Psychoneuroendocrinology 107 (2019) 124–132

Contents lists available at ScienceDirect

Psychoneuroendocrinology

journal homepage: www.elsevier.com/locate/psyneuen

Null results of oxytocin and vasopressin administration across a range of social cognitive and behavioral paradigms: Evidence from a randomized controlled trial

Benjamin A. Tabak[a,*], Adam R. Teed[a], Elizabeth Castle[b], Janine M. Dutcher[c], Meghan L. Meyer[d], Ronnie Bryan[e], Michael R. Irwin[b,f,g], Matthew D. Lieberman[b,f], Naomi I. Eisenberger[b]

[a] Department of Psychology, Southern Methodist University, Dallas, TX, United States
[b] Department of Psychology, University of California, Los Angeles, CA, United States
[c] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States
[d] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, United States
[e] Los Angeles Trade-Tech College, Los Angeles, CA, United States
[f] Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, CA, United States
[g] Cousins Center for Psychoneuroimmunology, Jane and Terry Semel Institute for Neuroscience, David Geffen School of Medicine, University of California, Los Angeles, CA, United States

ARTICLE INFO

Keywords:
Oxytocin
Vasopressin
Social behavior
Social cognition
Social processes

ABSTRACT

Research examining oxytocin and vasopressin in humans has the potential to elucidate neurobiological mechanisms underlying human sociality that have been previously unknown or not well characterized. A primary goal of this work is to increase our knowledge about neurodevelopmental and psychiatric disorders characterized by impairments in social cognition. However, years of research highlighting wide-ranging effects of, in particular, intranasal oxytocin administration have been tempered as the fields of psychology, neuroscience, and other disciplines have been addressing concerns over the reproducibility and validity of research findings. We present a series of behavioral tasks that were conducted using a randomized, double-blind, placebo controlled, between-subjects design, in which our research group found no main effects of oxytocin and vasopressin on a host of social outcomes. In addition to null hypothesis significance testing, we implemented equivalence testing and Bayesian hypothesis testing to examine the sensitivity of our findings. These analyses indicated that 47–83% of our results (depending on the method of post-hoc analysis) had enough sensitivity to detect the absence of a main effect. Our results add to evidence that intranasal oxytocin may have a more limited direct effect on human social processes than initially assumed and suggest that the direct effects of intranasal vasopressin may be similarly limited. Randomized controlled trial registration: NCT01680718.

1. Introduction

In mammals, the neuropeptides oxytocin (OT) and vasopressin (AVP) are produced in magnocellular neurons of the hypothalamus and released into the periphery via the posterior pituitary (Donaldson and Young, 2008). In addition to their role in mediating many physiological functions (e.g., lactation, bone density, water reabsorption, metabolism, homeostasis), it is the contribution of both neuropeptides to social processes such as pair bond formation that has spurred the greatest interest in the fields of psychology and psychiatry (Donaldson and Young, 2008). Based on animal research documenting a host of changes in social behavior that can be produced by altering one or both of these biological systems, researchers have hoped to improve our understanding of disorders characterized by social cognitive deficits (e.g., autism spectrum disorders and schizophrenia) through studies investigating human OT and AVP. Indeed, these neuropeptides, and in particular OT, have generated tremendous enthusiasm for their potential translational influence on human social cognition and behavior (Insel, 2016).

However, early enthusiasm for human research on the effects these neuropeptides, particularly OT, has diminished due to a host of methodological issues (Leng and Ludwig, 2016; Szeto et al., 2011) and

* Corresponding author at: Department of Psychology, Southern Methodist University, 6116 N. Central Expressway, P.O. Box 750442, Dallas, TX, 75206, United States.
E-mail address: btabak@smu.edu (B.A. Tabak).

https://doi.org/10.1016/j.psyneuen.2019.04.019
Received 30 January 2019; Received in revised form 25 April 2019; Accepted 26 April 2019

analytical concerns that are receiving increasing emphasis across psychology (Walum et al., 2016). A healthy dialogue between engaged skeptics and nuanced proponents is emerging, prompting research groups to recalibrate their efforts toward more careful and systematic approaches to understanding the complex relationships between OT, AVP, and human sociality (Leng and Ludwig, 2016; Quintana and Woolley, 2016). Skeptics point to several outstanding questions including the different methods (and tissues) used to measure endogenous OT (McCullough et al., 2013; Szeto et al., 2011), the lack of relationship between basal concentrations of peripheral and central OT but a positive association after OT administration and stress induction (Valstad et al., 2016), and the unknown pharmacokinetics of intranasal neuropeptide administration (Leng and Ludwig, 2016). Research practices resulting in underpowered studies, inflated effects, and potential publication bias, though certainly not unique to this subfield of psychology and neuroscience, have been highlighted by others (Walum et al., 2016). Compared to the hundreds of studies on human OT, the limited amount of studies examining the social role of AVP presents a challenge in synthesizing our current state of knowledge. Nonetheless, the issues listed previously also apply to the study of AVP.

In the past 10–15 years, perhaps the most widely used method to examine the social effects of OT (and AVP) has been intranasal administration of these peptides. Recent meta-analyses in healthy samples have found a positive effect of OT on the recognition of basic emotions (Hedge's g = .13 [0.02, 0.24]), increased expression of positive emotions (Hedge's g = 0.25 [0.04. 0.47]; Leppanen et al., 2017), and increased physiological startle responses to threat (Hedge's g = 0.3 [0.07. 0.53]; Leppanen et al., 2018). In addition, a meta-analysis of individuals with neurodevelopmental and psychiatric disorders found a positive effect of OT on theory of mind (Hedges' g = 0.21 [0.01. 0.41]; Keech et al., 2018). However, the authors of two of these meta-analyses (Leppanen et al., 2017; 2018) point out that none of the studies using a between-subjects design were sufficiently powered to report the detected effects at 80% statistical power. In addition, given the problem of publication bias, meta-analyses conducted to date contain fewer studies finding null effects. Nonetheless, recent meta-analyses have shown no meta-analytic effects of OT on a variety of social outcomes in healthy samples including: theory of mind (Leppanen et al., 2017), the expression of negative emotions (Leppanen et al., 2017), and attentional or behavioral responses toward threat (Leppanen et al., 2018). In clinical samples, no meta-analytic effect was found in physiological, attentional, or behavioral measures relevant to threat (Leppanen et al., 2018), the interpretation or expression of emotion (Leppanen et al., 2017), or emotion recognition or empathy (Keech et al., 2018). A recent report (Lane et al., 2015) directly addressed the state of intranasal OT administration research with a meta-analysis of a set of their own predominantly null and previously unreported findings that prompted their conversion "from believers into skeptics." Quintana (2018) responded to Lane et al. (2015) by drawing attention to the inadequacy of null-hypothesis statistical testing (NHST) to support true null effects. While the results of Lane et al.'s (2015) meta-analysis also failed to attain significance, Quintana (2018) reanalyzed their data and applied equivalence testing which showed that 73.5% of the tests did not result from statistical equivalence between tested conditions, but rather from insensitivity due to lack of power.

