to true quantitative comparisons in the field. Such conclusions can also
be affected by researchers analysing their subjects at different levels (analysing
per dyad vs. analysing over the whole group of subjects). Synthesising results
over all studies (both those analysed at group and dyadic levels) returns a higher
overall effect size than analyses conducted at only the group level. Since the
meta analysis treats such analyses at the dyadic level as individual, low powered
studies, if the effects found at the dyadic level are high then they can pull the
overall effect size in that direction. This again illustrates the potential affect
that individual differences within species can have on potential future phylo-
genetic meta analyses [50, 30]. This is a potential barrier to making reliable
quantitative comparisons between species. When different species are analysed
at different levels, it is difficult to assess whether the differences in effect sizes
is due to individual or species differences. Thus quantitative inter-species com-
parison is difficult to interpret in the present analysis. Although analysing at
the level of the individual or dyad might be useful for assessing individual differ-
ences both within and between species, if authors are attempting to contribute
data towards asking evolutionary questions, then analysing at both levels is the
most useful. Phylogenetic meta analyses will require measures of heterogeneity
both within and between species to draw any reliable conclusions.

Our statistical tests for publication bias (namely Orwin’s fail-safe N, a rank
correlation test, and the trim-and-fill method applied to the meta-analysis of
success measures in the delay task) suggest little publication bias in studies as-
sessing animal coordination. Overall effect size of success in a delay task was
small, and occasionally negative, or standard errors ranged from negative to
positive. However, a cumulative meta analysis does suggest some pull towards
no effect being a result of studies with small samples, and the available unpub-
lished material did report much less successful performances from animals than
the published literature (but only 1 unpublished manuscript was obtained).
Unfortunately, a cumulative meta analysis is limited in this case by the fact
that statistical power in the studies assessed is less affected by sample size and
more reliant upon repeated measures. As such, the results of the meta analy-
sis are conservative. Until unpublished literature can be empirically assessed,
the presence of publication bias remains uncertain. If anything, the results from
statistical tests of publication bias suggest that these tests may be inappropriate
for the results tested in the present study. The field is largely based on per-
forming confirmatory tests, with little room for disconfirmatory analyses [57].
For example, statistical analyses of the behaviour of animals in cognitive tasks frequently frame the question in terms of whether there was a significant difference from chance in their behaviour. However, there is no measure of the extent of failure in the actual task, and an animal would be expected to perform at a chance level if they did not understand the task. This means that an animal can only performance at chance or better than chance, but rarely worse. Thus, it would be unsurprising to find that animals quite frequently demonstrate small positive rates of success in the task over many repeated tests despite not really ‘understanding’ the task. This would be some evidence towards a potential systematic, normally distributed bias towards small, false positive effects. Because many of the effect sizes extracted in the present study were small, statistical tests of publication bias may still find that results are evenly distributed, and thus not skewed (biased) towards a specific result, despite the effects found being small false positives. This is potentially evidence of some systematic bias primarily in study design, rather than at publication.

The above statistical results do clash somewhat with the conclusions drawn by authors. Published authors mostly provided clear conclusions regarding the success or failure of species tested in coordination tasks. This contrasts with recent findings in the field of animal physical cognition, where the conclusions drawn by authors were frequently ambiguous [1]. We also found that the proportion of studies attributing success to species tested in their coordination abilities was similar to that same survey of the conclusions drawn in studies of animal physical cognition [1]. That is, authors in animal physical cognition (when they do draw clear conclusions) are doing so at a rate similar to researchers studying animal coordination. Unfortunately, this uniformity tells us little at present, but may be a fruitful area of research in the future. It could be a signal of systematic bias in study designs, analyses and results, or it could be that the sample of species tested across both domains are simply proportionally good at both coordination and physical tasks, amongst other potential explanations. The latter is an interesting finding in and of itself. However, if the effect is a signal of bias, this should be checked.
But, despite some authors’ claims, the effect sizes extracted from studies of animal coordination did not always correspond to success. Frequently, error values traversed zero, which means evidence may not be extraordinary enough to attribute extraordinary cognitive abilities (such as being able to understand a complex cooperative problem) to the species tested [30]. This may be further evidence of a systematic bias in study designs due to a lack of methods for disconfirming theory and hypotheses. Note that the rate of successful claims from authors is quite high (72%), despite a high proportion of low positive effect sizes. Although, assessments of other subfields of comparative cognition should assess the presence of such phenomena based on whether the subfield does make use of potentially biased study designs. A change in statistical approach could help to eliminate this bias by assessing the proportion of data in support of success vs. the proportion of data against success, as demonstrated in a recent study with marmosets [58]. There is likely substantial between species variation in measures of failure at a task. Until such measures of failure are developed and widely applied, highly reliable quantitative comparisons will be difficult. The present study illustrates this difficulty, with a handful of species performing at a similar low level of success.

