Maastricht, The Netherlands
April 6, 2023

Dear Dr. Espinosa,

We would like to thank you and the reviewers for the very helpful comments and suggestions that allowed us to improve the overall quality of our manuscript, and we appreciate the opportunity to re-submit a revised version to PCI-RR.

Please find our detailed responses to the comments raised by both you and the reviewers below. In short, the main changes include the following:

We have incorporated a power spectrum analysis for our second and third hypotheses as suggested, and uploaded an *R* script of the power simulations to OSF. We have further clarified how we plan to deal with potential ceiling effects. Moreover, we explain why we chose the p-value thresholds for our replication hypotheses as stated in the manuscript, and why, based on those strict thresholds, we deem additional multiple hypothesis correction as not necessary. We have further slightly adjusted the abstract of our manuscript. In addition, we have dealt with all other comments as described in our point-by-point responses.

Modifications to our initial manuscript are indicated in Track Changes. Recommender and reviewer's comments are numbered, with our point-by-point responses below; changes in the manuscript are presented in italic. We have further cited additional literature used to incorporate those changes.

With kind regards,

Charlotte F. Kroll, on behalf of all other co-authors

Department of Psychiatry & Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, Minderbroedersberg 4-6. P.O. Box 616, Maastricht, MD, 6200 The Netherlands
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD, Maastricht, The Netherlands
Department of Microeconomics and Public Economics (MPE), P.O. Box 616, Maastricht 6200 MD, The Netherlands
Email: charlotte.kroll@maastrichtuniversity.nl

Kroll et al. Oxytocin, individual differences, and trust game behavior: a registered large-scale replication

Recommender's comments

1. The Participants section mentions several elements that are mentioned in the design section. I would advice to put the design section before the participants section. For instance, you talk about Experimental Monetary Units (EMU) that you haven't introduced yet.

We agree that it is a good idea to re-arrange the sections and have followed your advice to put the study design and trust game section before the participant section. After rearranging the sections, we carefully checked that the order of introducing important study/design elements is consistent.

2. Please define a procedure for potential floor or ceiling effects. I think that the very purpose of RR is to specify in advance what you plan to do. So, we should avoid as much as possible to leave it to ex-post decisions. (Page 18)

Thank you for this important comment. Since our manuscript describes a replication attempt, we have chosen to not make considerable design/paradigm changes to avoid potential floor or ceiling effects. Instead, we will use statistical approaches in our analysis to account for the potential presence of such effects. We will formally evaluate the presence of a ceiling effect using the proportional odds assumption, i.e., an assumption stating that the effect of an independent variable is constant for each increase in the level of the response (here, investment levels). We will use the *brant* (Brant, 1990) and *nnet* (Venables & Ripley, 2002) packages, implemented in *R* (R Core Team, 2020), to assess the proportional odds assumption. In case of the presence of a ceiling effect based on this criterion, we will conduct three separate binary logistic models, comparing the individual investments (I: 0=0 and 4/8/12=1; II: 0/4=0 and 8/12=1; III: 0/4/8=0 and 12=1) between the OXT and placebo group. We realize that this procedure may have not been completely clear in the initial version of our manuscript. We have therefore added the definition of a ceiling effect to the manuscript using the mentioned packages. Please note: since a floor effect would be unproblematic (our one-sided hypothesis predicts an increase, not a decrease of trust) we would not have to deviate from our original OLR model and, therefore, procedures for floor effects are not discussed in the manuscript. The revised text reads as follows:

"*First,* visual inspection of floor and ceiling effects will be performed as quality checks. A floor effect (i.e., no investment made in the placebo condition is frequent) would be unproblematic, since our directional, one-sided, hypothesis predicts an increase, not a decrease, of interpersonal trust after OXT administration. In contrast, a ceiling effect (i.e., the highest investment is chosen frequently in the placebo condition), would decrease the power to detect an effect of OXT on investments. We will report the distribution of investments and *formally evaluate the presence of a ceiling effect using the proportional odds assumption*[65,67]*, i.e., that the effect of an independent variable is constant for each increase in the level of the response (here, investment levels). In case of the presence of a ceiling effect based on this criterion (brant*[67] *and nnet*[65] *test reveal a* $p \leq .05$*) we will conduct three separate binary logistic (BL) models, comparing the individual investments (I: 0=0 and 4/8/12=1; II: 0/4=0 and 8/12=1; III: 0/4/8=0 and 12=1) between the OXT and the placebo group (see Hypothesis 1a).*"

3. I do not understand why you do not want to correct for multiple comparisons for your registered analyses. By definition, correcting for multiple hypothesis testing requires to know how many tests you will run. It is only possible with registered hypotheses (as we do not know what exploratory analyses you will do and how much of it you will report). You say in the design table that you do not need to correct because "no related analyses are conducted using this variable". Well, you look at the same dependent variable in three hypotheses (H1, H2, H3), with the same dataset. In my view, these hypotheses are part of the same family of hypotheses (the OXT has an effect), which would justify correction.

Thank you for this important comment. We indeed, incorrectly, stated in the design table that no related analyses are conducted using this variable, which we have corrected. However, we want to emphasize that we are already using a relatively low alpha of .02 for testing our replication-related hypotheses 1-3 (as suggested by the *Cortex* guidelines), which is almost equivalent to a Bonferroni correction of the usual alpha of .05; thus, very close to the strictest correction for the usual threshold. We believe that further correcting this already-stringent threshold (e.g., by dividing it again by the number of hypotheses) would result in an unconventional, overly conservative, threshold that would result in a bias towards false negative results.

