THE ROYAL SOCIETY
PUBLISHING

PROCEEDINGS B

Does testosterone impair men's cognitive empathy? Evidence from two large-scale randomized controlled trials

Amos Nadler, Colin F. Camerer, David T. Zava, Triana L. Ortiz, Neil V. Watson, Justin M. Carré and Gideon Nave

Article citation details
Proc. R. Soc. B **286**: 20191062.
http://dx.doi.org/10.1098/rspb.2019.1062

Review timeline

Original submission:	8 May 2019
1st revised submission:	31 July 2019
2nd revised submission:	12 August 2019
Final acceptance:	12 August 2019

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2019-1062.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> No
>
> **Is it clear?**
> N/A
>
> **Is it adequate?**
> N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
The present manuscript reports two large-scale experiments to evaluate the hypothesis that testosterone (partly via organizational effects and partly via activational effects) impairs cognitive empathy in humans, which was prominently proposed by van Honk et al. 2011 in PNAS.

The reported studies have several strengths (and few weaknesses if any) compared to earlier work on this topic. The large sample size provides a lot of statistical power, and accordingly both large-scale studies come to the same conclusion that testosterone has no noteworthy effects on cognitive empathy. Both studies can rule out effect sizes of a magnitude of d=0.2 (or even less) which would be worth talking about, thereby clearly rejecting the earlier findings by van Honk et al. 2011 as a likely false-positive result. The authors also identify a couple of weaknesses of the earlier study which makes the interpretation as a false-positive finding even more plausible. Overall, I think that the current study is extremely valuable in terms of clarifying existing hypotheses and contributing to the scientific progress in the field, so my overall evaluation is very positive.

This said, I should also point out that the manuscript could still be improved notably in terms of presentation of results (effect sizes, direction of effects, possibly a forest plot, descriptive statistics on T-levels, power calculations, and especially Table 1) and maybe also a bit in terms of ensuring objectivity. Regarding the latter, I see a risk that the Authors are subjectively compiling all arguments that speak for their interpretation, not noticing that some of these may be weak arguments. Given the overwhelming empirical evidence presented, I think this pleading is unnecessary and potentially a bit irritating. Below, I give specific hints where I see room for improvement in the order of appearance in the text (next time, please add continuous line numbers to the MS).

(1) Page 3, Abstract, Line 6 (incl. header): consider adding "putative" in front of "biomarker"
(2) P. 4: Maybe you want to explain "cognitive empathy" in more detail for readers (like me) from other fields such as biology. Do I understand correctly that this has little to do with feeling compassion with the other? For instance, someone high in Machiavellian intelligence may be

good at mind reading but feel little compassion? Is that why this is called cognitive empathy (maybe in contrast to emotional empathy)?
(3) P. 5, L. 3: “detect a relationship”: clarify between which variables.
(4) P. 5, L. 17-19: Clarify here that this study found a significant main effect and an interaction, for which you need to describe the direction (e.g. something like “T causes loss of empathy but only in females with the most masculine digit ratio”, or whatever they found).
(5) P. 5, L. 20: insert “females” after “N=33” and “main” after “much smaller”
(6) P. 6, L. 4: Clarify exactly how study [14] differs from [17]. Does the difference just lie in whether the left or the right-hand digit ratio reaches significance (then this difference may be minor because L and R are typically highly correlated and effect sizes may be similar), or is the direction of the effect (low but not high digit ratio, meaning that effect sizes are in opposing direction)?
(7) P. 6, L. 8: “with a 95% confidence” sound like an incorrect interpretation of a significant p-value.
(8) P. 6, L. 7-10: Here you highlight the discrepancies, but isn’t it that always 2 out of 4 studies provided a hint into the same direction (once regarding the main effect, once regarding the interaction)?
(9) P. 6, L. 10: Note that when you cite Table 1 at this place, it will be shown in the Introduction including all your own results. Either do not cite it here, or consider a separate presentation of the literature and your findings (yet having everything in one place seems better to me).
(10) P. 8, L. 10: If there are 4 samples, then all 4 should also be shown in Figure 1 (not just 2 of them). Why is this not even presented in the Supplement (or did I miss it)?
(11) P. 9, L. 10: Explain why these mood measurements were conducted.
(12) P. 9, L. 14: clarify that this is age in years
(13) P. 9, L. 20: “were brought individually”
(14) P. 12, L. 2: Clarify that the 2 measurements were averaged.
(15) P. 12: “Sample size determination”: This sounds weird to say that you determined to require 126 participants but then realize much larger sample sizes. Maybe better to say what power you achieved in each of the 2 experiments. Here you report for the first time a Cohen’s D of d=0.50 for the van Honk study. Note that in Table 1 this is 0.49, but more importantly you need to be consistent with the direction of effects (which can be highly confusing) so you need to speak consistently of an effect size of d=-0.49.
(16) P. 12, L. 8-9: I do not understand what you are presenting here, and I think it is more misleading than helping. The realized power of a study for finding the effect that was found is never really informative. I am not sure what you mean by “veritable effect”. Note that there is no need to over-emphasize that the van Honk study was massively underpowered. I would just remove those two lines.
(17) Figure 1: Please bring the two y-axes to the same scale (one appears ln-transformed, the other log10-transformed). If you use log10 throughout, then it should be easy to label the axis with meaningful numbers (1, 10, 100, 1000, 10000) and indicate units (like pg/ml or whatever). In experiment 1 all 4 samples should be indicated in the top row and results shown below, in experiment 2 the bars should be aligned under the sample.
(18) P. 12-13: Consider moving this comparison to the Discussion section
(19) P. 12, bottom: This difference in findings is not well explained at all. First, there is a reference missing (which study served as justification?) and then it is unclear how that unspecified study differed in its findings from [24].
(20) P. 13, L. 8-9: I am not familiar with these studies, but consider that some of them might be equally shaky as the studies you are criticizing. Just make sure you are treating all studies according to their merit only and not according to whether they help your argument. My point is just to avoid subjectivity whenever there is a risk.
(21) P. 13, L. 12: Hypogonodal males may be not a good comparison to refer to. I guess your data show this clearly anyway.
(22) P. 13, second-last sentence is unclear. Also I am not sure about the arc in Fig. 1