Quintana and Williams (2018) separately utilized Bayesian hypothesis testing to assess the quality of an intranasal OT study with a positive, but underpowered effect. Bayesian testing is particularly beneficial for efficiently providing information on the relative degree of evidence that the data provide in favor of either the alternative or null hypotheses. The ability to detect the sensitivity of data may help assess whether a null result of uncertain quality may represent a question worth revisiting, and when cumulative results from multiple studies may be needed to adequately assess the veracity and stability of an effect, or lack thereof.

In the present article, we follow Lane et al. (2015) in reporting a lack of behavioral main effects for several paradigms using intranasal OT *and* AVP and use techniques employed by Quintana (2018; Quintana and Williams, 2018) to assess the quality of these analyses. However, unlike Lane et al. (2015) we chose not to conduct an internal meta-analysis of our findings (for a discussion of potential problems associated with conducting internal meta-analyses see Vosgerau et al., 2018). Like many research groups who were and/or are interested in studying the effects of intranasal OT and AVP, when we developed the experimental tasks described in the present paper in 2011–2012, we sought to investigate the effect of these neuropeptides on several social processes by extending work that had already been conducted, and also including several novel outcomes.

Six behavioral tasks will be presented in no specific order as all tasks were completed by the same participants during the same session in randomized order. Two of the tasks have been previously reported and examined the effect of OT or AVP on: 1) self-reported empathic concern when viewing uplifting or distressing videos (Tabak et al., 2015), and 2) performance on social and non-social working memory tasks (Tabak et al., 2016). Importantly, the goal of our two previous reports was to focus on main and interaction effects, whereas the present analyses focus specifically on main effects across all tasks. In the present study we also present results from four previously unreported tasks that examined the main effects of OT or AVP on: 3) deception detection (as in Israel et al., 2014; Pfundmair et al., 2017), 4) perceptions of trustworthiness or threat based on interpersonal distance (similar to outcomes studied in Cohen et al., 2018; Perry et al., 2015; Scheele et al., 2012), 5) a hypothetical bystander intervention, and 6) written reflections on a supportive interaction or a conflict with a close other (see Supplemental Materials for the background rationale of each task). It is important to note that our study began before the conceptually similar studies cited above were published.

Across the six behavioral tasks, we present separate analyses of main effects of OT vs. placebo (PLA) and AVP vs. PLA on 18 dependent variables (i.e., a total of 36 statistical comparisons). In the Supplementary Material, we describe the results of analyses using NHST, which showed null results for both drugs vs. PLA across all dependent variables. In the main text, we detail the results of equivalence testing and Bayesian hypothesis testing based on the null findings.

2. Methods

2.1. Participants

As noted in our previous work (Tabak et al., 2015), to achieve > 80% power, we sought to recruit more than 34 participants in each condition (i.e., OT, AVP, and PLA) based on previous studies showing moderate to large main effects of AVP (e.g., d = 0.7) on social cognitive processes (Uzefovsky et al., 2012). As described in Tabak et al. (Tabak et al., 2015), participants included 125 undergraduate students from the University of California, Los Angeles (90 female; 35 male, age range = 18–31 years, Mean age = 20.88, SD = 2.71) who were randomly assigned to receive intranasal OT (n = 42; 30 female, 12 male), AVP (n = 42; 30 female, 12 male), or PLA (n = 41; 30 female, 11 male). The number of overall participants fluctuated slightly depending on the task following removal of outliers and, for some tasks, data lost due to computer error (see Supplementary Material for complete details).

Exclusion criteria included current or history of medical illness, current psychiatric diagnosis, current use of medications (e.g., SSRIs), pregnancy, breastfeeding, and smoking > 15 cigarettes per day (for further details see Tabak et al., 2015 and Supplementary Fig. 1). Participants were asked to refrain from using all medication (e.g., Advil) or alcohol for 24 h, caffeine for 4 h, and food or drinks (except water) for 2 h preceding the experiment. Participants self-identified as Asian (58.2%), White (19.4%), Hispanic (12.2%), Black or African American (5.1%), and "Other" (5.1%). Participants who completed all aspects of

Table 1
NHST main effects *t*-test results from all OT and AVP tasks.

Study	Analysis	Treatment Condition	NHST			Smallest detectable effect size (80% power)
			T-Score	P-value	Cohen's d	
Empathic Concern	Empathy for distressing video	OT	−0.2210	0.8260	−0.0530	0.6200
	Empathy for uplifting video		0.1830	0.8550	0.0440	
Working memory	Nonsocial memory: high – low load	OT	0.4520	0.6530	0.1130	0.7000
	Non-social memory: moderate – low load		1.9520	0.0553	0.4750	0.6900
	Social memory: high – low load		0.5760	0.5670	0.1390	0.6800
	Social memory: moderate – low load		0.6400	0.5250	0.1510	0.7000
Deception	Detection Accuracy	OT	−0.6870	0.4940	−0.1580	0.6300
Personal Distance	Change in perceived threat for far vs. near faces – males	OT	0.4770	0.6350	0.1280	0.7200
	Change in trust for far vs. near faces – males		0.9330	0.3540	0.2350	0.7100
	Change in perceived threat for far vs. near faces – females		−0.2970	0.7680	−0.0710	0.7100
	Change in trust for far vs. near faces – females		−0.6710	0.5050	−0.1690	0.7300
Bystander Effect	Change in intent to help or confront	OT	0.0428	0.9660	0.0130	0.6800
	Change in angry & hostile feelings		−1.6450	0.1040	−0.3630	0.6300
	Change in personal distress		−0.5740	0.5670	−0.1290	0.6300
	Change in empathic concern		0.3990	0.6910	0.0920	0.6600
Essay	Empathy for person supported	OT	−0.4170	0.6780	−0.0900	0.6200
	Empathy & perspective taking for person argued with		−0.2580	0.7970	−0.0590	0.6200
	Negatively affected by argument		−0.8060	0.4220	−0.1710	0.6200
Empathic Concern	Empathy for distressing video	AVP	−0.9960	0.3220	−0.2160	0.6200
	Empathy for uplifting video		−0.7280	0.4680	−0.1640	
Working memory	Nonsocial memory: high – low load	AVP	−0.0874	0.9310	−0.0270	0.7300
	Non-social memory: moderate – low load		0.0000	1.0000	0.0030	
	Social memory: high – low load		0.1440	0.8860	0.0320	0.7600
	Social memory: moderate – low load		0.9490	0.3470	0.2620	0.8000
Deception	Detection Accuracy	AVP	−1.539	0.128	−0.3460	0.6300
Personal Distance	Change in perceived threat for far vs. near faces – males	AVP	−0.5610	0.5770	−0.1390	0.6900
	Change in trust for far vs. near faces – males		−0.1600	0.8730	−0.0360	0.6800
	Change in perceived threat for far vs. near faces – females		0.9330	0.3540	0.2260	
	Change in trust for far vs. near faces – females		2.0230	0.0472	0.4910	0.6900
Bystander	Change in intent to help or confront	AVP	1.2770	0.2050	0.2970	0.6500
	Change in angry & hostile feelings		0.7680	0.4450	0.1720	0.6300
	Change in personal distress		−0.6200	0.5370	−0.1410	0.6300
	Change in empathic concern		2.0520	0.0443	0.4990	0.7000
Essay	Empathy for person supported	AVP	−0.2180	0.8280	−0.0460	0.6300
	Empathy & perspective taking for person argued with		0.5660	0.5730	0.1270	0.6300
	Negatively affected by argument		−0.9770	0.3310	−0.2090	0.6200