For H1 the table now reads:

"We believe that one-sided hypothesis testing (investments oxytocin > investments placebo) at $p<0.02$ is justified *since this represents a stricter threshold alpha which we consistently use for testing our replication-related hypotheses 1-3. That is, an alpha of .02 is close to a stringent Bonferroni-corrected alpha of .05 corrected for the number of hypotheses".*

For both H2 and H3 the table now reads:
"We believe that testing for the presence of this group (oxytocin, placebo) by questionnaire score interaction at $p<0.02$ is justified *since this represents a stricter threshold alpha which we consistently use for testing our replication-related hypotheses 1-3. That is, an alpha of .02 is close to a stringent Bonferroni-corrected alpha of .05 corrected for the number of hypotheses."*

4. As a side note: I think that you can use a linear model to analyze the investment decision (eventually a Tobit model in case you fear ceiling or floor effect) because the investment decision is a number of tokens. I do not see here why you would need a non-linear model.

We agree that there are various statistical models that one could use to test our main hypotheses. Importantly, we opted for an OLR since our dependent variable (investments) has only four levels (0, 4, 8, and 12) and the subjectively perceived difference between those investment values might not be equal across individuals (as we discussed in the *Participants* section). We have now added a clarifying sentence in the participant section to make this point more explicit:

"*We chose an OLR because our dependent variable (investments) has only four levels (0, 4, 8, and 12) and the (subjectively perceived) difference between those investment values may not be equal across individuals*."

Furthermore, OLRs offer an intuitive way to assess and account for potential ceiling effects via the proportional odds assumption, as we clarified in response to your second comment and that we now discuss in the *Statistical analyses* section. However, if you feel that our approach is in some way suboptimal, we are happy to discuss this in further detail.

5. Page 21: You say "In case the effect of OXT on investments falls within the range ..." You talk about the confidence interval in the design table, which is clearer than this section. I would suggest to complement it.

Thank you for pointing this out. We agree and have added the rationale from the design table to the *Statistical analyses* section as suggested.

"*In case the 90% confidence interval of the effect of OXT on investments falls within the range of the upper and lower bound of the pre-defined interval, we have evidence to conclude that the OXT* *and placebo group are equivalent (where equivalence would imply that the OXT effect is not meaningful enough to further pursue lab-based studies).”*

6. Why setting alpha=0.02 in the power analysis?

Our decision for an alpha threshold of .02 is based on the mandatory threshold of the PCI-RR friendly journal *Cortex* that we have indicated as preferred outlet when submitting this registered report. Moreover, and as we discuss in the *Statistical analyses* section, the benefit of adopting an alpha of .02 is that it also accounts for the multiple comparisons that are necessary for testing hypotheses 1-3.

“Statistical tests will be considered significant at *p < .02 for our replication analyses (Hypotheses 1-3), and p<.05 for our pooled analysis. For the latter, we deviate from the stricter threshold since we aim to balance specificity and sensitivity for our analysis; maintaining an alpha of .02 would necessitate a sample size that exceeds our data collection capacity. For example, detecting a small effect size of Cohen’s* d=.2 f*or our pooled analysis would require around 850 participants with power=.8 and alpha=.02. We will supplement the test statistics* with Partial Eta Squared values ($\eta_p{}^2$) as a measure of effect size ($\eta_p{}^2$ of .01 indicates small effects, $\eta_p{}^2$ of .06 medium effects, and $\eta_p{}^2$ of equal to or greater than .14 large effects). When testing our registered hypotheses, we will not *additionally* correct for multiple comparisons *since we are using a stricter threshold alpha of .02 for all replication-related hypotheses testing. This alpha of .02 is close to a strict Bonferroni-corrected alpha of .05, corrected for the number of hypotheses.* In any exploratory analyses, we will correct for multiple comparisons and report uncorrected and corrected test results.”

7. Power analysis is made on H1 only. The main issue is that heterogeneity analysis (which is the case here for H2 and H3) requires more observations to maintain a sufficiently high statistical power. It would be nice to compute the expected statistical power of hypotheses 2 and 3.

Thank you for this comment. In this reply, we would like to elaborate on our initial line of reasoning, as well as discuss the changes we have made in response to your comment.

In the original version of this manuscript, we chose to include only a power analysis/simulation for H1 (i.e., the main effect of treatment on investments): our primary hypothesis and the most important reason for conducting this replication attempt. We were able to conduct the power simulation for this hypothesis in a well-informed manner due to the abundance of prior data on the effect of oxytocin on investments in the trust paradigm. Importantly, however, we do not have reliable prior data to conduct an informed power simulation for the interaction between oxytocin treatment and trait scores: only a recent replication of Declerck et al. (2020) has looked at interactions with trust propensity, and there are no well-powered datasets for interactions with reward or punishment sensitivity.