(23) P. 13, last sentence: “Various studies” needs references.
(24) P. 14, L. 9: “higher saliva T”: How many times higher (e.g. median of after treatment / median after placebo)?
(25) P. 14, L. 11-13: This has to be shown in the Supplement and referred to here (I did not check the Supplement)
(26) P. 15: “controlling for baseline performance”: add “as a covariate”. Although this will not affect the results, I am not so fond of analyses that use the first measurement as a covariate. Ordinary least squares regression assumes that this covariate is measured without error and contains no biological noise, so any deviation from the prediction gets misattributed to the second measurement. I think a mixed-effect model with all measurements and time within treatment group as fixed and individual ID as random would be a more appropriate model, but as I said, this does not really matter much.
(27) P. 16, L. 5 and second-last line: I would highlight that these two 95%CIs clearly exclude the previously observed d=-0.49.
(28) P. 16, L. 7: with 97.5% confidence again sounds like inaccurate wording (but people will understand what you mean).
(29) P. 16, L. 7-11: I would remove this sentence. There can be no confidence in such numbers (like 870), because you could easily have found that d=0.001 and then you might claim that trillions of participants would be needed to detect that effect. The general point that very large samples would be required for studying any remaining effect is already sufficiently clear.
(30) P. 16, L. 16: For this 95%CI, is it also possible to say what the corresponding estimate was in the van Honk study? Can you confidently exclude their effect size (I presume yes)?
(31) P. 17, L. 4-13: Please remove (see above).
(32) P. 17, 3rd-last line: Consider “Putative proxies of prenatal T”
(33) P. 18, last sentence of Results: This would better fit to the Discussion. I have the feeling you are over-emphasizing the “inconsistency” here. There are not just type 1 errors but also type 2 errors. Imagine that a real effect exists, but studies are underpowered, then they will sometimes pick this up in the left hand and sometimes in the right hand. Surely, I agree that this is a paradise for fishing (people have even looked at “left minus right”), but still the mentioned inconsistency is relatively minor.
(34) P. 18, L. 15: “negative bound of our confidence intervals”: clarify which one d=-0.15 or d=-0.19
(35) Table 1: This table needs to be improved a lot. First, maybe consider making the table only for aspects of study design and have a forest plot to illustrate effect sizes and 95%CIs. This could lead to a meta-analytic summary with an even narrower 95%CI even when including the initial van Honk study. Would it also be possible to extract effect sizes (from the literature) and directions of slopes for the digit ratio interaction? Within the Table, I would order the studies like this: Honk, Olsson, Bos, Carre, Exp1, Exp2. There is no need to show bars to illustrate sample sizes. The Table needs a detailed legend explaining e.g. “repeated task”, “ES” (note that this abbreviation is not widely known in biology) etc. The “Main effect” column contains effect sizes that deviate from the next column (justify!). For all effect sizes you need to clarify the sign (currently they all seem incorrect) and the magnitude (e.g. 0.22 should be -0.19). For the last column it would be better to have quantitative measures and a direction.
(36) P. 19 end and P. 20 start: I think one could formulate a stronger review of the existing digit ratio literature (but I admit this might be quite some reading effort). What is the main positive evidence that we have and what are the main criticisms? For instance, the observation that Hadza people show no sexual dimorphism is a very weak argument compared to the possibility that the apparent sexual dimorphism may be completely spurious, thereby abolishing the reason for why this trait caught attention and was attributed to sex hormones (see BioRxiv 298786).
(37) P. 21, L. 6: How about mentioning the limitation that females were not studied here?
(38) Please check carefully whether Supplementary Table Legends have the necessary explanations (like I noted for Table 1).

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> Yes
>
> **Is it clear?**
> Yes
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This manuscript reports two large studies of the effects of testosterone treatment on cognitive empathy in men. In addition, it investigates the role of finger ratios, thought be some to reflect prenatal androgen exposure, in any relations observed. The samples sizes and methodologies are impressive. The findings suggest that testosterone treatment does not influence cognitive empathy in men. The authors also found no evidence for a role of 2D:4D in influences on cognitive empathy.
These are important results. They counteract widely-publicized findings from small samples suggesting that testosterone treatment and 2D:4D play a substantial role in cognitive empathy. I have only a few comments that I offer to improve the manuscript.
1. On page 6, the authors say “We conducted a powerful direct test of the activational and developmental effects of T on cognitive empathy” I think this statement would benefit from rephrasing. The authors clearly successfully manipulated T in adulthood and so likely measured

activational effects. As they themselves note in their discussion, the measure of developmental effects, finger ratios, is probably not a reliable measure of early androgen exposure. It might be useful to modify this to say that the current study used similar methodology in larger samples to attempt to replicate prior findings, or something similar.
2. On page 9, the authors say that for experiment 2, “samples of males with low ethnic heterogeneity”. Do they mean that the men were largely Caucasian? If so, that would be easier to understand.
3. On page 20, the authors say the prior results were “not generalizable”. Can they say a bit more here, e.g., “not generalizable across methodologies that increase T concentration” if that is what they mean.

Decision letter (RSPB-2019-1062.R0)

17-Jun-2019

Dear Dr Nadler:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers’ comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Associate Editor have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

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 Associate Editor, 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. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb 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.

When submitting your revision please upload a file under "Response to Referees" - in the "File Upload" section. This should document, point by point, how you have responded to the reviewers’ and Editors’ comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as ‘tracked changes’ to be included in the ‘response to referees’ document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article. Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Victoria Braithwaite

Professor V A Braithwaite
mailto: proceedingsb@royalsociety.org

======

Associate Editor
Comments to Author:

This is a very important contribution that with a large sample size demonstrates that earlier studies demonstrating a link between testosterone and 2D/4D digit ratios play an important role in cognitive empathy were underpowered. With an impressive sample size, a solid experimental setup and state-of-the-art methods the authors demonstrate that there is no link between cognitive empathy and testosterone in humans. The two referees provided very constructive comments to further improve this manuscript.
Sincerely,
Wolfgang Goymann

===

Reviewers' Comments to Author:

Referee: 1

The present manuscript reports two large-scale experiments to evaluate the hypothesis that testosterone (partly via organizational effects and partly via activational effects) impairs cognitive empathy in humans, which was prominently proposed by van Honk et al. 2011 in PNAS.

The reported studies have several strengths (and few weaknesses if any) compared to earlier work on this topic. The large sample size provides a lot of statistical power, and accordingly both large-scale studies come to the same conclusion that testosterone has no noteworthy effects on cognitive empathy. Both studies can rule out effect sizes of a magnitude of d=0.2 (or even less) which would be worth talking about, thereby clearly rejecting the earlier findings by van Honk et al. 2011 as a likely false-positive result. The authors also identify a couple of weaknesses of the earlier study which makes the interpretation as a false-positive finding even more plausible. Overall, I think that the current study is extremely valuable in terms of clarifying existing hypotheses and contributing to the scientific progress in the field, so my overall evaluation is very positive.

This said, I should also point out that the manuscript could still be improved notably in terms of presentation of results (effect sizes, direction of effects, possibly a forest plot, descriptive statistics on T-levels, power calculations, and especially Table 1) and maybe also a bit in terms of ensuring objectivity. Regarding the latter, I see a risk that the Authors are subjectively compiling all arguments that speak for their interpretation, not noticing that some of these may be weak arguments. Given the overwhelming empirical evidence presented, I think this pleading is unnecessary and potentially a bit irritating. Below, I give specific hints where I see room for improvement in the order of appearance in the text (next time, please add continuous line numbers to the MS).