Note. OT = oxytocin and AVP = vasopressin.

the study were paid $40-$50. Informed consent was obtained from all participants and the UCLA Institutional Review Board approved this study.

2.2. Procedure

Participants arrived in groups of 2–15 at a computer lab where they each had their own computer terminal. Sessions were conducted between 2:00 − 5:30pm. Participants first completed a set of questionnaires pre-administration including measures of positive/negative affect and state anxiety (described below) and also provided a urine sample, which was tested for drug use and possible pregnancy. Research nurses then checked all participants' temperature, heart rate, and blood pressure to ensure that they were within acceptable limits: systolic blood pressure: 90–130, diastolic blood pressure: 60–90, heart rate: 55–100 BPM, and temperature < 100° F. If vital signs were out of range, participants rested for 10–15 minutes and measurements were repeated until readings were within acceptable limits; one participant was excluded on basis of abnormal vital signs (i.e., after waiting and re-testing, they did not fall within the pre-established ranges) and did not receive OT, AVP, or PLA.

In preparation for each drug-administration session, a third-party research coordinator unrelated to the study used an online random number generator (www.random.org) to randomly assign participants to the OT, AVP, or PLA condition (blocked on gender) and communicated this information to the UCLA pharmacy. A UCLA pharmacist prepared the drug or PLA for each participant with no indication on the label as to its contents (to maintain the blind). Approximately one hour after arriving, participants received OT, AVP, or PLA using a randomized, double-blind, placebo-controlled, between-subjects procedure. We used sterile 6 ml amber glass bottles with metered nasal pumps from Advantage Pharmaceuticals, Inc. Participants first received instructions on how to use the nasal sprays from the first author and a UCLA research nurse. Participants were then instructed to deliver one spray per nostril in an alternating fashion when prompted (every 30 s).

OT (Syntocinon) was provided by Novartis Pharmaceuticals, Switzerland. OT (24 IU/ml) was transferred into the bottles with attached intranasal applicators (1 puff = 0.1 ml). Participants self-administered 5 puffs per nostril (2.4 IU/puff) for a total dose of 24 IU. AVP was provided by American Regent Laboratories, Shirley, NY, USA. The pharmacist transferred AVP (20 IU/ml) into the bottles with attached intranasal applicators (1 puff = 0.1 ml). Participants self-administered 5 puffs per nostril (2 IU/puff) for a total dose of 20 IU. PLA consisted of 2mls glycerine and 3mls purified water (methylparaben and propylparaben mixed according to purified water formula) for a total of 5 ml. This was filtered with a 5 mu filter and transferred to the bottles with attached intranasal applicators (1 puff = .1 ml). Participants self-administered 5 puffs per nostril.

Table 2
Equivalence test results from all OT and AVP tasks.

Study	Analysis	Treatment Condition	Equivalence Test		
			Score	Bonferroni P-value	Holm P-value
Empathic Concern	Empathy for distressing video	OT	2.6030	**0.0055**	**0.0099**
	Empathy for uplifting video		−2.6410	**0.0050**	**0.0099**
Working memory	Nonsocial memory: high – low load	OT	−2.3910	**0.0099**	**0.0395**
	Non-social memory: moderate – low load		−0.8720	0.1930	0.1930
	Social memory: high – low load		−2.2680	0.0133	**0.0399**
	Social memory: moderate – low load		−2.2210	0.0151	**0.0399**
Deception	Detection Accuracy	OT	2.1470	**0.0174**	**0.0174**
Personal Distance	Change in perceived threat for far vs. near faces – males	OT	−2.3560	**0.0109**	**0.0327**
	Change in trust for far vs. near faces – males		−1.9050	0.0307	0.3070
	Change in perceived threat for far vs. near faces – females		2.5650	**0.0064**	**0.0254**
	Change in trust for far vs. near faces – females		2.1760	0.0168	**0.0336**
Bystander Effect	Change in intent to help or confront	OT	−2.7970	**0.0034**	**0.0134**
	Change in angry & hostile feelings		1.1880	0.1190	0.1190
	Change in personal distress		2.2590	0.0133	**0.0266**
	Change in empathic concern		−2.4360	**0.0087**	**0.0260**
Essay	Empathy for person supported	OT	2.4080	**0.0092**	**0.0183**
	Empathy & perspective taking for person argued with		2.5660	**0.0061**	**0.0182**
	Negatively affected by argument		2.0180	0.0235	**0.0235**
Empathic Concern	Empathy for distressing video	AVP	1.8280	0.0356	**0.0392**
	Empathy for uplifting video		2.0960	**0.0196**	**0.0392**
Working memory	Nonsocial memory: high – low load	AVP	2.7730	**0.0037**	**0.0116**
	Non-social memory: moderate – low load		−2.8610	**0.0029**	**0.0116**
	Social memory: high – low load		−2.7230	**0.0043**	**0.0116**
	Social memory: moderate – low load		−1.9170	0.0304	**0.0304**
Deception	Detection Accuracy	AVP	1.2780	0.1020	0.1020
Personal Distance	Change in perceived threat for far vs. near faces – males	AVP	2.2790	0.0129	**0.0387**
	Change in trust for far vs. near faces – males		2.6980	**0.0044**	**0.0175**
	Change in perceived threat for far vs. near faces – females		−1.9240	0.0292	0.0584
	Change in trust for far vs. near faces – females		−0.8170	0.2080	0.2080
Bystander	Change in intent to help or confront	AVP	−1.5470	0.0630	0.1260
	Change in angry & hostile feelings		−2.0650	0.0211	0.0633
	Change in personal distress		2.2130	0.0149	0.0596
	Change in empathic concern		−0.7920	0.2160	0.2160
Essay	Empathy for person supported	AVP	2.6340	**0.0051**	**0.0152**
	Empathy & perspective taking for person argued with		−2.2240	**0.0145**	**0.0290**
	Negatively affected by argument		1.8920	0.0310	**0.0310**

Note. Statistics presented use Bonferroni and Holm methods of correction for multiple correction. Significant *p*-values (in bold) indicate data sensitive for detecting a null result.