Nevertheless, the power simulation for H2 and H3, as you rightly indicate, remains interesting. In response to your comment, we have therefore adopted the following approach. Because we cannot conduct this power simulation (especially for H3) in a well-informed manner, we have decided to conduct the H2 and H3 power analysis as a "power spectrum analysis", in line with previous work from Quintana (2022). That is, rather than selecting a poorly-informed effect size that represents the (treatment by trait score) interaction, we simulate the power for a range of interaction effect sizes (using our alpha of 0.02 and our planned sample size).

In short, the results of our power spectrum analysis reveal that we would only be well-powered for those situations in which the treatment effect (i.e., the odds ratios for the oxytocin versus placebo effect on investments) is very different for people with a high versus low propensity to trust (or high versus low sensitivity to rewards/punishments). Importantly, this is not an entirely unrealistic situation: Let us assume that approximately 50% of individuals score low and 50% score high on a trait. Further, Declerck et al. (2020) reported that the effect of oxytocin on investment diminishes with higher trust disposition scores. Consequently, to retrieve the effect reported in Kosfeld et al. (2005), the discrepancies between odds ratios in the low and high trait groups would have to be large. We have now added the description and results of the power spectrum analysis in the *Participants* section and added a matrix ("heatmap") that summarizes the result of the power spectrum analysis to the manuscript. The text now reads:

"As opposed to the power simulation for the main hypothesis, the power simulation for the treatment-by-trait score interaction hypotheses is more challenging, primarily because we do not have reliable priors for data simulation. The only well-powered study to date that examined an interaction between OXT (vs. placebo) and trust propensity on investments is the recent replication by Declerck and colleagues[4]. Moreover, to our knowledge, there are no well-powered datasets to conduct an informed power analysis for interactions with sensitivity to reward and punishment. We therefore chose to report a power spectrum analysis for Hypotheses 2 and 3 by estimating the power for a range of (treatment-by-trait) interaction effect sizes, in line with previous work from Quintana[69]. To this end, we simulated the power for a range of differences in odds ratios for the

treatment effect (i.e., OXT vs. placebo effect on investments) between people that score high vs. people that score low on the trait.

Similar to the main power analysis, we used the investment probabilities of the placebo condition from Kosfeld et al.[3] as a starting point to simulate investment probabilities for people who score low, compared to people who score high on a hypothetical trait that may reflect trust propensity, sensitivity to punishment, or reward (for simplicity, a median split is used to create low versus high trait groups). Next, we simulated 1000 random datasets to derive our power for each unique combination of a difference in the treatment effect (odds ratio of 1=very small, odds ratio of 6=very large) for the low and high score trait groups. With alpha set to .02, we obtain the power spectrum results as presented in Figure 3.

As expected, this analysis demonstrates that the power to detect a significant interaction depends on the magnitude of the difference of the treatment effect for individuals who score high, compared to individuals who score low, on the trait. With our planned sample size of N=220, only large differences in the odds ratio would yield sufficient power to detect a treatment-by-trait interaction on investments. Importantly, this is not an entirely unrealistic situation: Let us assume that approximately 50% of individuals score low and 50% score high on a trait and take into account that Declerck et al. (2020) reported that the effect of oxytocin on investments diminishes with higher dispositional trust ratings. Consequently, to retrieve the effect reported in Kosfeld et al.[3], i.e., an effect size of Cohen's d of .51 which approximately translates to an odds ratio of 2.52[68], the discrepancies between odds ratios in the two groups would have to be large."

8. Design Table: I fear that the interpretation is a bit too strong for the first line. If you fail at rejecting H1, does it really mean that it contradicts the theory or does it mean instead that you fail at finding supporting evidence for the theory?

Thank you for making us aware of this. We have changed the sentence in the design table to:

"In case we find no evidence that oxytocin increases trust as measured by the trust game (both H1a and pooled analysis)*, we will have found no supporting evidence for the theory that oxytocin serves a trust-promoting role."*

9. Last, an important issue for me is the clustering of standard errors. Indeed: participants in your experiment spend some time discussing in small groups. This discussion session inevitably generates dependence between observations within the same session. In this case,

we would usually cluster the standard errors at the session level. Note that clustering will drastically reduce your degree of freedom, but it seems the correct way to proceed.

Thank you for pointing this out. This is also an aspect of the study that we have discussed in the research team. We note that, when participants start the minimal social contact procedure, they will have received no information about the trust game, as this step of the experiment is only introduced later. Participants will therefore not undergo the minimal social condition with the realization that they will have to interact once more with one of these participants. We are therefore confident that the observations in the trust game are sufficiently independent. However, if you strongly feel that accounting for clustering is necessary, we would be consider this as an additional “robustness analysis” for the main hypothesis.

10. I agree with the anonymous referee who suggests to put the code of the power analysis online or in the manuscript.

Thank you for bringing this to our attention; we will add a full script containing all power simulations to OSF.

Reviewer 1's comments:

1. The abstract is long and does not read straight to the point. I believe that the authors should consider shortening the abstract by providing only the most important information.