(1) Page 3, Abstract, Line 6 (incl. header): consider adding “putative” in front of “biomarker”
(2) P. 4: Maybe you want to explain “cognitive empathy” in more detail for readers (like me) from other fields such as biology. Do I understand correctly that this has little to do with feeling

compassion with the other? For instance, someone high in Machiavellian intelligence may be good at mind reading but feel little compassion? Is that why this is called cognitive empathy (maybe in contrast to emotional empathy)?
(3) P. 5, L. 3: "detect a relationship": clarify between which variables.
(4) P. 5, L. 17-19: Clarify here that this study found a significant main effect and an interaction, for which you need to describe the direction (e.g. something like "T causes loss of empathy but only in females with the most masculine digit ratio", or whatever they found).
(5) P. 5, L. 20: insert "females" after "N=33" and "main" after "much smaller"
(6) P. 6, L. 4: Clarify exactly how study [14] differs from [17]. Does the difference just lie in whether the left or the right-hand digit ratio reaches significance (then this difference may be minor because L and R are typically highly correlated and effect sizes may be similar), or is the direction of the effect (low but not high digit ratio, meaning that effect sizes are in opposing direction)?
(7) P. 6, L. 8: "with a 95% confidence" sound like an incorrect interpretation of a significant p-value.
(8) P. 6, L. 7-10: Here you highlight the discrepancies, but isn't it that always 2 out of 4 studies provided a hint into the same direction (once regarding the main effect, once regarding the interaction)?
(9) P. 6, L. 10: Note that when you cite Table 1 at this place, it will be shown in the Introduction including all your own results. Either do not cite it here, or consider a separate presentation of the literature and your findings (yet having everything in one place seems better to me).
(10) P. 8, L. 10: If there are 4 samples, then all 4 should also be shown in Figure 1 (not just 2 of them). Why is this not even presented in the Supplement (or did I miss it)?
(11) P. 9, L. 10: Explain why these mood measurements were conducted.
(12) P. 9, L. 14: clarify that this is age in years
(13) P. 9, L. 20: "were brought individually"
(14) P. 12, L. 2: Clarify that the 2 measurements were averaged.
(15) P. 12: "Sample size determination": This sounds weird to say that you determined to require 126 participants but then realize much larger sample sizes. Maybe better to say what power you achieved in each of the 2 experiments. Here you report for the first time a Cohen's D of d=0.50 for the van Honk study. Note that in Table 1 this is 0.49, but more importantly you need to be consistent with the direction of effects (which can be highly confusing) so you need to speak consistently of an effect size of d=-0.49.
(16) P. 12, L. 8-9: I do not understand what you are presenting here, and I think it is more misleading than helping. The realized power of a study for finding the effect that was found is never really informative. I am not sure what you mean by "veritable effect". Note that there is no need to over-emphasize that the van Honk study was massively underpowered. I would just remove those two lines.
(17) Figure 1: Please bring the two y-axes to the same scale (one appears ln-transformed, the other log10-transformed). If you use log10 throughout, then it should be easy to label the axis with meaningful numbers (1, 10, 100, 1000, 10000) and indicate units (like pg/ml or whatever). In experiment 1 all 4 samples should be indicated in the top row and results shown below, in experiment 2 the bars should be aligned under the sample.
(18) P. 12-13: Consider moving this comparison to the Discussion section
(19) P. 12, bottom: This difference in findings is not well explained at all. First, there is a reference missing (which study served as justification?) and then it is unclear how that unspecified study differed in its findings from [24].
(20) P. 13, L. 8-9: I am not familiar with these studies, but consider that some of them might be equally shaky as the studies you are criticizing. Just make sure you are treating all studies according to their merit only and not according to whether they help your argument. My point is just to avoid subjectivity whenever there is a risk.
(21) P. 13, L. 12: Hypogonodal males may be not a good comparison to refer to. I guess your data show this clearly anyway.

(22) P. 13, second-last sentence is unclear. Also I am not sure about the arc in Fig. 1
(23) P. 13, last sentence: “Various studies” needs references.
(24) P. 14, L. 9: “higher saliva T”: How many times higher (e.g. median of after treatment / median after placebo)?
(25) P. 14, L. 11-13: This has to be shown in the Supplement and referred to here (I did not check the Supplement)
(26) P. 15: “controlling for baseline performance”: add “as a covariate”. Although this will not affect the results, I am not so fond of analyses that use the first measurement as a covariate. Ordinary least squares regression assumes that this covariate is measured without error and contains no biological noise, so any deviation from the prediction gets misattributed to the second measurement. I think a mixed-effect model with all measurements and time within treatment group as fixed and individual ID as random would be a more appropriate model, but as I said, this does not really matter much.
(27) P. 16, L. 5 and second-last line: I would highlight that these two 95%CIs clearly exclude the previously observed d=-0.49.
(28) P. 16, L. 7: with 97.5% confidence again sounds like inaccurate wording (but people will understand what you mean).
(29) P. 16, L. 7-11: I would remove this sentence. There can be no confidence in such numbers (like 870), because you could easily have found that d=0.001 and then you might claim that trillions of participants would be needed to detect that effect. The general point that very large samples would be required for studying any remaining effect is already sufficiently clear.
(30) P. 16, L. 16: For this 95%CI, is it also possible to say what the corresponding estimate was in the van Honk study? Can you confidently exclude their effect size (I presume yes)?
(31) P. 17, L. 4-13: Please remove (see above).
(32) P. 17, 3rd-last line: Consider “Putative proxies of prenatal T”
(33) P. 18, last sentence of Results: This would better fit to the Discussion. I have the feeling you are over-emphasizing the “inconsistency” here. There are not just type 1 errors but also type 2 errors. Imagine that a real effect exists, but studies are underpowered, then they will sometimes pick this up in the left hand and sometimes in the right hand. Surely, I agree that this is a paradise for fishing (people have even looked at “left minus right”), but still the mentioned inconsistency is relatively minor.
(34) P. 18, L. 15: “negative bound of our confidence intervals”: clarify which one d=-0.15 or d=-0.19
(35) Table 1: This table needs to be improved a lot. First, maybe consider making the table only for aspects of study design and have a forest plot to illustrate effect sizes and 95%CIs. This could lead to a meta-analytic summary with an even narrower 95%CI even when including the initial van Honk study. Would it also be possible to extract effect sizes (from the literature) and directions of slopes for the digit ratio interaction? Within the Table, I would order the studies like this: Honk, Olsson, Bos, Carre, Exp1, Exp2. There is no need to show bars to illustrate sample sizes. The Table needs a detailed legend explaining e.g. “repeated task”, “ES” (note that this abbreviation is not widely known in biology) etc. The “Main effect” column contains effect sizes that deviate from the next column (justify!). For all effect sizes you need to clarify the sign (currently they all seem incorrect) and the magnitude (e.g. 0.22 should be -0.19). For the last column it would be better to have quantitative measures and a direction.
(36) P. 19 end and P. 20 start: I think one could formulate a stronger review of the existing digit ratio literature (but I admit this might be quite some reading effort). What is the main positive evidence that we have and what are the main criticisms? For instance, the observation that Hadza people show no sexual dimorphism is a very weak argument compared to the possibility that the apparent sexual dimorphism may be completely spurious, thereby abolishing the reason for why this trait caught attention and was attributed to sex hormones (see BioRxiv 298786).
(37) P. 21, L. 6: How about mentioning the limitation that females were not studied here?
(38) Please check carefully whether Supplementary Table Legends have the necessary explanations (like I noted for Table 1).