As in previous research (Rilling et al., 2012), following completion of administration, participants waited approximately 40 min before beginning the tasks. During this time, participants were asked to sit quietly and read from a stack of 10 magazines (e.g., Newsweek). They were also instructed to turn off their phones and refrain from speaking to one another. Participants then completed measures of positive/negative affect and state anxiety. All tasks were presented in randomized order to minimize potential order effects. Study personnel and research nurses were blind to the drug condition (there were participants in each condition in each group session). The first author supervised the procedure and was present throughout every session. State positive and negative affect, along with state anxiety was measured and did not differ pre- and post-administration (see Supplemental Materials for further details).

2.3. Task designs

Our six tasks used the following stimuli and dependent variables. Please see the Supplementary Materials for further details and descriptions of the null hypothesis statistical tests for each task. As described in Tabak et al. (2015), in task one examining empathic concern, participants responded to two videos in randomized order provided by Sze et al. (2012): an "uplifting" video that depicted non-profit workers teaching children with autism how to swim and surf, and a "distressing" video that gave an overview of the crisis in Darfur while presenting upsetting images from the conflict. Following each video, participants rated the extent to which they experienced empathic concern as reflected by 5-point Likert scale ratings ranging from "not at all" to "extremely" for how moved, sympathetic, and compassionate they felt after viewing the videos. In the second task (described in Tabak et al., 2016), participants completed a social and non-social working memory paradigm (Meyer et al., 2012). Several weeks prior to the lab portion of the task, participants completed a trait-rating questionnaire regarding 10 close friends in which they were asked to rate the extent to which each friend possessed several traits (presented one at a time) on a 1–100 scale (1 = the least, 100 = the most). During the lab portion of the task, following administration of drug or placebo, participants viewed a slide with two, three, or four names of their friends (i.e., to vary cognitive load during the 'encoding' phase, 4 s) which was followed by an adjective (e.g., funny; 1.5 s), and then a delay period (6 s) in which they were asked to mentally rank the friends presented from most to least on the given adjective (e.g., most to least funny). Participants were then given a true/false question that asked whether or not a specific name in the list was in a certain position in the ranked list (e.g., in a list containing the names Claire, Kristin, Rebecca, the true/false question would read: "second funniest?—Rebecca"). The responses were compared to the initial rank order list given weeks earlier to determine the accuracy of the participant's response. In the non-social working memory condition, everything from the social task was held constant except that participants were asked to rank their friends' names in alphabetical order.

In the third task, participants completed a paradigm assessing deception detection accuracy. Participants viewed 16 videos of truthful (n = 8) or deceptive (n = 8) investment pitches in randomized order

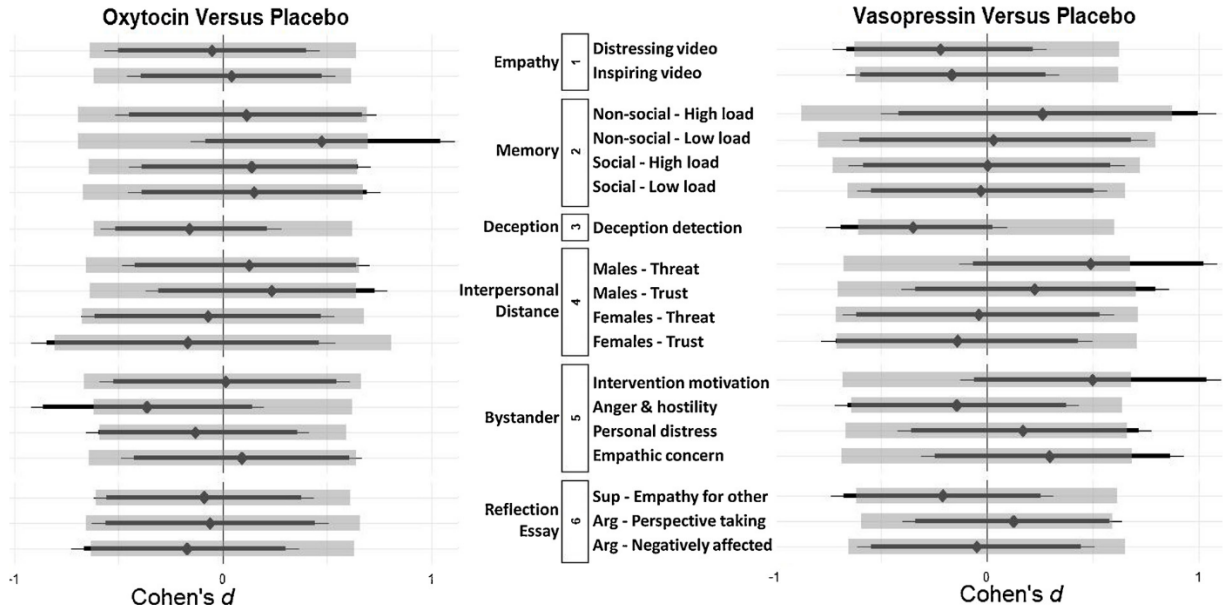

Fig. 1. Visualized equivalence tests are shown for each of the 18 comparisons, respectively for OT and AVP versus PLA, across the six tasks. For each comparison, the Cohen's *d* (diamonds), confidence intervals for mean difference (MD interval; thick lines), effect size confidence intervals (ES interval; thin lines), and raw score equivalence bounds (grey rectangles) are shown. Both confidence intervals were Bonferroni corrected (see Table 1 for corrected alpha levels). If an ES interval does not include zero, the NHST test is statistically significant. If an MD interval lies within the equivalence bounds and ES interval includes zero then the test supports the null hypothesis. If an ES interval includes zero and a MD interval crosses one of its equivalence bounds, then the test is insensitive for any decision. Adapted from Quintana (2018).

and indicated after each whether they thought the person was telling the truth. In task four, participants rated 18 (9 female) grayscale facial images (Bryan et al., 2012) manipulated to appear physically closer or farther away on three different traits: trustworthiness, threatening, and attractiveness (included only for statistical control purposes). Outcome variables were computed to assess the extent to which differences in interpersonal distance influenced perceptions of trustworthiness and threat. In task five we examined the potential influence of OT or AVP administration on bystander helping. Participants read two vignettes from Levine and Crowther (2008) describing a physical assault that varied only in terms of how many bystanders were present. They then rated how they felt based on adjectives from which we derived three composite scores reflecting empathic concern, personal distress, and anger and hostility. In the sixth task, participants were asked to complete a writing task in which they reflected on a supportive interaction and an argument or conflict with a close other. Independent coders then rated the content of the passages on several rating scales assessing the extent to which participants expressed empathy in the supportive and conflict passages, and lingering negative feelings about the conflict.