Phylogenetic (meta) analyses may provide a further potential tonic to such issues for certain questions in comparative cognition and psychology [59, 46]. Conducting more comparable studies across species and using data from these studies to perform phylogenetic analyses would allow researchers to assess which species perform better than others in a certain task, as opposed to basing conclusions simply on the success or lack of success of a single species. The complex nature of cognition and factors affecting success in such tasks means that asking whether or not a certain species ‘understands’ a task is a much more complicated question than a simple binary outcome can answer. However, information regarding which species perform better in a task than others will perhaps provide a more fruitful avenue for understanding factors involved in task success, both evolutionary and cognitive.

At present, results in studies of animal coordination are just not yet compa-
table enough to make phylogenetic analyses very reliable. In the small field of animal coordination, studies need to be replicated to ensure validity of conclusions for populations of the samples studied, and more species need to be studied more systematically and with more standardised methods in order to make asking evolutionary questions, quantitative comparisons, and phylogenetic analyses possible. This would be the ideal situation, but we recognise that there are extensive practical barriers to such research. Large and varied samples are often difficult to access, especially where species are endangered or difficult to keep in captivity, and studying species in the wild, despite being more ecologically valid, means a loss of experimental control. Comparative cognition as a field at large has faced similar issues in the past [60, 61]. Thankfully, some subfields of comparative cognition have been able to make some reliable quantitative comparisons [59, 46]. Furthermore, initiatives like the Many Primates project are making such analyses more possible for certain animal taxa [54]. However, to make reliable estimates of the effects of selective pressures vs. evolutionary history on certain cognitive traits, data from species from across the evolutionary bush are required [62]. Phylogenetic methods are integral to these efforts. They can be used to synthesise results from both highly controlled experiments and studies with animal groups in the wild. They can also give some added control to studies with wild groups, whilst also accounting for evolutionary relationships when testing species experimentally.

It is important to continually assess the specific challenges and possibilities that each subfield and comparative paradigm presents. We hope we have demonstrated that for animal coordination paradigms in the present study. We also hope that we have demonstrated how methods of systematic review (quality assessment in particular), and meta analysis can be used to assess the extent to which comparative questions can reliably be asked at a given point in time. This will be an ongoing process until it is deemed that there are enough results that are both reliable and similar enough to be synthesised into a phylogenetic analysis. Even then, such quantitative analyses should be continually assessed in this manner to detect (and ultimately mitigate) bias and increase reliability.
This is a vital result as it can be used to direct research in comparative cognition in a reliably productive and specific direction. Thus, at present, comparative and evolutionary questions cannot be asked of animal coordination studies using the data available in the current, sampled literature. Although many of these existing studies can be used to assess other questions of Tinbergen’s [55], the evolutionary questions cannot be assessed quantitatively based on the currently available data. Further, large scale collaborative research can remedy this situation [54, 59].

Data Accessibility

Data and analysis script for this study is available on the Open Science Framework: https://osf.io/hr6ma/.

Declaration of interests

The authors report no conflicts of interest regarding the authorship or publication of this article.

Authors’ contributions

The conceptual and methodological details of the present study were developed by all authors. LK carried out the search strategy, data extraction and statistical analyses. The results were written by LK. All authors contributed to the discussion. An individual external to the study authors double coded a proportion of papers collected for the study.