Referee: 2

This manuscript reports two large studies of the effects of testosterone treatment on cognitive empathy in men. In addition, it investigates the role of finger ratios, thought be some to reflect prenatal androgen exposure, in any relations observed. The samples sizes and methodologies are impressive. The findings suggest that testosterone treatment does not influence cognitive empathy in men. The authors also found no evidence for a role of 2D:4D in influences on cognitive empathy.
These are important results. They counteract widely-publicized findings from small samples suggesting that testosterone treatment and 2D:4D play a substantial role in cognitive empathy. I have only a few comments that I offer to improve the manuscript.
1. On page 6, the authors say "We conducted a powerful direct test of the activational and developmental effects of T on cognitive empathy" I think this statement would benefit from rephrasing. The authors clearly successfully manipulated T in adulthood and so likely measured activational effects. As they themselves note in their discussion, the measure of developmental effects, finger ratios, is probably not a reliable measure of early androgen exposure. It might be useful to modify this to say that the current study used similar methodology in larger samples to attempt to replicate prior findings, or something similar.
2. On page 9, the authors say that for experiment 2, "samples of males with low ethnic heterogeneity". Do they mean that the men were largely Caucasian? If so, that would be easier to understand.
3. On page 20, the authors say the prior results were "not generalizable". Can they say a bit more here, e.g., "not generalizable across methodologies that increase T concentration" if that is what they mean.

Author's Response to Decision Letter for (RSPB-2019-1062.R0)

See Appendix A.

Decision letter (RSPB-2019-1062.R1)

05-Aug-2019

Dear Dr Nadler

I am pleased to inform you that your manuscript RSPB-2019-1062.R1 entitled "Does testosterone impair men's cognitive empathy? Evidence from two large-scale randomized controlled trials" has been accepted for publication in Proceedings B. However, the Associate Editor requests some minor revisions to your manuscript. Therefore, I invite you to respond to these comments and revise your manuscript. Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days. If you do not think you will be able to meet this date please let us know.

To revise your manuscript, log into https://mc.manuscriptcentral.com/prsb 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. 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 Editor and upload a file "Response to Editor". You can use this to document any changes you make to the original manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Before uploading your revised files please make sure that you have:

1) A text file of the manuscript (doc, txt, rtf or tex), including the references, tables (including captions) and figure captions. Please remove any tracked changes from the text before submission. PDF files are not an accepted format for the "Main Document".

2) A separate electronic file of each figure (tiff, EPS or print-quality PDF preferred). The format should be produced directly from original creation package, or original software format. PowerPoint files are not accepted.

3) Electronic supplementary material: this should be contained in a separate file and where possible, all ESM should be combined into a single file. All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

4) A media summary: a short non-technical summary (up to 100 words) of the key findings/importance of your manuscript.

5) Data accessibility section and data citation
It is a condition of publication that data supporting your paper are made available either in the electronic supplementary material or through an appropriate repository.

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should be fully cited. To ensure archived data are available to readers, authors should include a 'data accessibility' section immediately after the acknowledgements section. This should list the database and accession number for all data from the article that has been made publicly available, for instance:

- DNA sequences: Genbank accessions F234391-F234402
- Phylogenetic data: TreeBASE accession number S9123
- Final DNA sequence assembly uploaded as online supplemental material
- Climate data and MaxEnt input files: Dryad doi:10.5521/dryad.12311

NB. From April 1 2013, peer reviewed articles based on research funded wholly or partly by RCUK must include, if applicable, a statement on how the underlying research materials – such as data, samples or models – can be accessed. This statement should be included in the data accessibility section.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link http://datadryad.org/submit?journalID=RSPB&amp;manu=(Document not available) which will take you to your unique entry in the Dryad repository. If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link. Please see https://royalsociety.org/journals/ethics-policies/data-sharing-mining/ for more details.

6) For more information on our Licence to Publish, Open Access, Cover images and Media summaries, please visit https://royalsociety.org/journals/authors/author-guidelines/.

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

Sincerely,
Victoria Braithwaite

Professor V A Braithwaite
Editor, Proceedings B
mailto:proceedingsb@royalsociety.org

Associate Editor:

Comments to Author:

Thank you very much for thoroughly revising your manuscript according to the suggestions of the referees. Upon reading the final version I noticed a few further issues that need to be addressed:
1) Since you already have a large number of acronyms in your manuscript consider to spell out "testosterone" instead of using "T"
2) I noticed that the sencence in line 12 on page 10 does not make sense, please change.
3) Also, if you compare your results section with the entries in Table 1, then the numbers for Cohen's D and the 95% confidence intervals (signs) do not match.
4) The font type and size changes in the main text and in the supplement. For the main text this is not so much of an issue since it will be reformatted anyway, but the supplement will be printed as is. You may consider changing this.
Regards
Wolfgang Goymann

Decision letter (RSPB-2019-1062.R2)

12-Aug-2019

Dear Dr Nadler

I am pleased to inform you that your manuscript entitled "Does testosterone impair men’s cognitive empathy? Evidence from two large-scale randomized controlled trials" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Paper charges
An e-mail request for payment of any related charges will be sent out shortly. The preferred payment method is by credit card; however, other payment options are available.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,

Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Appendix A

Dear Professor Braithwaite,

It is with excitement that we resubmit to you a revised version of manuscript RSPB-2019-1062, “Does testosterone impair men’s cognitive empathy? Evidence from two large-scale randomized controlled trials”, for the *Proceedings of the Royal Society B: Biological Sciences*.

Thank you for giving us the opportunity to revise and resubmit this manuscript. In keeping with your communication to us, we are resubmitting this revision following the request for extension.

We appreciate the time and detail provided by each reviewer and by you and have incorporated the suggested changes into the manuscript; the paper has certainly benefited from these insightful revision suggestions.

We believe this revision satisfies outstanding comments and prepares this manuscript to publication in the *Proceedings of the Royal Society B: Biological Sciences*.

Thank you,

Amos Nadler, Colin F. Camerer, David T. Zava, Triana L. Ortiz, Neil V. Watson, Justin M. Carré, & Gideon Nave

Reviewers' Comments to Author:

Referee: 1

The present manuscript reports two large-scale experiments to evaluate the hypothesis that testosterone (partly via organizational effects and partly via activational effects) impairs cognitive empathy in humans, which was prominently proposed by van Honk et al. 2011 in PNAS.

The reported studies have several strengths (and few weaknesses if any) compared to earlier work on this topic. The large sample size provides a lot of statistical power, and accordingly both large-scale studies come to the same conclusion that testosterone has no noteworthy effects on cognitive empathy. Both studies can rule out effect sizes of a magnitude of d=0.2 (or even less) which would be worth talking about, thereby clearly rejecting the earlier findings by van Honk et al. 2011 as a likely false-positive result. The authors also identify a couple of weaknesses of the earlier study which makes the interpretation as a false-positive finding even more plausible. Overall, I think that the current study is extremely valuable in terms of clarifying existing hypotheses and contributing to the scientific progress in the field, so my overall evaluation is very positive.

This said, I should also point out that the manuscript could still be improved notably in terms of presentation of results (effect sizes, direction of effects, possibly a forest plot, descriptive statistics on T-levels, power calculations, and especially Table 1) and maybe also a bit in terms of ensuring objectivity.

Regarding the latter, I see a risk that the Authors are subjectively compiling all arguments that speak for their interpretation, not noticing that some of these may be weak arguments. Given the overwhelming empirical evidence presented, I think this pleading is unnecessary and potentially a bit irritating.

Below, I give specific hints where I see room for improvement in the order of appearance in the text (next time, please add continuous line numbers to the MS).

(1) Page 3, Abstract, Line 6 (incl. header): consider adding “putative” in front of “biomarker”
Thank you, we added “putative” as suggested.