Thank you for pointing this out. We have shortened the abstract and it now reads as follows:

"The neuropeptide oxytocin (OXT) is thought to modulate important aspects of prosocial behavior. In a seminal paper, Kosfeld et al.(2005) reported that intranasally administered OXT modulated trusting behavior in an economic trust game. *Several attempts to conceptually replicate these findings yielded mixed results, which might be partly due to small sample sizes that* can reduce the ability to detect, or reject, meaningful effects. Here, we propose to perform a large-scale replication (*N*=220) of Kosfeld et al. (2005) with specific attention for small effects and subpopulations whose trusting behavior may be sensitive to OXT manipulations. Moreover, we will conduct the largest-ever pooled analysis by merging our data with data from a previous replication by Declerck et al. (2020). *Using additional (equivalence) analyses, we aim* to refute effect sizes of OXT on interpersonal trust that will not be worthwhile pursuing in most *lab-based* contexts. *Our* study will contribute to a more refined understanding of OXT's involvement in human social behavior, for example by identifying boundary conditions that will delineate when OXT-induced effects on prosocial behavior may occur. Critically, we anticipate that our work will offer a more realistic perspective on the effect sizes that can be expected when using intranasal OXT to modulate prosocial behavior."

2. Clarifying question. Are participants told the potential effects of Oxytocin on prosocial behavior? I am assuming that they are not as it would reveal the purpose of the experiment, but I am not familiar with experimental procedures using an intranasal administration.

Participants will not be informed about any potential effects of OXT on prosocial behavior. To make this explicit, we have added a clarifying sentence in the experimental procedure section.

"*They are not informed about any potential effects of OXT on prosocial behavior since this would reveal the purpose of this experiment*."

3. Are the items from IGTS and SPSRQ-RC randomized? Is the order of display of the IGTS, SPSRQ-RC and NOSE randomized? The authors should mention it in the text (please indicate where it is mentioned in the text in case I missed it).

Thank you for your comment. We will randomize the order of the questionnaires in the first (online) session and have added a sentence below Table 1 to clarify this. We will however keep the order of items within the questionnaires as suggested in the original studies.

“**Table 1.** Self-reported symptoms and traits assessed during session 1. *Questionnaires will be presented randomized at the individual level.*”

4. The authors should be more accurate on the statistical threshold alpha. In the power analysis, the authors start with alpha = 0.02, then increase to 0.05 when changing the Cohen’s D. I understand that the authors have increased the alpha threshold to 0.05 because they have decreased the Cohen’s D (from 0.51 to 0.2) and want to keep the probability to detect an effect above 0.8. I believe that the authors should consider keeping one statistical threshold alpha for all the statistical analyses in the paper, including the power analysis and the planned analyses. Simply increasing alpha to 0.05 and re-running the first part of the power analysis for the Cohen’s D at 0.51 should solve this issue.

Thank you for raising this important point. For all three hypotheses related to the replication, we use an alpha of .02 (as suggested by the *Cortex* guidelines). Moreover, the benefit of using this threshold is that it accounts for the multiple comparisons that we conduct as part of testing hypotheses 1-3 (also see our reply to the recommender’s comment 6). The only deviation from this threshold is for our pooled analysis, where we balance the specificity and sensitivity of our analysis. Maintaining an alpha of .02 here would necessitate a sample size that exceeds our data collection capacity. For example, detecting a small effect size of Cohen’s d=.2 or our pooled analysis would require around 850 participants with power=.8 and alpha=.02. In the *Participants* section, we now explicitly mention when we rely on an alpha of 0.02 (i.e., for hypotheses H1-H3), and when we rely on an alpha of 0.05 (i.e., only for the pooled analysis). Moreover, we have clarified in the text which power analyses belong to which hypothesis or pooled analysis.

5. From what I understand, the authors have only run the power analysis for hypothesis 1a. The authors need to run their power analysis for each outcome that is to be tested, define a

minimum effect size of interest for each tested outcome and report the probability to detect for each outcome. For example, hypothesis 2 tests whether the effect of oxytocin on investments will decrease with increasing trust propensity scores. Here, the authors should define a minimum effect size of interest for the IGTS (see Dienes, Z. (2021a) Obtaining evidence for no effect. https://doi.org/10.31234/osf.io/yc7s5), run the power analysis for this outcome variable and report the probability to detect.

Thank you for this comment, which we have addressed (by conducting another power analysis) in our response to the recommender's comment 7.

6. The authors should also include multiple hypothesis correction in their power analysis (see my comment below). The power analysis R code needs to be included in the stage-1.

Thank you for this important comment. As also discussed in our response to the recommender's comment 3, we have adopted a lower alpha than commonly used (as also suggested by *Cortex* guidelines). Our alpha of 0.02 is close to a Bonferroni-corrected alpha of .05, given our number of comparisons (i.e., our alpha approximates the strictest multiple comparisons correction). We consistently use this threshold for testing our replication-related hypotheses 1-3. We hope that the reviewer agrees with our rationale. We will also add a full script containing all power simulations to OSF.

7. The authors should clarify how they will deal with potential floor or ceiling effects. While the authors anticipate the effects "We will report the distribution of investment and will take potential ceiling effects into account in our statistical analyses", they do not provide any detail on how they will deal with them in case they happen. I believe that the authors should provide outcome neutral tests for each outcome variable they plan to analyze in the confirmatory section (for more information on outcome neutral tests please see https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4110803 by Espinosa and Arpinon, and "The inner workings of Registered Reports" by Zoltan Dienes).

Thank you for this important comment which was also made by the recommender. We refer to our justification discussed in our response to the recommender's point 3. In brief, we have now added information about how we test for the presence of ceiling effects in the *Statistical analyses* section.