Acknowledgements

The authors would like to thank Ben Farrar and one other, anonymous reviewer for their very helpful comments on study protocol and the paper itself. We
would also like to thank Robin Watson (RW) for double coding part of our dataset.

**Funding**

This work was supported by the NINE Doctoral Training Partnership.

**References**

[1] B. G. Farrar, D. M. Altschul, J. Fischer, J. van der Mescht, S. Placi, C. A. Troisi, A. Vernouillet, N. S. Clayton, L. Ostojić, Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition, Animal behavior and cognition 7 (3) (2020) 419.

[2] E. Fehr, U. Fischbacher, The nature of human altruism, Nature 425 (6960) (2003) 785.

[3] M. Tomasello, A. P. Melis, C. Tennie, E. Wyman, E. Herrmann, Two Key Steps in the Evolution of Human Cooperation, Current Anthropology 53 (6) (2012) 673–692. doi:10.1086/668207 URL http://www.journals.uchicago.edu/doi/10.1086/668207

[4] R. Boyd, P. J. Richerson, Culture and the evolution of human cooperation, Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1533) (2009) 3281–3288. doi:10.1098/rstb.2009.0134

[5] A. P. Melis, D. Semmann, How is human cooperation different?, Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553) (2010) 2663–2674. doi:10.1098/rstb.2010.0157

[6] T. Clutton-Brock, Cooperation between non-kin in animal societies, Nature 462 (7269) (2009) 51.
[7] A. P. Melis, B. Hare, M. Tomasello, Engineering cooperation in chimpanzees: tolerance constraints on cooperation, Animal Behaviour 72 (2) (2006) 275–286. doi:10.1016/j.anbehav.2005.09.018

[8] S. Hirata, K. Fuwa, Chimpanzees (pan troglodytes) learn to act with other individuals in a cooperative task, Primates 48 (1) (2007) 13–21.

[9] K. Jaakkola, E. Guarino, K. Donegan, S. L. King, Bottlenose dolphins can understand their partner’s role in a cooperative task, Proceedings of the Royal Society B: Biological Sciences 285 (1887) (2018) 20180948. doi:10.1098/RSPB.2018.0948

[10] J. M. Plotnik, R. Lair, W. Suphachoksahakun, F. B. M. de Waal, Elephants know when they need a helping trunk in a cooperative task, Proceedings of the National Academy of Sciences 108 (12) (2011) 5116–5121.

[11] M. Heaney, R. D. Gray, A. H. Taylor, Keas perform similarly to chimpanzees and elephants when solving collaborative tasks, PLoS ONE 12 (2) (2017) 1–13. doi:10.1371/journal.pone.0169799

[12] A. M. Seed, N. S. Clayton, N. J. Emery, Cooperative problem solving in rooks (Corvus frugilegus), Proceedings of the Royal Society B: Biological Sciences 275 (1641) (2008) 1421–1429. doi:10.1098/rspb.2008.0111.

[13] J. J. Massen, C. Ritter, T. Bugnyar, Tolerance and reward equity predict cooperation in ravens (Corvus corax), Scientific Reports 5 (2015) 1–11. doi:10.1038/srep15021.

[14] F. Péron, L. Rat-Fischer, M. Lalot, L. Nagle, D. Bovet, Cooperative problem solving in African grey parrots (Psittacus erithacus), Animal Cognition 14 (4) (2011) 545–553. doi:10.1007/s10071-011-0389-2
[15] L. Ostojić, N. S. Clayton, Behavioural coordination of dogs in a cooperative problem-solving task with a conspecific and a human partner, Animal Cognition 17 (2) (2014) 445–459, Academic journal search engine. doi:10.1007/s10071-013-0676-1.

[16] B. Hare, A. P. Melis, V. Woods, S. Hastings, R. Wrangham, Tolerance Allows Bonobos to Outperform Chimpanzees on a Cooperative Task, Current Biology 17 (7) (2007) 619–623. doi:10.1016/j.cub.2007.02.040.

[17] C. Tassin de Montaigu, K. Durdevic, D. Brucks, A. Krasheninnikova, A. von Bayern, Blue-throated macaws (ara glaucogularis) succeed in a cooperative task without coordinating their actions, Ethology 126 (2) (2020) 267–277.