2.4. Statistical analysis

A description of each study, the variables included, task-specific statistical analyses, including omnibus tests that include both study drugs and PLA, as well as tests examining OT vs. PLA and AVP vs. PLA separately can be found in the Supplementary Materials. Here we report on results from equivalence testing and Bayesian hypothesis testing after determining that all main effects were non-significant using NHST. JASP (v0.8.5.1) was used for NHST and Bayesian hypothesis testing as in Quintana and Williams (2018). All tests were performed as two-tailed tests of z-scored mean responses for either independent or paired samples.

Equivalence testing was conducted in RStudio (version 1.1.453) using the TOSTER package (version 0.2.3; Lakens, 2017). To provide lower and upper bounds for equivalence tests, the "pwr" R package (version 1.2.2; Champely, 2018) was used to determine the smallest effect size detectable at 80% power in each study as in Quintana (2018). The TOSTER package implements the TOST procedure (Schuirmann, 1987) of equivalence testing in which the lower and upper bounds for effect size quantities, Cohen's *d* in our case, are specified to represent the range within which potential effect sizes are considered equivalent to a non-meaningful result.

In each Bayesian *t*-test, we compared a model representing different means for the treatment condition versus PLA with a null hypothesis model for which the effect sizes of each group are equal. Bayesian inferences were made by calculating the Bayes factor (BF), which is the ratio of the posterior odds of the alternative and null hypothesis to its prior odds (Jeffreys, 1961; Kass and Raftery, 1995). When the alternative and null hypotheses are equally probable a priori, the Bayes factor is equal to the posterior odds in favor of the alternative hypothesis. In a Bayesian *t*-test, this is calculated as the ratio of the marginal likelihoods of the data under the alternative and null hypothesis given an experimenter selected distribution of prior odds. The BF product of this ratio is interpreted as how many times more likely the alternative or null hypothesis is over the other model.

For the purpose of adapting BF calculation as an alternative to the NHST *t*-test, Rouder and colleagues (Rouder et al., 2009) combined a Cauchy distribution of effect sizes based on an inverse chi-square function proposed by (Zellner and Siow, 1980) and the Jeffreys distribution of variances (Jeffreys, 1961). This default prior distribution (named Jeffreys' prior) had an interquartile range of $r = 1$ indicative of having 50% confidence in the true effect size being between -1 and 1. However, the weight this Cauchy distribution assigns to large effect sizes has been viewed by some as likely biasing the statistical decisions of behavioral psychology studies biased toward the null hypothesis given the relatively modest effect sizes typically observed in that field (Wagenmakers et al., 2018). Thus, in addition to employing Jeffreys' prior, we also conduct Bayesian t-tests using the $r = 0.707$ distribution that is the default prior in JASP.

Table 3
Bayesian *t*-test results from all OT and AVP tasks.

Study	Analysis	Treatment Condition	Bayes Factor	
			r = 0.707	r = 1
Empathic Concern	Empathy for distressing video	OT	**4.2620**	**5.8020**
	Empathy for uplifting video		**4.2980**	**5.8540**
Working memory	Nonsocial memory: high – low load	OT	**3.6250**	**4.8750**
	Non-social memory: moderate – low load		0.8120	1.0130
	Social memory: high – low load		**3.5140**	**4.7170**
	Social memory: moderate – low load		**3.4340**	**4.6010**
Deception	Detection Accuracy	OT	**3.4740**	**4.6650**
Personal Distance	Change in perceived threat for far vs. near faces – males	OT	**3.4690**	**4.6470**
	Change in trust for far vs. near faces – males		2.6960	**3.5440**
	Change in perceived threat for far vs. near faces – females		**3.8040**	**5.1340**
	Change in trust for far vs. near faces – females		**3.1910**	**4.2460**
Bystander Effect	Change in intent to help or confront	OT	**4.0540**	**5.4990**
	Change in angry & hostile feelings		1.3660	1.7330
	Change in personal distress		**3.7340**	**5.0380**
	Change in empathic concern		**3.8840**	**5.2530**
Essay	Empathy for person supported	OT	**4.0600**	**5.5100**
	Empathy & perspective taking for person argued with		**4.2350**	**5.7630**
	Negatively affected by argument		**3.3580**	**4.5010**
Empathic Concern	Empathy for distressing video	AVP	2.8620	**3.7990**
	Empathy for uplifting video		**3.4290**	**4.6020**
Working memory	Nonsocial memory: high – low load	AVP	2.4870	**3.2330**
	Non-social memory: moderate – low load		**3.7060**	**4.9910**
	Social memory: high – low load		**3.8480**	**5.1980**
	Social memory: moderate – low load		**3.8300**	**5.1710**
Deception	Detection Accuracy	AVP	2.7750	1.9510
Personal Distance	Change in perceived threat for far vs. near faces – males	AVP	0.7190	1.1510
	Change in trust for far vs. near faces – males		**4.0420**	**3.6630**
	Change in perceived threat for far vs. near faces – females		**3.4850**	**5.4820**
	Change in trust for far vs. near faces – females		1.5290	**4.6740**
Bystander	Change in intent to help or confront	AVP	0.7080	1.1730
	Change in angry & hostile feelings		**3.6280**	**4.8860**
	Change in personal distress		**3.3360**	**4.4680**
	Change in empathic concern		2.0510	2.6580
Essay	Empathy for person supported	AVP	2.9450	**3.9150**
	Empathy & perspective taking for person argued with		**3.7520**	**5.0640**
	Negatively affected by argument		**4.2680**	**5.8100**

Note. Statistics presented using either JASP's default Cauchy distribution (r = .707) or the more standard Jeffreys' prior distribution (r = 1). Significant p-values (in bold) are those that are BF = 3 or higher and indicate moderate evidence in support of the null hypothesis.

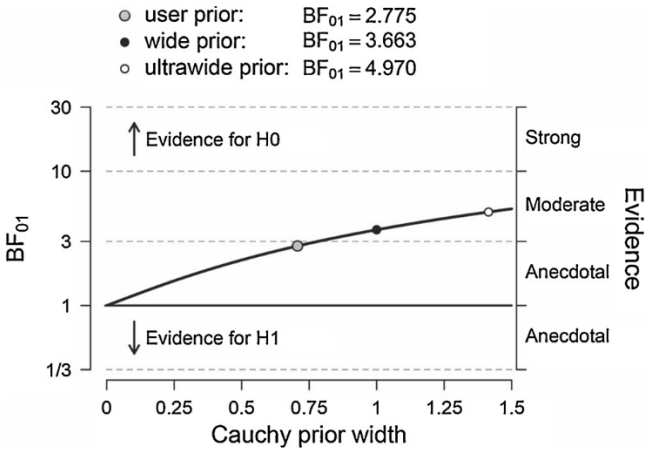

Fig. 2. Graph showing how the width (r value, shown on the X axis) of different prior distributions influences Bayes Factor scores (BF, shown on Y axis) from the Bayesian *t*-test results. Here the "user prior" is JASP's default Cauchy distribution (r = .707) yields relative evidence in favor of the null hypothesis that is below BF = 3, while the more standard Jeffreys' prior (r = 1) is greater than three, indicating evidence analogous to a moderate effect. Data shown is from the interpersonal distance task in which we tested the effects of receiving AVP versus PLA nasal sprays on the change in female participants' trustworthiness judgments between faces that appeared nearby and far away.