(2) P. 4: Maybe you want to explain “cognitive empathy” in more detail for readers (like me) from other fields such as biology. Do I understand correctly that this has little to do with feeling compassion with the other? For instance, someone high in Machiavellian intelligence may be good at mind reading but feel little compassion? Is that why this is called cognitive empathy (maybe in contrast to emotional empathy)?

We appreciate this comment, as it is difficult to judge familiarity with a topic with which one becomes intimately involved. We elaborate on cognitive empathy and how it differs from empathy for a broad readership in a footnote in the Introduction.

(3) P. 5, L. 3: “detect a relationship”: clarify between which variables.
Thank you, added **(bold)**: “...several others failed to detect a relationship **between digit ratio and cognitive empathy**...”

(4) P. 5, L. 17-19: Clarify here that this study found a significant main effect and an interaction, for which you need to describe the direction (e.g. something like “T causes loss of empathy but only in females with the most masculine digit ratio”, or whatever they found).
Point taken. We seek to be clearer about the direction of effect while specifying the hand in that section (**bold** added): “**In addition to reporting a main effect of exogenous T reducing cognitive empathy**, more than 50% of the individual differences in the effect on the RMET were explained by the participants’ variation in the right-hand 2D:4D, implying involvement of prenatal T exposure in the causal effect [17].”

(5) P. 5, L. 20: insert “females” after “N=33” and “main” after “much smaller”
Thank you, added.

(6) P. 6, L. 4: Clarify exactly how study [14] differs from [17]. Does the difference just lie in whether the left or the right-hand digit ratio reaches significance (then this difference may be minor because L and R are typically highly correlated and effect sizes may be similar), or is the direction of the effect (low but not high digit ratio, meaning that effect sizes are in opposing direction)?
We elaborate on each study following L. 4 and concatenate them all in Table 1.

(7) P. 6, L. 8: “with a 95% confidence” sound like an incorrect interpretation of a significant p-value.
Thank you, this sentence has been re-written for clarity.

(8) P. 6, L. 7-10: Here you highlight the discrepancies, but isn’t it that always 2 out of 4 studies provided a hint into the same direction (once regarding the main effect, once regarding the interaction)?
Thank you for this suggestion, we now elaborate on the directional trend as well as associated challenges in the section **Do the data support the hypothesis?.**

(9) P. 6, L. 10: Note that when you cite Table 1 at this place, it will be shown in the Introduction including all your own results. Either do not cite it here, or consider a separate presentation of the literature and your findings (yet having everything in one place seems better to me).

This is a good point which was not apparent to us from a reader's perspective. We have omitted references to our results and to Table 1 in the Introduction and moved the table's location to the Discussion section for the reason you point out.

(10) P. 8, L. 10: If there are 4 samples, then all 4 should also be shown in Figure 1 (not just 2 of them). Why is this not even presented in the Supplement (or did I miss it)?
Additional samples were taken to obtain high-resolution measures for tasks occurring over the course of the entire day-long study. We now added them to the Supplementary Material: "Two additional samples were taken later during this day-long experiment to provide measures for other tasks: sample C T levels in the T group were 9.23, SD = 1.45 and the placebo group mean was 5.28, *SD* = 0.962; two-sided t-test: *P* < 0.0001, *t*(240) = 24.8. The fourth samples were similar, with the T group mean T levels of 9.16, *SD* = 0.13, placebo mean T levels of 5.19, *SD* = 0.92; two-sided t-test: *P* < 0.0001, *t*(240) = 25.8."

(11) P. 9, L. 10: Explain why these mood measurements were conducted.
Thank you for prompting us to elaborate on our intentions in the Methods section of manuscript: "Because there are various feasible channels through which T could affect RMET performance (and affect being one of them), We measured mood using the PANAS-X scale [22], both pre- and post-treatment (see Table S1a in Supplementary Material for aggregated responses)."

(12) P. 9, L. 14: clarify that this is age in years
Agreed: we replaced M with "mean age" for both studies' **Participants and experimental procedure** section.

(13) P. 9, L. 20: "were brought individually"
Added "brought", thank you

(14) P. 12, L. 2: Clarify that the 2 measurements were averaged.
Good call, now clarified for both experiments: "inter-rater correlation was 0.96 **and their scores were averaged**."

(15) P. 12: "Sample size determination": This sounds weird to say that you determined to require 126 participants but then realize much larger sample sizes. Maybe better to say what power you achieved in each of the 2 experiments.
We concur. This was recontextualized and moved to the Discussion to specify the degree to which said study was underpowered.

Here you report for the first time a Cohen's D of d=0.50 for the van Honk study. Note that in Table 1 this is 0.49, but more importantly you need to be consistent with the direction of effects (which can be highly confusing) so you need to speak consistently of an effect size of d=-0.49.
Thank you for pointing this out, we now use the rounding rules consistently in the paper and correct the signs to reflect the correct direction of reported effect.

(16) P. 12, L. 8-9: I do not understand what you are presenting here, and I think it is more misleading than helping. The realized power of a study for finding the effect that was found is never really informative. I am not sure what you mean by "veritable effect". Note that there is no need to over-emphasize that the van Honk study was massively underpowered. I would just remove those two lines.
We can see that those two lines contribute little and thus removed them

(17) Figure 1: Please bring the two y-axes to the same scale (one appears ln-transformed, the other log10-transformed). If you use log10 throughout, then it should be easy to label the axis with meaningful numbers (1, 10, 100, 1000, 10000) and indicate units (like pg/ml or whatever).
Thank you for pointing this out. Y-axes have been standardized for both datasets for this two-part figure

In experiment 1 all 4 samples should be indicated in the top row and results shown below, in experiment 2 the bars should be aligned under the sample.
The two additional measures occurred long after the task was completed (by design) yet have no functional connection to the results or analysis in this paper so were excluded from the manuscript. However, we added them to the Supplementary Material in the "**Hormonal Changes Following Treatment and Manipulation Check"** section**:**

"Two additional samples were taken later during this day-long experiment to provide measures for other tasks: sample C T levels in the T group were 9.23, SD = 1.45 and the placebo group mean was 5.28, SD = 0.962; two-sided t-test: P < 0.0001, t(240) = 24.8. The fourth samples were similar, with the T group mean T levels of 9.16, SD = 0.13, placebo mean T levels of 5.19, SD = 0.92; two-sided t-test: P < 0.0001, t(240) = 25.8."

(18) P. 12-13: Consider moving this comparison to the Discussion section
We appreciate this suggestion. We considered moving this section to Discussion yet respectfully decided to place it in the Methods section for the following reasons:

1. Provide a transparent comparison between the studies as we expect readers to naturally compare/contrast the studies. Having this comparison in the Methods is intended to allow readers to adjudicate the Results section more readily.
2. Elaborate on the experimental design and provide rationale for various factors therein
3. The Discussion section does not have subsections, and we feared that introducing a subsection will have a negative effect on the paper's reading flow.

(19) P. 12, bottom: This difference in findings is not well explained at all. First, there is a reference missing (which study served as justification?) and then it is unclear how that unspecified study differed in its findings from [24].
Thank you for pointing out this is unclear, we now refer to the study (cited earlier in the same paragraph) by author name: "Moreover, the **Tuiten et al.** study that served as a justification for using a 4 hour delay had only 8 participants, and reported a statistically weak treatment effect…"

(20) P. 13, L. 8-9: I am not familiar with these studies, but consider that some of them might be equally shaky as the studies you are criticizing. Just make sure you are treating all studies according to their merit only and not according to whether they help your argument. My point is just to avoid subjectivity whenever there is a risk.