[18] K. Asakawa-Haas, M. Schiestl, T. Bugnyar, J. J. Massen, Partner choice in raven (corvus corax) cooperation, PLoS ONE 11 (6) (2016) 1–15. doi:10.1371/journal.pone.0156962.

[19] S. Marshall-Pescini, C. Basin, F. Range, A task-experienced partner does not help dogs be as successful as wolves in a cooperative string-pulling task, Scientific reports 8 (1) (2018) 1–12.

[20] M. Suchak, T. M. Eppley, M. W. Campbell, F. B. de Waal, Ape duos and trios: spontaneous cooperation with free partner choice in chimpanzees, PeerJ 2 (2014) e417. doi:10.7717/peerj.417. URL https://peerj.com/articles/417

[21] M. Suchak, J. Watzek, L. F. Quarles, F. B. de Waal, Novice chimpanzees cooperate successfully in the presence of experts, but may have limited understanding of the task, Animal Cognition 21 (1) (2018) 87–98. doi:10.1007/s10071-017-1142-2.

[22] A. D. Davies, Z. Lewis, L. R. Dougherty, A meta-analysis of factors influencing the strength of mate-choice copying in animals, Behavioral Ecology 31 (6) (2020) 1279–1290.
[23] A. M. Seed, K. Jensen, Large-scale cooperation, Nature 472 (7344) (2011) 424–425.

[24] L. Ostojic, The evolution of social cognition: the case of animal cooperative problem solving (2020).

[25] J. C. Valentine, Judging the quality of primary research, The handbook of research synthesis and meta-analysis 2 (2009) 129–146.

[26] W. Wood, A. H. Eagly, Advantages of certainty and uncertainty, The Handbook of Research Synthesis and Meta-Analysis (2009) 455.

[27] M. P. Crawford, The cooperative solving of problems by young chimpanzees, Comp Psychol Monogr 14 (1937) 1–88.

[28] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, et al., Preferred reporting items for systematic reviews and meta-analyses: the prisma statement, PLoS med 6 (7) (2009) e1000097.

[29] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein, Introduction to meta-analysis, John Wiley & Sons, 2011.

[30] B. Farrar, K. Voudouris, N. Clayton, Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research.

[31] N. Borrego, Socially tolerant lions (panthera leo) solve a novel cooperative problem, Animal Cognition 23 (2) (2020) 327–336.

[32] N. I. for Health, C. E. G. Britain), Methods for the development of NICE public health guidance, National Institute for Health and Care Excellence, 2012.

[33] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1) (1960) 37–46.

[34] A. Del Re, M. A. Del Re, Package `compute.es' (2012).
[35] E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r, Bioinformatics 35 (3) (2019) 526–528.

[36] K. P. Schliep, phangorn: phylogenetic analysis in r, Bioinformatics 27 (4) (2011) 592–593.

[37] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, Genbank, Nucleic acids research 41 (D1) (2012) D36–D42.

[38] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, Nucleic acids research 32 (5) (2004) 1792–1797.

[39] C. Heibl, Phyloch: R language tree plotting tools and interfaces to diverse phylogenetic software packages, Available online at: http://www.christophheibl.de/Rpackages.html (2008).

[40] O. Gascuel, Bionj: an improved version of the nj algorithm based on a simple model of sequence data., Molecular biology and evolution 14 (7) (1997) 685–695.

[41] K. P. Burnham, D. R. Anderson, K. P. Huyvaert, Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons, Behavioral ecology and sociobiology 65 (1) (2011) 23–35.

[42] E.-J. Wagenmakers, S. Farrell, Aic model selection using akaike weights, Psychonomic bulletin & review 11 (1) (2004) 192–196.

[43] R. Freckleton, P. Harvey, M. Pagel, Phylogenetic dependence and ecological data: a test and review of evidence, Am. Nat 160 (2002) 716–726.

[44] M. Pagel, Inferring the historical patterns of biological evolution, Nature 401 (6756) (1999) 877–884.