8. First, the authors should include some form of multiple hypothesis correction. The authors mention the following: "When testing our registered hypothesis, we will not correct for multiple comparisons. In any exploratory analyses, we will correct for multiple comparisons and report uncorrected and corrected test". The fundamental goal of a Registered Report is to test a hypothesis or set of hypotheses to draw strong statistical conclusions, while leaving any unregistered set of results as exploratory results, upon which no clear conclusions are drawn. Here, the authors should reverse their approach and include a correction for multiple hypothesis testing when testing for the set of registered analyses. The authors are then free to correct, or not, for multiple comparisons in the exploratory section.

We understand your point and admit that our original formulation may have been confusing. As also mentioned in the response to your comment 6 above, the use of an alpha=0.02 serves also the purpose of correcting for multiple comparisons for pre-registered hypothesis 1-3. However, we did not refer to it as "corrected" because this threshold was also recommended by *Cortex*, the PCI-RR journal that we have indicated as the preferred outlet when submitting this registered report. We feel that employing an additional correction would be overly stringent and unusual. Therefore, we have decided to keep an (uncorrected) alpha of 0.02 for hypotheses 1-3.

9. Additionally, the presentation of the hypotheses could be improved. I think that the pooled analysis could be included as hypothesis 4, as it is an important analysis to be conducted. If the authors do not believe that this analysis is central, they should remove it from the stage-1 and simply include it in the exploratory analysis.

Thank you for this comment. Our first three hypotheses are related to the replication part of our study; we also use the same significance threshold of 0.02 for those hypotheses. The pooled analysis is an addition to the replication hypotheses, which uses a different threshold, and should be seen as being separate from the first three hypotheses. Nevertheless, we believe that it is essential to keep this pooled analysis and our expectations about it as a pre-registered part of the manuscript. Therefore, we have decided to keep the labels as they are and hope that the reviewer agrees with our rationale.

10. Hypothesis 1a should be reduced. The role of the covariate NOSE should be explored in the exploratory analysis and should not be included in the hypothesis section. If the authors wish to analyze the role of NOSE in the confirmatory analysis, they should formulate an

> additional hypothesis. I believe that the following could also be removed from hypothesis 1a: « the data will first be (visually) explored using summary statistics and frequency tables as well as distribution characteristics » as it will not be formally tested using statistical tests.

Thank you for the suggestion to remove the part on visually exploring the data and have done so. At the same time, we feel that the NOSE is an important covariate that has the potential to affect our results (see e.g., Declerck *et al*. who report a non-trivial number of participants that scored relatively high on this assessment). Given that, *a priori*, we expect that this variable may affect our results we strongly feel that it should always be included in our analyses. If the reviewer feels that the effect of the NOSE score should be purely exploratory, then our suggestion is to repeat H1 after excluding participants that score above an arbitrary NOSE score, as an exploratory analysis.

Reviewer 2's comments:

1. The term "minimal effects analysis" in the abstract is unclear to me. Do you mean equivalence testing?

Thank you for pointing this out. We have revised and shortened the abstract to better clarify our study aims. We specifically now clarify that a minimal effects analysis (also known as "equivalence testing") serves to demonstrate whether the observed effect of oxytocin is even worthwhile pursuing. We further explain the concept of minimal effects analysis in the introduction. The abstract now reads:

"The neuropeptide oxytocin (OXT) is thought to modulate important aspects of prosocial behavior. In a seminal paper, Kosfeld et al.(2005) reported that intranasally administered OXT modulated trusting behavior in an economic trust game. *Several attempts to conceptually replicate these findings yielded mixed results, which might be partly due to small sample sizes that* can reduce the ability to detect, or reject, meaningful effects. Here, we propose to perform a large-scale replication (*N*=220) of Kosfeld et al. (2005) with specific attention for small effects and subpopulations whose trusting behavior may be sensitive to OXT manipulations. Moreover, we will conduct the largest-ever pooled analysis by merging our data with data from a previous replication by Declerck et al. (2020). *Using additional (equivalence) analyses, we aim* to refute effect sizes of OXT on interpersonal trust that will not be worthwhile pursuing in most *lab-based* contexts. *Our* study will contribute to a more refined understanding of OXT's involvement in human social behavior, for example by identifying boundary conditions that will delineate when OXT-induced effects on prosocial behavior may occur. Critically, we anticipate that our work will offer a more realistic perspective on the effect sizes that can be expected when using intranasal OXT to modulate prosocial behavior."

2. "The hormone oxytocin (OXT) is a nine amino acid neuropeptide that is synthesized in the hypothalamus…" It is mainly synthesized in the hypothalamus, it is also synthesized at other sites in the body, but in much smaller amounts.

Thank you for pointing this out. We have incorporated this change in the introduction section. We now write:

"The hormone oxytocin (OXT) is a nine amino acid neuropeptide that is *mainly* synthesized in the hypothalamus and that acts both centrally and peripherally[1,2]."

3. It should be briefly noted why small sample sizes bias true effect sizes. It may also help to describe the types of effect sizes the original study was powered to reliably detect.