3. Results

3.1. Null hypothesis significance testing

As shown in Table 1 and detailed in the Supplementary Material, across all experimental tasks, main effects of all dependent variables were non-significant using NHST. Specifically, there were no effects of OT or AVP (vs. PLA) on empathic concern, social versus non-social working memory, deception detection, interpersonal distance, hypothetical bystander intervention, or reflections of interpersonal interactions.

3.2. Equivalence testing

Equivalence testing was then conducted in conjunction with NHST. Given that all t-tests were non-significant (i.e. the mean difference between conditions was not statistically different from zero), we followed these NHST tests with equivalence tests to assess whether each individual test was sensitive enough to reject effects large enough to be considered worthwhile (as defined by the smallest detectable effect size) and determine whether the conditions compared were statistically equivalent. Both NHST and equivalence tests were corrected for multiple comparisons across the set of 18 unique variables in separate analyses for OT vs. PLA and AVP vs. PLA. The smallest effect size detectable with 80% power was calculated using the pwr package in R for each study as shown in Table 1. The average smallest effect size detectable across all 36 main effects tests conducted in the study,

regardless of the neuropeptide, was $d = 0.67$ and ranged from $d = 0.62$ to $d = 0.8$. Using the average effect size as an example, an equivalence test with bounds of .67 means we can only reject the presence of effects that are as large or larger than .67. If the "true" effect is smaller than .67, which is likely for this area of research, then an equivalence test would not be sensitive enough to reject the presence of this effect. When using the most conservative form of multiple test correction (i.e., Bonferroni), results of equivalence testing demonstrated that 47.22%, or 17 of 36 of all null results were due to statistical equivalence (see Table 2), while 52.78% lacked sensitivity. Using a less conservative Holm correction for multiple testing, 72.22%, or 26 of 36 of all null results were due to statistical equivalence which resulted in only 27.78% of analyses that lacked sensitivity.

Fig. 1 visualizes the Bonferroni corrected results separately for the OT and AVP treatment conditions. These results show that OT vs. PLA resulted in 10 of 18 (55.56%) null results that had the sensitivity to reveal statistical equivalence, whereas AVP vs. PLA resulted in 7 of 18 (38.89%) null results demonstrating sensitivity. Using Holm correction for multiple testing, this changed to 15 of 18 (83.33%) for OT vs. PLA and 11 of 18 (61.11%) AVP vs. PLA.

3.3. Bayesian hypothesis testing

As expected, Bayesian hypothesis testing of our 36 non-significant NHST results found no support for the alternative hypothesis in any case. As shown in Table 3, when comparing the results of the equivalence testing with the Bayesian hypothesis testing, results varied greatly based on the prior distribution supplied. Using the $r = 0.707$ JASP prior, 25 tests (69.44%) reached BF = 3, which is typically considered "moderate" evidence (i.e., nearly all the tests provided at least three times more evidence for the null hypothesis relative to the alternative hypothesis). Results using Jeffreys' $r = 1$ prior were the most extreme of all approaches (see Fig. 2 for a graphic comparison of the influence of prior choice on a comparison) with 30 tests (83.33%) reaching BF = 3. The BFs corresponding to Jeffreys' $r = 1$ and the $r = 0.707$ default prior in JASP are provided in Table 3 for all dependent variables.

4. Discussion

We presented six tasks that investigated the main effects of OT or AVP on several social processes represented by 18 dependent variables. Using NHST, no main effects were found for all analyses. After conducting equivalence testing and Bayesian hypothesis testing on all tests, we found that approximately 47–83% of our analyses provided evidence for the null model relative to the alternative model. Our results add to increasing evidence that intranasal OT and AVP may have a more limited direct effect on human social processes than previously assumed. The present results also highlight the importance of conducting well-powered studies (Lane et al., 2015; Walum et al., 2016). Indeed, while 47–83% of our analyses had adequate sensitivity to detect statistical equivalence, 17–53% of our analyses lacked the sensitivity necessary to have confidence in our findings.

Our equivalence testing results are similar to those reported by Quintana (2018) for the Lane et al. (2015) study set when comparing the average smallest detectable effect size ($d = 0.67$ in the present study and $d = 0.75$ for Lane et al. (2015), but our data did have a lower proportion of tests deemed to be insensitive (i.e. mean differences were not significantly different from zero but statistically equivalence could not be determined) due to lack of power (52.78% for our work and 73.5% for Lane et al., 2015). Regardless, the figure was 52.78%, suggesting that other labs' unpublished null findings from between-subjects designs with similar sample-sizes are likely from data in which around 50% are statistically insensitive.

On the other hand, our data being frequently on the margin of significance for statistical equivalence also impacted the Bayesian results as it meant that the choice for the size of the r value was in many cases influential if we are to adopt the convention of considering a BF value of 3 as moderate evidence. Jeffreys' $r = 1$ prior has been criticized as being too sensitive to the null hypothesis by granting too much weight to potential effect sizes that are unlikely (Wagenmakers et al., 2018). Indeed, in the present study Jeffreys' $r = 1$ prior suggested that nearly all tests provided moderate evidence of in favor of the null hypothesis.

We followed Quintana (2018) in using Bonferroni correction for multiple comparisons, which is known to be a conservative approach. This is noteworthy because several tests trended close enough to significance that any other correction method would have changed the decision. For instance, using Holm correction instead, another of the more conservative correction methods, shows 72.22% of the equivalence test to be significant, a rate nearly identical to the 69.44% of tests above BF = 3 using the $r = 0.707$ default prior in JASP. This correspondence across testing methods affords some confidence in the true number of tests demonstrating evidence of null results across these tests. However, this should be qualified by the facts that the smallest detectable effect sizes across the NHSTs were rather large and the BFs from the Bayesian tests remain at the low end of the range considered to represent a moderate evidence. Specifically, the largest BF using a .707 Cauchy was 4.27, which suggests the null model was only 4.27 times more favored than the alternative model in our data, demonstrating modest evidence. In that context, the 69.44% rate may be the high end of what is reasonable, while the Bonferroni corrected equivalence tests figure of 47.22% may represent the conservative end of the number of tests sensitive enough to reveal statistical equivalence. In either case, the advantage of the BF metric as an indicator of the relative degree of evidence for a model, rather than the discrete decision between statistical equivalence and data insensitivity obtained via p-values means that the majority of reported results offered *some* support for the null hypothesis.