[45] A. N. Iwaniuk, J. E. Nelson, Can endocranial volume be used as an estimate of brain size in birds?, Canadian Journal of Zoology 80 (1) (2002) 16–23.
[46] E. L. MacLean, B. Hare, C. L. Nunn, E. Addessi, F. Amici, R. C. Anderson, F. Aureli, J. M. Baker, A. E. Bania, A. M. Barnard, et al., The evolution of self-control, Proceedings of the National Academy of Sciences 111 (20) (2014) E2140–E2148.

[47] A. R. DeCasien, S. A. Williams, J. P. Higham, Primate brain size is predicted by diet but not sociality, Nature ecology & evolution 1 (5) (2017) 1–7.

[48] M. W. Colbert, R. Racicot, T. Rowe, Anatomy of the cranial endocast of the bottlenose dolphin, tursiops truncatus, based on hrxct, Journal of Mammalian Evolution 12 (1-2) (2005) 195–207.

[49] W. Viechtbauer, Conducting meta-analyses in r with the metafor package, Journal of statistical software 36 (3) (2010) 1–48.

[50] L. R. Dougherty, L. M. Guillette, Linking personality and cognition: a meta-analysis, Philosophical Transactions of the Royal Society B: Biological Sciences 373 (1756) (2018) 20170282.

[51] R. G. Orwin, A fail-safe n for effect size in meta-analysis, Journal of educational statistics 8 (2) (1983) 157–159.

[52] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willigen, R. Maintainer, Package ‘nlme’, Linear and nonlinear mixed effects models, version 3 (1) (2017).

[53] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian journal of statistics (1979) 65–70.

[54] M. Primates, D. M. Altschul, M. J. Beran, M. Bohn, J. Call, S. DeTroy, S. J. Duguid, C. L. Egelkamp, C. Fichtel, J. Fischer, et al., Establishing an infrastructure for collaboration in primate cognition research, PLoS One 14 (10) (2019) e0223675.
[55] P. Bateson, K. N. Laland, Tinbergen’s four questions: an appreciation and an update, Trends in ecology & evolution 28 (12) (2013) 712–718.

[56] P. L. Smith, D. R. Little, Small is beautiful: In defense of the small-n design, Psychonomic bulletin & review 25 (6) (2018) 2083–2101.

[57] B. Farrar, L. Ostojic, The illusion of science in comparative cognition (2019).

[58] J. S. Martin, S. Koski, T. Bugnyar, A. V. Jaeggi, J. J. Massen, Prosociality, social tolerance and partner choice facilitate mutually beneficial cooperation in common marmosets, callithrix jacchus, Animal Behaviour 173 (2021) 115–136.

[59] E. L. MacLean, L. J. Matthews, B. A. Hare, C. L. Nunn, R. C. Anderson, F. Aureli, E. M. Brannon, J. Call, C. M. Drea, N. J. Emery, et al., How does cognition evolve? phylogenetic comparative psychology, Animal cognition 15 (2) (2012) 223–238.

[60] S. J. Shettleworth, The evolution of comparative cognition: is the snark still a boojum?, Behavioural processes 80 (3) (2009) 210–217.

[61] F. A. Beach, The snark was a boojum., American Psychologist 5 (4) (1950) 115.

[62] A. Krasheninnikova, P. K. Y. Chow, A. M. von Bayern, Comparative cognition: Practical shortcomings and some potential ways forward., Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 74 (3) (2020) 160.
Appendix F

Data Dictionary

Liam Keeble, Joel Wallenberg, Bess Price

Emboldened variables are to be double coded and analysed for agreement. If any variables are not presented in paper, code as "NA".

