Referee Report Thibaut Arpinon

"Oxytocin, individual differences, and trust game behavior: a registered large-scale replication"
by Charlotte F. Kroll, Koen Schruers, Wolfgang Viechtbauer, Claudia Vingerhoets, Leonie Seidel, Arno Riedl, and Dennis Hernaus

I would like to thank the authors for the very interesting read and the quality of the proposed stage-1. The work the authors are doing is very important in terms of replication and a Registered Report is well suited to provide a strong and clean scientific contribution to the literature on the role of Oxytocin. I found the introduction to be very informative, accurate, and straightforward on what Oxytocin is and its role. It also highlights very well the limitations of the literature and how the current paper could improve the evidence on Oxytocin. The procedure is thorough, and the authors have planned out carefully how this study will be conducted.

However, I believe that the authors have missed some fundamental elements of Registered Reports, and considering those elements will greatly improve the quality of the stage-1. I have described below some minor points and major points that the authors should consider to improve the quality of the stage-1 manuscript.

Minor 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.

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.

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).

Major comments

Statistical threshold alpha

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.

Power analysis

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.

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.

Outcome neutral tests

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).

Multiple hypothesis testing

I believe that the authors could greatly improve the statistical validity of their Registered Report by improving the hypothesis section.

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.

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.

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.