Overall, we found that inputting reasonable parameters into equivalence tests and Bayesian hypothesis t-tests allowed us to evaluate the lack of main effects found in four previously unpublished tasks, along with two published tasks, for the quality of their support for the null hypothesis. These accessible techniques may aid others in gaining better understanding of unpublished research. Conversely, many published studies finding main effects of either drug may actually lack the sensitivity to detect the effect. In fields such as the study of the social effects of OT and AVP that have included many under-powered studies due to logistical hurdles (e.g., where to obtain the drugs, the cost of the study, etc.), additional analyses such as these can provide better conclusions and guide hypothesis generation.

The present study has several strengths including the use of equivalence testing and Bayesian hypothesis testing in addition to NHST. In addition, the fact that all tasks were conducted in a randomized order during the same session with the same participants helps to minimize outside factors that could influence the results. Limitations include a relatively small sample size, a predominantly female sample (a ratio of approximately 3:1), the inclusion of a healthy (non-clinical) sample, and the use of only one dose of OT and AVP. In addition, the present results reflect varied timing of the tasks relative to the doses and the specific incubation period selected. The interaction between some or all of these factors may not be optimal to find main effects (Quintana and Woolley, 2016). Moving forward, well-powered studies including variable dosages and incubation periods will help to disentangle the potential main effects of these neuropeptides.

It should be noted that post-hoc null results that are not preregistered (as in the case of the present manuscript) may be subject to researcher bias (e.g. reverse p-hacking; see Chuard et al., 2019). Moreover, in advocating that principal investigators open their "file drawers" of unreported (and potentially not preregistered) findings, we are aware that we are making the unorthodox recommendation for post-hoc analyses to help advance hypothesis generation and increase transparency between labs. As such, we wish to acknowledge that

testing parameters such as equivalence bounds and prior distributions are often most effectively determined when specified a priori and emphasize that we make this recommendation not out of general advocacy for post-hoc analysis, but out of particular consideration for the amount of uncertainty that exists in human neuropeptide research, given concern about replication in the field of psychology. The field would be best served if researchers were willing to preregister post-hoc analyses of "file-drawer" null results and if journals gave full consideration in the evaluation and publication of null results.

Future studies should prioritize recruiting appropriate sample sizes, the determination of which may be performed using accessible software, such as G*Power (Faul et al., 2007), and include an equal number of men and women to facilitate analyses of gender-specific effects, which are common in this field. In addition, future studies including OT or AVP administration would benefit from incorporating equivalence testing and/or Bayesian hypothesis testing, as this would improve our understanding of the validity of effects (or null findings) in healthy and clinical populations. Last, it should be noted that our original preregistration for the study (NCT01680718) listed four of the six tasks described in the present manuscript. The final two tasks (i.e., the bystander intervention task and the task assessing written reflections of a supportive interaction or a conflict with a close other) were late additions to the protocol and, unfortunately, were not added to the preregistration as they should have been.

5. Conclusions

In sum, the present findings add to a growing body of work demonstrating a lack of main effects of intranasal OT on certain social processes (Bartz et al., 2011; Keech et al., 2018; Lane et al., 2016). In particular, our study found a consistent pattern of null results across different assessments of empathy in three tasks. In addition, results also provide evidence for a lack of main effects of intranasal AVP (for which there is considerably less published evidence) on certain social processes; however, our analyses determined that these results had less sensitivity relative to OT. Last, we also underscore the importance of conducting well-powered research on OT and AVP and describe two different types of post-hoc tests that can be used to assess the sensitivity of findings.

Funding and disclosures

Funding was provided by the UCLA Jeffrey/Wenzel Term Chair in Behavioral Neuroscience (to N.I.E.). During the original data collection, a postdoctoral fellowship for B.A.T. was supported by the MH15750 training fellowship in Biobehavioral Issues in Mental and Physical Health at the University of California, Los Angeles.

Conflict of interest statement

The authors declare no conflicts of interest.

CRediT authorship contribution statement

Benjamin A. Tabak: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Data curation, Project administration. **Adam R. Teed:** Formal analysis, Writing - original draft, Writing - review & editing. **Elizabeth Castle:** Methodology, Investigation, Writing - review & editing, Data curation. **Janine M. Dutcher:** Methodology, Investigation, Writing - review & editing, Data curation. **Meghan L. Meyer:** Methodology, Investigation, Writing - review & editing, Data curation. **Ronnie Bryan:** Methodology, Data curation. **Michael R. Irwin:** Methodology, Writing - review & editing. **Matthew D. Lieberman:** Conceptualization, Methodology, Writing - review & editing. **Naomi I. Eisenberger:** Conceptualization, Methodology, Writing - review & editing, Funding acquisition.

Acknowledgements

We would like to thank Daniel Quintana for his assistance in using the equivalence testing and Bayesian hypothesis testing software packages.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.psyneuen.2019.04.019.

References

Bartz, J.A., Zaki, J., Bolger, N., Ochsner, K.N., 2011. Social effects of oxytocin in humans: context and person matter. Trends Cogn. Sci. 15, 301–309. https://doi.org/10.1016/j.tics.2011.05.002.

Bryan, R., Perona, P., Adolphs, R., 2012. Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. PLoS One 7. https://doi.org/10.1371/journal.pone.0045301.

Champely, S., 2018. pwr: Basic Functions for Power Analysis (Version 1.2.2) [Computer Software]. Retrieved from f. https://CRAN.R-project.org/package=pwr.

Chuard, P.J.C., Vrtílek, M., Head, M.L., Jennions, M.D., 2019. Evidence That Nonsignificant Results Are Sometimes Preferred: Reverse P-Hacking or Selective Reporting? https://doi.org/10.1371/journal.pbio.3000127.

Cohen, D., Perry, A., Mayseless, N., Kleinmintz, O., Shamay-Tsoory, S.G., 2018. The role of oxytocin in implicit personal space regulation: an fMRI study. Psychoneuroendocrinology 91, 206–215. https://doi.org/10.1016/j.psyneuen.2018.02.036.

Donaldson, Z.R., Young, L.J., 2008. Oxytocin, vasopressin, and the neurogenetics of sociality. Science 322, 900–904.

Faul, F., Erdfelder, E., Lang, A.G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191. https://doi.org/10.3758/BF03193146.

Insel, T.R., 2016. Translating oxytocin neuroscience to the clinic: a national institute of mental health perspective. Biol. Psychiatry 79, 153–154. https://doi.org/10.1016/J.BIOPSYCH.2015.02.002.

Israel, S., Hart, E., Winter, E., 2014. Oxytocin decreases accuracy in the perception of social deception. Psychol. Sci. 25, 293–295. https://doi.org/10.1177/0956797613500794.

Jeffreys, H., 1961. Theory of Probability, 3rd edn. Oxford University Press, Oxford.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Keech, B., Crowe, S., Hocking, D.R., 2018. Intranasal oxytocin, social cognition and neurodevelopmental disorders: a meta-analysis. Psychoneuroendocrinology 87, 9–19. https://doi.org/10.1016/j.psyneuen.2017.09.022.