- **Title**: Title of publication.
- **Authors**: Authors of publication.
- **Year**: Year of publication.
- **Removed**: Removed during assessing studies against inclusion criteria (Yes/No). (Categorical variable)
- **Reason**: Reason for removal during assessing studies against inclusion criteria. (Description)
- **Latin**: The latin name of the species in study.
- **Code**: GenBank code identifier for species genetic information. (Categorical variable)
- **ExperienceEst**: Estimate for the relationship between success in a simultaneous release task and number of trials/Attempts at the task (Correlation).
- **ExperienceVar**: Variance for experience estimate for the relationship between success in task and number of trials/Attempts at task.
- **OrigEst**: Description of estimate taken directly from a paper for the relationship between success in a simultaneous release task and number of trials/Attempts at the task from paper before conversion (e.g. "correlation coefficient of 0.7"). (Description)
- **DelayPValues**: P value reported for a statistical test on success in a delay task. (e.g. "p=0.0012") (Numeric)
- **PDescription**: Description of what p-value is representing in study. (e.g. "Difference between groups", or "Performed better than chance"). This is often unclear in studies. In which case, it can be described as such (e.g. "unclear"). (Description)
- **DelayEst**: Estimate of success for delay coordination task (rate of success). (Numerical variable: percentage/fraction)
• DelayVar: Variance for estimate of success for delay coordination task.

• **OrigEst1**: Description of estimate taken directly from paper for success rate in a delay task, where a participant’s partner has been held back from the apparatus (e.g. "50% success rate", or "Dyads were successful 10 out of 20 trials on average"). (Description)

• TolEst: Estimate for the relationship between success in a simultaneous release and the tolerance score of a dyad (Correlation).

• TolVar: Variance for tolerance estimate for the relationship between success in a simultaneous release task and the tolerance score of a dyad.

• **OrigEst2**: Description of estimate taken directly from paper for relationship between dyad tolerance and coordination success (e.g. "Correlation coefficient of 0.5"). (Description)

• AffEst: Estimate for the relationship between success in a simultaneous release and the affiliation score of a dyad. (Correlation)

• AffVar: Variance for tolerance estimate for the relationship between success in a simultaneous release task and the affiliation score of a dyad.

• **OrigEst3**: Description of estimate taken directly from paper for relationship between dyad affiliation and coordination success (e.g. "Beta coefficient = 1.4, s.e.=0.4"). (Description)

• TrainingTrials: Number of training trials per individual/number of trials of experience with string-pull apparatus per individual prior to test phase. If unclear, then just code as "unclear". (Numerical)

• SuccessCriterion1: Description of criterion(s) for success for passing on to testing. (Description)

• TestTrials: Test trials per individual. If unclear, then just code as "unclear". (Numerical)

• SuccessCriterion2: Success criterion for success in coordination task. (Description)

• RepeatedMeasures: Repeated measures or not (Yes/No). (Binary categorical variable)

• AnalysisLevel: Level of analysis - if a paper does both, then add a new row for each dyad and code for each dyad (group/individual). (Binary categorical variable)

• RopeLength: Total length of rope used (metres). If total length is not given, then leave as "NA". (Numerical)
- **SampleSize**: Sample size/number of individual animals tested. (Numerical)

- **DelayTime**: Delay time used in delay task (s). If not specified then leave as "NA". (Numerical)

- **Av.SuccTrialsDelay**: Average number of successful trials in delay task. If not clear from paper, leave as "NA". (Numerical)

- **LengthOfTrials(m)**: Length of test trials in minutes. (Numerical)

- **DelayOrSim**: Does study incorporate delay release, simultaneous release, neither or both simultaneous and delay release iterations of task? (simultaneous/delay/neither/both). (Categorical variable)

- **ApparatusDesc**: Description of apparatus. (Description)

- **ParticipatedBefore**: Have subjects participated in the task before? (Description)

- **AuthorConclusion**: Description of author conclusion from study. (Description)

- **ConclusionCode**: Did authors conclude species were successful or not (successful/not/NA). This variable is only applicable to studies that used a delay task. If authors did not use delay task, the code as "NA". (Categorical variable)

- **PublicationType**: Was publication a journal article, unpublished manuscript, PhD thesis etc. (Categorical variable)

- **PubStatus**: Whether or not the data comes from a study that is published in a journal or not. (Published/Unpublished) (Binary categorical variable)
Reply to Reviewer

Liam Keeble, Joel Wallenberg, Elizabeth E. Price

December 29, 2021

Once again we thank both the editor and reviewer for both their time and their helpful comments. We hope we have responded to them in full. Our responses are indicated by the bold text.