We thank you for highlighting this and recognize the value of remaining objective and being transparent about our experimental design choices. We now elaborate on rationale that informed our decision to use transdermal T that we previously left out for succinctness:

"In Experiment 1 we chose to administer T using an FDA-approved transdermal gel for three reasons. First, transdermal gel had been extensively studied in the medical literature both prior and following its approval [25,26]. Second, one of our laboratories found reliable treatment effects in serum [27], and third, the pharmacokinetics of a single-dose of this T administration method were mapped prior to the inception of our experiments by a study suggesting that plasma T levels peaked 3 hours after single-dose exogenous transdermal administration, and stabilized at high levels between 4 and 7 hours following administration [21].[1] Therefore, we had all participants return to the lab 4.5 hours after receiving gel, when androgen levels were elevated and stable. We used a 100 mg transdermal dose, which quickly elevates then holds T levels high and stable for approximately 24 hours [25] and was shown to generate effects on cognition, decision making, and other behaviours [27,29–31].

…

The doses in both experiments are commonly prescribed daily to men with low circulating T levels and serve as two distinct physical transport channels (transdermal and intranasal, respectively) to reduce the probability that behavioural effects are transport channel specific. Various studies show significant heterogeneity in change in T levels depending on delivery method, location of application in the body, and biofluid measured [14,21,24,27,29,32]. However, all the exogenous delivery methods in this particular literature cause a common hormonal trajectory characterized by a rapid initial rise, a peak above typical circulating levels, and eventual return to baseline."

[1] Subsequent studies measuring T in serum in significantly larger sample sizes demonstrate an earlier hormonal peak at 60 minutes post administration with subsequent stabilization [14,25].

(21) P. 13, L. 12: Hypogonodal males may be not a good comparison to refer to. I guess your data show this clearly anyway.

True, there are a multitude of differences between eugonadal and hypogonadal men (most of which likely stem from the etiology of the condition itself plus the hormonal deficit). This is a new, recently approved T delivery system, however, it was chosen to test whether delivery method *per se* matters. As shown, we find that it does not affect RMET and can rule out that the lack of effect is due to a single administration method/experimental choice. Recent literature using this intranasal delivery method found effects in cooperation (Bird et al. 2019) and responses to threatening others (Geniole et al. 2019), which suggests that testosterone has domain-specific, rather than catholic effects.

(22) P. 13, second-last sentence is unclear. Also I am not sure about the arc in Fig. 1

Thank you, we agree. The sentence has been removed and the paragraph has been re-worked to better communicate the intended messages.

(23) P. 13, last sentence: “Various studies” needs references.

Thank you, references have been added.

(24) P. 14, L. 9: “higher saliva T”: How many times higher (e.g. median of after treatment / median after placebo)?

Full details are provided in the unabridged supplemental material, yet we now also provide the logged T measures both pre- and post-treatment in that part of the manuscript (Results > Manipulation check) and provide the following footnote: “raw median T levels of the T group were 33.5 times that of the placebo group post-treatment”. The spread of T on surfaces, door knobs, and pens even after being cleaned increased the range of sample values without corresponding physiological levels (discussed in **Hormonal Changes Following Treatment and Manipulation Check** in Supplementary Material, which includes the measures we took to address this problem mid-study). After the experiment we learned that this issue has occurred in other experiments using topical hormonal administration (Du et al. 2013), and have implemented strict controls to avoid this in future experiments we run as well as providing transparent documentation for others.

(25) P. 14, L. 11-13: This has to be shown in the Supplement and referred to here (I did not check the Supplement)

Thank you for pointing this out. We now note in the manuscript for easy reference/retrieval that mood measurements pre- and post-treatment in Table S1a for study one and similarly for study two in Table S1b. Also, we provide an exhaustive table of pre- and post-treatment hormone levels across treatments in Fig. S2a (along with differences in mean t-test reporting) for study one and similar data in Table S2b for study two.

(26) P. 15: “controlling for baseline performance”: add “as a covariate”. Although this will not affect the results, I am not so fond of analyses that use the first measurement as a covariate. Ordinary least squares regression assumes that this covariate is measured without error and contains no biological noise, so any deviation from the prediction gets misattributed to the second measurement. I think a mixed-effect model with all measurements and time within treatment group as fixed and individual ID as random would be a more appropriate model, but as I said, this does not really matter much.

We agree that baseline RMET measures are not “pure” baselines due to noise. Our analytical choice was motivated by a recent blog by Uri Simonsohn (http://datacolada.org/39), which suggested to a regression with control for baseline in order to account for baseline performance measured with low test-retest reliability, while keeping statistical power high. We agree that a mixed-effect model would also be appropriate for the analysis here and would yield similar results. Respectfully, we decided to keep the original analysis, yet if the reviewer insists we will be happy to provide a mixed-model.

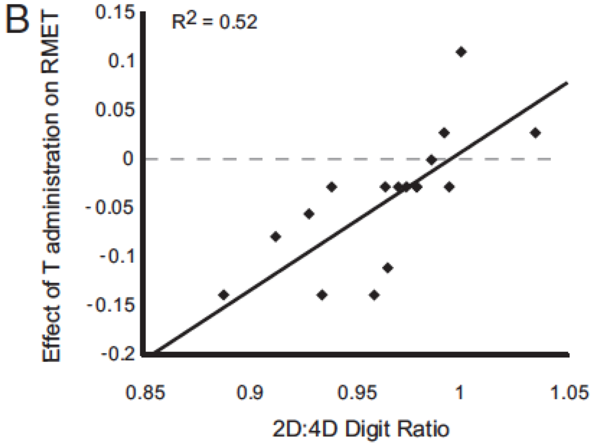
(27) P. 16, L. 5 and second-last line: I would highlight that these two 95%CIs clearly exclude the previously observed d=-0.49.

Thank you for this suggestion. We changed the second sentence in that paragraph to state, “The effect’s point estimate was positive and excludes the *d*=-0.49 reported in (van Honk et al. 2011).”

(28) P. 16, L. 7: with 97.5% confidence again sounds like inaccurate wording (but people will understand what you mean).
We see your point and changed that sentence to the following for experiment 2: *“... point estimate of the effect in Experiment 2 was positive, and the 95% CI did not include negative effects of T administration on the RMET that were greater in magnitude than 0.15.”*
Also, we provide include the following in relation to findings from experiment 1: *“Thus, the effect’s point estimate was positive and the 95% CI excluded the d=-0.49 reported in (van Honk et al. 2011) or any negative effects that are greater in magnitude than d=0.19.”*

(29) P. 16, L. 7-11: I would remove this sentence. There can be no confidence in such numbers (like 870), because you could easily have found that d=0.001 and then you might claim that trillions of participants would be needed to detect that effect. The general point that very large samples would be required for studying any remaining effect is already sufficiently clear.
We appreciate the suggestion and agree with the logic put forth; we have removed this sentence as it appears that the evidence speaks for itself, *res ipsa loquitur.*

(30) P. 16, L. 16: For this 95%CI, is it also possible to say what the corresponding estimate was in the van Honk study? Can you confidently exclude their effect size (I presume yes)?
Unfortunately, the van Honk et al. paper did not provide the 95% CI of the effect. The paper states: “However, Spearman correlations showed that the relation between 2D:4D ratio and the impairment in cognitive empathy induced by testosterone administration was highly significant [**ρ**(14) = 0.85; P < 0.0001].” And this associated graphic was provided:



(31) P. 17, L. 4-13: Please remove (see above).
Agreed, removed.