Thank you for this comment. We have incorporated a brief explanation in the Introduction section:

"*If only studies with* small sample sizes *that happen to achieve statistical significance are published, which can only happen when the effect size is large, i.e., under publication bias*[41]*, biased effect estimates can arise.*"

We have also described the effect size for the main effect reported in the original study in the *Participants* section:

"Critically, however, mixed results from previous OXT-trust studies suggest that the true effect regarding an increase in interpersonal trust following OXT administration, e.g., [4,37,39] is smaller than the effect size of Cohen's *d* of .51 (corresponding to an odds ratio of 2.52[68]) reported in the original study by Kosfeld et al.[3]."

4. "there is the possibility that OXT may have a smaller effect, perhaps limited to particular subpopulations": This brings to mind a recent paper describing the importance of

recognising the heterogeneity of study populations (https://www.nature.com/articles/s41562-021-01143-3). However, I will leave it up to the authors if they wish to mention this paper and/or the broader issue of recognising population heterogeneity

Thank you for suggesting this interesting paper. This is an important subject and we have included a sentence in the introduction section and cited the paper:

“*This is in line with the importance of considering heterogeneity in study populations to establish comprehensive explanations of potential causal mechanisms*[58]*.”*

5. “…and applying proper statistical techniques to improve interpretability”: some examples of these techniques should be named.

We thank the reviewer for mentioning this and have added an example in the *Introduction* section:

“The discussed considerations, most notably the growing skepticism regarding OXT effects in the context of social behavior, e.g.,[39], clearly indicate the need for further transparent investigation and replication of intranasal OXT studies using a sufficiently large sample size and applying proper statistical techniques to improve interpretability, *such as equivalence testing to more reliably detect or reject meaningful effects*[47](e.g.,[39,40,44]).”

6. “…which they suspect (and we confirmed) to result from a clipped or misprinted aspect of the figure”: How was this confirmed? By the authors?

Thank you for this question. This was indeed not clearly formulated in the submitted version of our manuscript. We now clarify in the *Participants* section that we confirmed this by changing the background color of the figure:

“Secondly, as noted by Calin-Jagermann and Cumming[62], the bar heights for the placebo group in the figure sum up to less participants than reported in the text, which they suspect (and we confirmed *by changing the background color of the pdf*) to result from a clipped or misprinted aspect of the figure.”

7. Table 1 - there is a gap just before “inattentiveness”

Thank you for pointing this out. We corrected it.

8. I like the author’s approach for power analysis (i.e., simulations). Is this code available? Perhaps I missed the link? I would recommend posting this code on OSF.

Thank you for bringing this to our attention. We will add a full script containing all power simulations to OSF.

9. Regarding dose, I understand the choice of 24IU, given this is a replication, but potential dose-dependent effects of oxytocin should also be mentioned in the article.

Thank you for pointing this out. We have added a sentence in the oxytocin administration section to clarify this as follows.

“*Although previous studies have reported dose-dependent effects of OXT on psychological measures*[76] *consistent with the original study by Kosfeld et al.*[3]*, and for feasibility reasons, we use a single dose of 24 I/U here.*”

10. Figure 1b - It would be helpful to make the axis tick labels with a slightly larger font

We enlarged the axis tick labels. Thank you for mentioning this.

11. “where it can cross over to the hypothalamus…” Would be more accurate to say something like “where it can travel to the hypothalamus…”

Thank you for pointing this out. We have incorporated this suggestion in the oxytocin administration section:

“In previous research, it has been assumed that intranasal administration of OXT crosses the blood-brain barrier via a layer of cells in the nasal cavity epithelium, where it can *travel* to the hypothalamus[1], thereby exerting central nervous system effects.”

12. Why was alpha set to .02 and not .025?

Thank you for this question. We selected an alpha of .02 since this is the mandatory significance threshold for *Cortex*, the PCI-RR friendly journal that we indicated as our preferred choice. As also addressed in the response to comment 3 of the recommender, the use of this threshold has the additional benefit that it accounts for multiple comparisons that we conduct when testing hypotheses 1-3.

13. We used extracted probabilities for the placebo condition from Kosfeld et al. and assumed a more reasonable true minimal effect size, a Cohen's d =.2." This Cohen's d value seems reasonable, but can the authors provide a justification for this particular value?

Thank you for this important comment. The reason why we attempted to power our study to find an effect of (Cohen's *d*) equal to .2 is that, in our view, effects smaller than .2 are less interesting to pursue as it is infeasible to detect such effects in lab studies due to the large number of observations needed. We added a sentence in the participant section to clarify this:

"*We further believe that demonstrating an effect smaller than d =.2 is infeasible in most human lab studies. For instance, a Cohen's d of .1 would require a sample size of around 2500 participants."*

14. When using the present experimental design for a different population, it should be kept in mind that OXT administration can induce uterine contractions": While the risk of inducing uterine contractions in pregnant women is certainly a consideration for not including females, this is usually mitigated by administering pregnancy tests prior to administration. I would assume the more likely reason that females have not been usually included is the potential impact of different hormone levels across the menstrual cycle on oxytocin effects.

Thank you for pointing this out. Our explanation was meant as a rationale for using a safety measure, rather than as a rationale for using male subjects only (we use male subjects only since the original study did so). We have decided to delete this sentence to avoid confusion and will dedicate a section regarding generalization of results in our discussion section.

15. Our experimental procedure will be partially based on the replication by Declerck and colleagues, with adjustments made to serve the purpose of our study". I think it would be

very useful to have a table or text box that summarizes the adjustments made and why these were made.