1 Reviewer 2

• The authors have completed the study in-line with the Stage 1 submission, with their noted change to the search after finding Animal Behaviour did not accept wild cards in the search string. This could have been overcome by using a search engine (e.g., Scopus, Web of Science) to perform the searches but the authors might not have had access.

• Overall I believe the authors have performed a useful study and generated a useful dataset. However I believe the inferences the authors make, and general structure of the results and discussion could be much improved to ensure the value of the study is communicated effectively and that the interpretation is justified.

1.1 Overall Structure

• I thought the overall results and discussion section could be made easier to follow, especially because it jumps between methodological challenges, quality control and analyses. Could subheadings be used corresponding to each of the objectives set out in the introduction to make the structure of the results clearer to the reader?

• We have realigned the structure of both our results and discussion sections according to our aims for this study.

1.2 Quality control

• The authors provided a well presented dataset of both the raw data and the quality control, but I was surprised not to the the reasons for the decisions in the quality control document not also included. While the

Appendix G

1
Stage 1 submission didn’t explicitly state that the reasons and relevant text for the decisions to the quality control questions would be recorded, I thought this would have been performed. For example, the authors state that populations were rarely representative, but understanding how this decision was made is important for the quality of the dataset (i.e., have a column in the datafile with the same description from the full text extract and a comment on why the decision has been made). If these data weren’t extracted, I think it would be really valuable for the authors to extract them and add them to the dataset for each question. Commenting on why the decision has been made for internal validity seems particularly important (and what criteria were screened for this), especially as the authors make the claim that “studies of animal coordination are largely internally valid”. This is not necessarily true, rather what the authors did find was that one coder assessing a small number of string pulling tasks did not find reasons to doubt their internal validity.

- **Reasons for the decisions accompanied by extracts from the relevant papers have been included in the quality assessment data frame (qualityAssessment.csv). Where short extracts could not be determined as enough evidence for a reason, readers have been directed to the relevant sections of the paper being assessed. This is usually where we have concluded that a quality assessment criteria is absent from the text, and thus there is no evidence for that reason other than the absence of the criteria in the paper. For example, where studies have not identified confounds, we have noted that there is no evidence that researchers have done so and invite readers to assess the relevant sections for themselves. We have also clarified the limitations of the conclusions that we can draw based on a subjective method such as quality assessment (ll. 589-593)**

1.3 Conclusions

- Another conclusion the authors made surprised me. About publication bias, the authors said their analysis suggested little publication bias in studies assessing animal co-ordination. This surprised me as I know of unpublished studies in this area, and that the statistical power of their analyses to detect publication bias is very low. If their overall sample size is 17 included studies, what number of unpublished similar studies would mean that there is little publication bias? Even 4 or 5 unpublished studies would be a lot here.. If they wanted to make the conclusion about little publication bias then I think power analyses to see the degree of e.g. asymmetry their analyses could detect would be necessary. Alternatively I see no issue in highlighting how difficult it can be to assess publication in very heterogenous datasets with small sample sizes (this is the conclusion I would draw, with a strong prior that some studies will be missing).
• We recognise that our inference came across too strong here. We did not mean to imply that we believed such an inference could be made on such a small scale analysis, but that one could only make such an inference if the analyses were valid (although we recognise that this was unclear in the text). We hope we have now clarified the limitations of our conclusions regarding our statistical tests of publication bias (ll. 621-633).

1.4 Discussion

• The authors discussion focuses on co-ordination tasks and what their data can tell us about string-pulling. As the other reviewer at Stage 1 predicted, this might not be so much. I think a section of the discussion should be devoted to the challenges the authors faced in collecting and synthesising data in this project, as I feel this is one of the major contributions it makes.

• A section has been added describing some difficulties we encountered whilst conducting the study (ll. 747-773).

1.5 Minor Comments

• Reasons for exclusion that are given in the text (Lines 442 - 450) also be presented in the PRISMA diagram

• Reasons have been included in the relevant box in figure 1.

• Could the raw agreement also be included in Table 2, i.e., 5/5 statements extract the same, or 3/5? I think this would be easier for people to understand given the small number of double coded variables.

• Raw agreement scores have been included in table 2.

• Could the quality control statements also be written in Table 1

• Quality control statements have been written in table 1.