(32) P. 17, 3rd-last line: Consider "Putative proxies of prenatal T"

We previously used the word "putative" in that context, yet backed off to reduce seeming as though we were attacking the measure at every opportunity. We have added it back for accuracy.

(33) P. 18, last sentence of Results: This would better fit to the Discussion. I have the feeling you are over-emphasizing the "inconsistency" here. There are not just type 1 errors but also type 2 errors. Imagine that a real effect exists, but studies are underpowered, then they will sometimes pick this up in the left hand and sometimes in the right hand. Surely, I agree that this is a paradise for fishing (people have even looked at "left minus right"), but still the mentioned inconsistency is relatively minor.

Fair point, now deleted. We already wrote about this in the Discussion ("A third reason concerns the validity of the 2D:4D biomarker…") so it was somewhat redundant.

(34) P. 18, L. 15: "negative bound of our confidence intervals": clarify which one d=-0.15 or d=-0.19

This was regarding Experiment 1's upper limit of beta of d=-0.19; now clarified in the manuscript.

(35) Table 1: This table needs to be improved a lot.

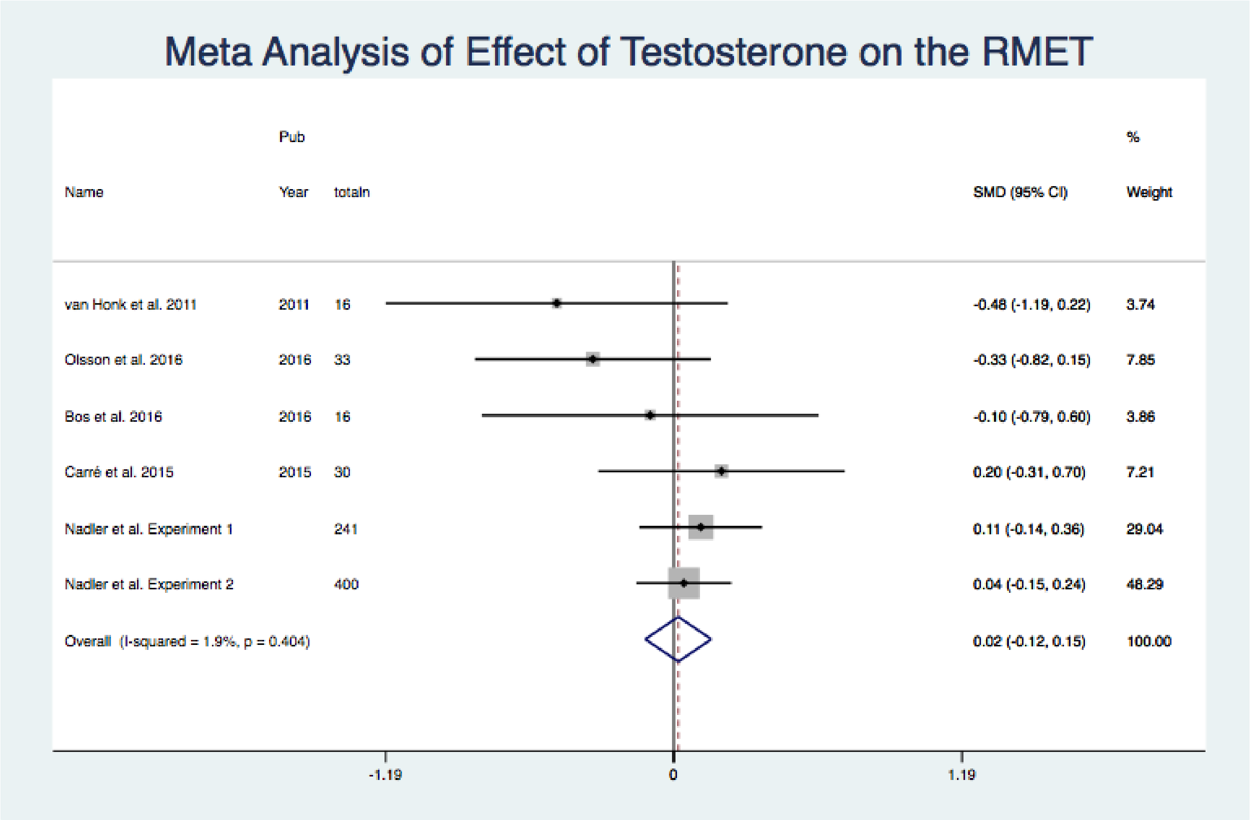
A. First, maybe consider making the table only for aspects of study design and have a forest plot to illustrate effect sizes and 95%CIs. This could lead to a meta-analytic summary with an even narrower 95%CI even when including the initial van Honk study.

We appreciate the recommendation to modify Table 1, which has helped to further improve the paper. We also agree, in principle, that a forest plot could serve a useful purpose. However, creating a meta analytic graph using the summary statistics that are currently available to us is problematic for two reasons.

1. Unfortunately, the statistical reporting of results in some of the previously published papers on the topic were not comprehensive, and we do not have the raw data of these studies. We therefore lack some of the information necessary for computing standard errors, change score, or any other sufficient statistics, which would allow us to correctly aggregate studies from disparate study designs (i.e., between/ within subject designs, as discussed in (Morris and DeShon 2002)).
2. The grand majority of the data that would contribute to the meta analysis of all available literature on the topic comes from our two studies (77% in total), which creates a lopsided presentation of results already contained in the manuscript.

We believe that the updated Table 1 provides both the study designs and key results and properly communicates the relevant information to readers in an unbiased fashion.

Below we append the forest plot based on the summary statistics gathered from the literature, yet the results are deficient for the aforementioned reasons.



B. Would it also be possible to extract effect sizes (from the literature) and directions of slopes for the digit ratio interaction?
We provide effect sizes calculated from the literature in the table and also include direction of 2D:4D influence along with attendant hand. We do not find explicit slopes/betas in the literature and do not have the necessary data to estimate digit ratio interactions ourselves (although we estimate the role of 2D:4D in our own data in Fig. S4 in the supplementary materials).

C. Within the Table, I would order the studies like this: Honk, Olsson, Bos, Carre, Exp1, Exp2. There is no need to show bars to illustrate sample sizes.
Thank you for these suggestions, we re-ordered thus and removed n bars for parsimony.

D. The Table needs a detailed legend explaining e.g. “repeated task”, “ES” (note that this abbreviation is not widely known in biology) etc.
Suggestion taken to heart, legend clarifying this design characteristic added.

E. The “Main effect” column contains effect sizes that deviate from the next column (justify!). For all effect sizes you need to clarify the sign (currently they all seem incorrect) and the magnitude (e.g. 0.22 should be -0.19). For the last column it would be better to have quantitative measures and a direction.
We appreciate the careful eye. We updated the effect size measure by using the more conservative Hedge’s bias-corrected effect size instead of the standard Cohen’s d (due to its n-1 pooled sample size and absence of assumption of equal variances between samples) and did

not uniformly update the table; we explain this in the legend as readers may be interested in this analytical detail. The incorrect sign was an oversight, and this too has been rectified in Table 1.