Thank you for this suggestion. We have discussed this in the research team and concluded that it is very hard to completely cover where we deviate from Declerck et al. (2020) in a table because the description of deviations also depends on the (micro-)level at which readers will define this (e.g., even time and space differ by nature in every replication). Therefore, we deem the presentation of a seemingly complete table infeasible. Instead, we added text that summarizes the – in our opinion – most important practical differences in setting up and running the sessions. This information should enable other researchers to smoothly re-replicate our study in the future. The sentence reads now:

"*Specifically, compared to the replication study by Declerck et al.*[4]*, the current replication effort differs in terms of questionnaires (omission of some questionnaires that did not directly address objectives of the current study; addition of questionnaires, most notably the SPSRQ-RC*[78] *for testing our third hypothesis), reimbursement (guaranteed payout of all trust game decisions to increase fairness, as opposed to only guaranteed payout of investor decision in Declerck et al.*[4]*), the collection of saliva samples (for intended analysis of hormonal markers), and the addition of a generalized dictator game*[79] *in session 2, of which participants will not be aware when playing the trust game.*"

16. "…with levels of chronic nasal obstruction as a covariate": The method of evaluation is only described further down, but it should be introduced earlier.

Thank you for pointing this out. Chronic nasal obstruction will be assessed by the NOSE questionnaires as introduced in Table 1 and used as a covariate in the further analyses. We added the name of the questionnaire after "with levels of further nasal obstruction" in the experimental procedure section for clarification. There it now reads:

"Baseline measures of dispositional trust and sensitivity to punishment and rewards will be used when testing our registered hypotheses, with levels of chronic nasal obstruction *(NOSE*[80]*)* as a covariate, whereas the remaining measures will serve as moderators in potential further exploratory analyses."

We further add a separate rating of current nose obstruction at the end of the study and will exclude those participants from data analysis since this score has been shown to correlate with inspiratory flow in the upper nasal cavity (see our clarifications in the participant and experimental procedure section). The texts now reads as follows:

“Lastly, participants will rate their current nose obstruction on a scale from 0 (incredibly clear) to 10 (incredibly blocked). *A score of >8 will be considered as severe nasal obstruction*[90].”

“Those who fail to complete the investment paradigm, *report to have severe nose obstruction on the day of the experiment, or did not comply with rules abstaining from alcohol, and smoking (see Experimental procedure)* will be excluded from data analysis.”

17. What is the purpose of the saliva samples? Are you evaluating oxytocin receptor SNPs and peripheral oxytocin levels? It seems oxytocin concentrations are being evaluating pre and post administration, but the utility of measuring oxytocin in saliva post intranasal administration is compromised by the “drip down” of exogenous oxytocin from the nasal cavity to the oral cavity. So rather than measuring circulating levels of oxytocin, this approach mainly measures exogenous oxytocin. Indeed, after intranasal administration, saliva oxytocin levels are not related to peripheral levels measured in blood plasma (https://doi.org/10.1016/j.yhbeh.2018.05.004). Are there any predictions regarding oxytocin receptor SNPs and the the effects of oxytocin on trusting behaviors?

Thank you for pointing out the issue regarding quantifying OXT in saliva and pointing us to this important paper. We collect saliva for protein or other hormonal assessments and believe the results are of potential interest. At this point, we have not developed strong a priori hypotheses regarding these markers, but there may be scenarios in which we conduct exploratory analyses (in the current or other manuscripts) with these data.

18. Consequently, the upper bound (ΔU) will be set to d =.33, and the lower bound (ΔL) will be set to d =-.33 (testing one-sided).“ I understand the justification for these equivalence test bounds, but these bounds are relatively large. The median summary effect size for oxytocin administration study meta-analyses is 0.14, and this doesn’t even account for publication bias, so the “true” effect is likely smaller (https://doi.org/10.1016/j.cpnec.2020.100014)

Thank you for this comment. We agree that these bounds (and the effects) are relatively large. However, it is important to point out the rationale behind this analysis. The aim is not to demonstrate that oxytocin has no effect at all. Indeed, this is what one would likely conclude if we would demonstrate equivalence for an extremely small effect size like Cohen's d=0.1. But this analysis would be infeasible as it would require an enormous sample (~2500 participants) that could never be collected in a single (or even multi) site study.

Rather, our aim is to demonstrate that oxytocin's effect on investments is not worthwhile pursuing in typical lab-based studies, where we collect much less data (e.g., somewhere between 20-50 per treatment group according to Nave et al., 2015). As such, we need to use larger bounds for equivalence testing, because the great majority of psychology studies are not powered to find small effects and, arguably, the effects we commonly observe in psychological research are much smaller (e.g., around Cohen's *d*=0.5 for non-preregistered studies; Schäfer & Schwarz, 2019).