(36) P. 19 end and P. 20 start: I think one could formulate a stronger review of the existing digit ratio literature (but I admit this might be quite some reading effort). What is the main positive evidence that we have and what are the main criticisms? For instance, the observation that Hadza people show no sexual dimorphism is a very weak argument compared to the possibility that the apparent sexual dimorphism may be completely spurious, thereby abolishing the reason for why this trait caught attention and was attributed to sex hormones (see BioRxiv 298786).

Thank you for highlighting the need to address the literature more broadly and for bringing up the allometric issue and sharing this helpful paper on BioRxiv. We now include this important facet of research in that section and provide more support for the statement regarding lack of ethnic universality in the Discussion section of the manuscript (and re-contextualize Apicella et al. (2015) within that discussion).:

"A third reason concerns the validity of the 2D:4D biomarker. The initial findings that prenatal T exposure correlates with 2D:4D are supported in non-clinical and clinical human populations [11], as well as in preliminary causal evidence in relative phalanx/tibia lengths in mice [18][38]. However, recent work highlights concerns regarding the reliability of 2D:4D as a biomarker [39,40]. The 2D:4D of complete androgen insensitivity syndrome patients were found to be only somewhat feminized, and had the same variance as in healthy controls, demonstrating that the preponderance of individual differences in the measure is not attributable to the influence of T exposure [19]. There is also longitudinal evidence that 2D:4D systematically changes during childhood [41,42], which is unconformable with the preposition that it accurately quantifies prenatal influences. Moreover, 2D:4D sexual dimorphism is arguably a necessary condition for the measure's validity due to robust prenatal androgenic differences by sex; although some studies support sexual dimorphism [43,44], some studies suggest lack of ethnic universality of dimorphism [45,46]. Finally, there is debate whether sexual dimorphism is the product of allometric shift in shape rather than hormonal influences [47,48]."

(37) P. 21, L. 6: How about mentioning the limitation that females were not studied here?

Fair point and thank you for the suggestion. We acknowledge the importance of the sex of the participant pool and make a recommendation to resolve potential uncertainty regarding generalization in the Discussion:

"...to both males and females. Future work with females could employ a similar approach as ours characterized by large samples from different geographics, distinct administration methods, and other design features that strongly inform whether a relationship (or its absence) generalizes across sexes."

(38) Please check carefully whether Supplementary Table Legends have the necessary explanations (like I noted for Table 1).

Legends for tables and graphs in the supplementary materials have been elaborated upon for the reader wherever ambiguity or lack of clarity may exist; we thank you for this comment and welcome any further suggestions you may have.

Referee: 2

This manuscript reports two large studies of the effects of testosterone treatment on cognitive empathy in men. In addition, it investigates the role of finger ratios, thought by some to reflect prenatal androgen exposure, in any relations observed. The samples sizes and methodologies are impressive. The findings suggest that testosterone treatment does not influence cognitive empathy in men. The authors also found no evidence for a role of 2D:4D in influences on cognitive empathy.
These are important results. They counteract widely-publicized findings from small samples suggesting that testosterone treatment and 2D:4D play a substantial role in cognitive empathy. I have only a few comments that I offer to improve the manuscript.

1. On page 6, the authors say "We conducted a powerful direct test of the activational and developmental effects of T on cognitive empathy" I think this statement would benefit from rephrasing. The authors clearly successfully manipulated T in adulthood and so likely measured activational effects. As they themselves note in their discussion, the measure of developmental effects, finger ratios, is probably not a reliable measure of early androgen exposure. It might be useful to modify this to say that the current study used similar methodology in larger samples to attempt to replicate prior findings, or something similar.

We appreciate this suggestion, and re-wrote the paragraph to be clearer:
"To this end, we conducted a powerful direct test of the activational and developmental effects of T on cognitive empathy by measuring causal effect of exogenous T and role of putative prenatal androgenic biomarkers in two studies of healthy young men. Our studies constitute the two largest behavioural T administration experiments conducted to date, with samples that were 15 and 25 times greater than that of the previous study conducted in females [17] and 7 and 12 times greater than the largest experiment in males [14], respectively. In both studies we used a computer-based version of the RMET to test the hypothesis that T administration and its purported developmental biomarkers affects cognitive empathy."

2. On page 9, the authors say that for experiment 2, "samples of males with low ethnic heterogeneity". Do they mean that the men were largely Caucasian? If so, that would be easier to understand.
Thank you for this comment, we now specify clearly that the sample is predominantly Caucasian Canadians in Methods > Experiment 2 > Participants: "Experiment 2 included both students and participants from the general public for a total sample of 400 participants (mean age=22.80, SD=4.68). The all-male sample was composed predominantly of Caucasians and overall ethnic heterogeneity was representative of the region (see *Participants* section and Table S1b in Supplementary Material)..."

3. On page 20, the authors say the prior results were "not generalizable". Can they say a bit more here, e.g., "not generalizable across methodologies that increase T concentration" if that is what they mean.

We agree that this was unclear and appreciate the comment. We rephrased this section to reduce ambiguity (**bold** added to denote new content**)**: “However, even if those design differences led to a complete abolishment of a “real” effect of T on cognitive empathy, our results demonstrate beyond a reasonable doubt that such an effect is not generalizable **to both males and females**. **Future work with females could employ a similar approach as ours characterized by large samples from different geographies, distinct administration methods, and other design features that strongly inform whether a relationship (or its absence) exists and whether it generalizes across sexes.**”

****

Journal Name: Proceedings of the Royal Society B
Journal Code: RSPB
Print ISSN: 0962-8452
Online ISSN: 1471-2954
Journal Admin Email: proceedingsb@royalsociety.org
MS Reference Number: RSPB-2019-1062
Article Status: SUBMITTED
MS Dryad ID: RSPB-2019-1062
MS Title: Does testosterone impair men’s cognitive empathy? Evidence from two large-scale randomized controlled trials
MS Authors: Nadler, Amos; Camerer, Colin; Zava, David; Ortiz, Triana; Watson, Neil; Carré, Justin; Nave , Gideon
Contact Author: Amos Nadler
Contact Author Email: amos.nadler@gmail.com
Keywords: testosterone, cognitive empathy, mind reading, prenatal priming, pharmacology, steroid hormones
Abstract: The capacity to infer others’ mental states (known as “mind reading” and “cognitive empathy”) is essential for social interactions across species, and its impairment characterizes psychopathological conditions such as autism spectrum disorder and schizophrenia. Previous studies reported that testosterone administration impaired cognitive empathy in healthy humans, and that a biomarker of prenatal testosterone exposure (finger digit ratios) moderated the effect. However, empirical support for the relationship has relied on small-sample studies with mixed evidence. We investigate the reliability and generalizability of the relationship in two large-scale double-blind placebo-controlled experiments in young men (N=243 and N=400), using two different testosterone administration protocols. We find no evidence that cognitive empathy is impaired by testosterone administration or associated with digit ratios. With an unprecedented combined sample size, these results counter current theories and previous high-profile reports, and demonstrate that previous investigations of this topic have been statistically underpowered.
EndDryadContent