We therefore settled on equivalence test bounds that are far more realistic for psychological/experimental economics research. Importantly, this implies that we can demonstrate equivalence for an effect size that is smaller than the median effect size commonly observed in psychological studies (Schäfer & Schwarz, 2019). Although we agree with the reviewer that using smaller bounds would certainly be interesting, we once more want to underline that it is not our aim to demonstrate that oxytocin has no effect at all; we simply want to demonstrate that the effect is not worthwhile pursuing in most lab-based studies that study the effect of oxytocin on decision-making. We have now clarified our line of reasoning in the *Statistical analyses* section. The section now reads:

"Hypothesis 1a will be supplemented with equivalence testing using the *TOSTER* package[91] implemented in *R*[66]. Here, the aim is to assess, in the event that the OXT group, relative to placebo group, shows a significant increase in monetary investments, whether OXT's effect on interpersonal trust is large enough to be considered *a meaningful psychological finding*[40,47]. We will set the upper bound (Δ_U) t to *d*=.33, and the lower bound (Δ_L) to *d*=-.33 (testing one-sided). *Although at first glance this may seem like a rather wide range, it is not our aim to demonstrate equivalence for the smallest possible effect sizes. Rather, we aim to demonstrate equivalence for a range of effect sizes that are smaller than effect sizes commonly encountered in psychological research*[92]*. Moreover, this effect size is in the range of the mean effect size reported by previous studies that have examined OXT's effects in trust-related contexts*[93-95].

In case the 90% confidence interval of the effect of OXT on investments falls within the range of the upper and lower bound of the pre-defined interval, we have evidence to conclude that the OXT *and placebo group are equivalent (where equivalence would imply that the OXT effect is not meaningful enough to further pursue in lab-based studies).*"

19. We will report the distribution of investment and will take potential ceiling effects into account in our statistical analyses." Can authors provide a suggestion for how this will be taken into account, if necessary?

Thank you for this important comment, which was also raised by the recommender. We kindly refer you to our response to comment 2 of the recommender, where we explain exactly how we will deal with ceiling effects in our statistical models.

Further comments from our side:

1. We have added the SVO as a questionnaire in the first session as a potential moderator in further exploratory analyses (see Table 1).
2. We slightly changed the order of saliva/questionnaire and start of the experiment in the experimental procedure section since we believe this order to be more appropriate (i.e., participants start the trust game immediately after receiving the instructions without being interrupted by providing a saliva sample).
3. We have deleted the text erroneously stating that the minimal social contact will be imposed and structured. To clarify: we use the minimal social condition used in Kosfeld et al., which was discussed in detail in Declerck et al (which involved most of the original authors of Kosfeld et al).

Newly added literature to account for the recommender and reviewer's comments:

(1) Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

(2) Quintana, D. S. (2022, July 11). A guide for calculating study-level statistical power for meta-analyses. https://doi.org/10.31219/osf.io/js79t

(3) Quintana, D. S., Westlye, L. T., Hope, S., Nærland, T., Elvsåshagen T., Dørum, E., Rustan, Ø, Valstad, M., Rezvaya, L., Lishaugen, H., Stensønes, E., Yaqub, S., Smerud, K. T.,

Mahmoud, R. A., Djupesland, P. G., & Andreassen, O. A. (2017). Dose-dependent social-cognitive effects of intranasal oxytocin delivered with novel Breath Powered device in adults with autism spectrum disorder: a randomized placebo-controlled double-blind crossover trial. *Translational Psychiatry, 7*(5), e1136–e1136. https://doi.org/10.1038/tp.2017.103

(4) Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*(813). https://doi.org/10.3389/fpsyg.2019.00813

(5) Teixeira, R. U. F., Zappelini, C. E. M., Oliveira, L. G., Basile, L. C. G., & da Costa, E. A. (2011). Correlation Between the Peak Nasal Inspiratory Flow and the Visual Analogue Scale Before and After Using a Nasal Decongestant. *International Archives of Otorhinolaryngol, 15(*2), 156–162. http://www.arquivosdeorl.org.br/additional/acervo_eng.asp?id=758

(6) Van Lange, P. A. (2000). Beyond self-interest: A set of propositions relevant to interpersonal orientations. *European Review of Social Psychology, 11*(1), 297–331. https://doi.org/10.1080/14792772043000068

(7) Mikolajczak M., Gross J. J., Lane A., Corneille O., de Timary P., Luminet O. (2010). Oxytocin makes people trusting, not gullible. *Psychological Science, 21*, 1072–4.

(8) Klackl J., Pfundmair M., Agroskin D., Jonas E. (2013). Who is to blame? Oxytocin promotes nonpersonalistic attributions in response to a trust betrayal. *Biological Psychology, 92*, 387–394

(9) Yao S., Zhao W., Cheng R., Geng Y., Luo L., Kendrick K. M. (2014). Oxytocin makes females, but not males, less forgiving following betrayal of trust. *The International Journal of Neuropsychopharmacology, 17*, 1785–92

Literature cited in this rebuttal letter:

(1) Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics, 46*(4), 1171–1178. https://doi.org/10.2307/2532457

(2) Declerck, C. H., Boone, C., Pauwels, L., Vogt, B., & Fehr, E. (2020). A registered replication study on oxytocin and trust. *Nature Human Behaviour, 4*(6), 646–655. https://doi.org/10.1038/s41562-020-0878-x

(3) Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673–676. https://doi.org/10.1038/nature03701

(4) Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science, 10*(6), 772–789. https://doi.org/10.1177/1745691615600138

(5) Quintana, D. S. (2022, July 11). A guide for calculating study-level statistical power for meta-analyses. https://doi.org/10.31219/osf.io/js79t

(6) R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

(7) Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*(813). https://doi.org/10.3389/fpsyg.2019.00813

(8) Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S (4th ed.).* Springer.