Lakens, D., 2017. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Personal. Sci. 8, 355–362. https://doi.org/10.1177/1948550617697177.

Lane, A., Mikolajczak, M., Treinen, E., Samson, D., Corneille, O., De Timary, P., Luminet, O., 2015. Failed replication of oxytocin effects on trust: the envelope task case. PLoS One 10, 1–10. https://doi.org/10.1371/journal.pone.0137000.

Lane, A., Luminet, O., Nave, G., Mikolajczak, M., 2016. Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. J. Neuroendocrinol. 28. https://doi.org/10.1111/jne.12384.

Leng, G., Ludwig, M., 2016. Intranasal oxytocin: myths and delusions. Biol. Psychiatry 79, 243–250. https://doi.org/10.1016/j.biopsych.2015.05.003.

Leppanen, J., Ng, K.W., Tchanturia, K., Treasure, J., 2017. Meta-analysis of the effects of intranasal oxytocin on interpretation and expression of emotions. Neurosci. Biobehav. Rev. 78, 125–144. https://doi.org/10.1016/j.neubiorev.2017.04.010.

Leppanen, J., Ng, K.W., Kim, Y.R., Tchanturia, K., Treasure, J., 2018. Meta-analytic review of the effects of a single dose of intranasal oxytocin on threat processing in humans. J. Affect. Disord. 225, 167–179. https://doi.org/10.1016/j.jad.2017.08.041.

Levine, M., Crowther, S., 2008. The responsive bystander: how social group membership and group size can encourage as Well as inhibit bystander intervention. J. Pers. Soc. Psychol. 95, 1429–1439. https://doi.org/10.1037/a0012634.

McCullough, M.E., Churchland, P.S., Mendez, A.J., 2013. Problems with measuring peripheral oxytocin: can the data on oxytocin and human behavior be trusted? Neurosci. Biobehav. Rev. 37, 1485–1492. https://doi.org/10.1016/J.NEUBIOREV.2013.04.018.

Perry, A., Mankuta, D., Shamay-Tsoory, S.G., 2015. OT promotes closer interpersonal distance among highly empathic individuals. Soc. Cogn. Affect. Neurosci. 10, 3–9. https://doi.org/10.1093/scan/nsu017.

Pfundmair, M., Erk, W., Reinelt, A., 2017. "Lie to me"—Oxytocin impairs lie detection between sexes. Psychoneuroendocrinology 84, 135–138. https://doi.org/10.1016/J.PSYNEUEN.2017.07.001.

Quintana, D.S., 2018. Revisiting non-significant effects of intranasal oxytocin using equivalence testing. Psychoneuroendocrinology 87, 127–130. https://doi.org/10.1016/j.psyneuen.2017.10.010.

Quintana, D.S., Williams, D.R., 2018. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. BMC Psychiatry 18, 178. https://doi.org/10.1186/s12888-018-1761-4.

Quintana, D.S., Woolley, J.D., 2016. Intranasal oxytocin mechanisms can be better understood, but its effects on social cognition and behavior are not to be sniffed at. Biol. Psychiatry 79, e49–e50. https://doi.org/10.1016/j.biopsych.2015.06.021.

Rilling, J.K., DeMarco, A.C., Hackett, P.D., Thompson, R., Ditzen, B., Patel, R., Pagnoni, G., 2012. Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. Psychoneuroendocrinology 37, 447–461. https://doi.org/10.1016/j.psyneuen.2011.07.013.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon. Bull. Rev. 16, 225–237. https://doi.org/10.3758/PBR.16.2.225.

Scheele, D., Striepens, N., Güntürkün, O., Deutschländer, S., Maier, W., Kendrick, K.M., Hurlemann, R., 2012. Oxytocin modulates social distance between males and females. J. Neurosci. 32, 16074–16079. https://doi.org/10.1523/JNEUROSCI.2755-12.2012.

Schuirmann, D.J., 1987. A comparison of the Two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinet. Biopharm. 15, 657–680. https://doi.org/10.1007/BF01068419.

Sze, J.A., Gyurak, A., Goodkind, M.S., Levenson, R.W., 2012. Greater emotional empathy and prosocial behavior in late life. Emotion 12, 1129–1140. https://doi.org/10.1037/a0025011.

Szeto, A., McCabe, P.M., Nation, D.A., Tabak, B.A., Rossetti, M.A., McCullough, M.E., Schneiderman, N., Mendez, A.J., 2011. Evaluation of enzyme immunoassay and radioimmunoassay methods for the measurement of plasma oxytocin. Psychosom. Med. 73, 393–400. https://doi.org/10.1097/PSY.0b013e31821df0c2.

Tabak, B.A., Meyer, M.L., Castle, E., Dutcher, J.M., Irwin, M.R., Han, J.H., Lieberman, M.D., Eisenberger, N.I., 2015. Vasopressin, but not oxytocin, increases empathic concern among individuals who received higher levels of paternal warmth: a randomized controlled trial. Psychoneuroendocrinology 51, 253–261. https://doi.org/10.1016/j.psyneuen.2014.10.006.

Tabak, B.A., Meyer, M.L., Dutcher, J.M., Castle, E., Irwin, M.R., Lieberman, M.D., Eisenberger, N.I., 2016. Oxytocin, but not vasopressin, impairs social cognitive ability among individuals with higher levels of social anxiety: a randomized controlled trial. Soc. Cogn. Affect. Neurosci. 11, 1272–1279. https://doi.org/10.1093/scan/nsw041.

Uzefovsky, F., Shalev, I., Israel, S., Knafo, A., Ebstein, R.P., 2012. Vasopressin Selectively Impairs Emotion Recognition in Men. https://doi.org/10.1016/j.psyneuen.2011.07.018.

Valstad, M., Alvares, G.A., Andreassen, O.A., Westlye, L.T., Quintana, D.S., 2016. The relationship between central and peripheral oxytocin concentrations: a systematic review and meta-analysis protocol. Syst. Rev. 5, 49. https://doi.org/10.1186/s13643-016-0225-5.

Vosgerau, J., Simonsohn, U., Nelson, L.D., Simmons, J.P., 2018. Internal meta-analysis makes false-positives easier to produce and harder to correct. SSRN Electron. J. https://doi.org/10.2139/ssrn.3271372.

Wagenmakers, E.J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q.F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J.N., Morey, R.D., 2018. Bayesian inference for psychology. Part II: example applications with JASP. Psychon. Bull. Rev. 25, 58–76. https://doi.org/10.3758/s13423-017-1323-7.

Walum, H., Waldman, I.D., Young, L.J., 2016. Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. Biol. Psychiatry 79, 251–257. https://doi.org/10.1016/j.biopsych.2015.06.016.

Zellner, A., Siow, A., 1980. Posterior odds ratios for selected regression hypotheses. Trab. Estad. Y Investig. Oper. 31, 585–603. https://doi.org/10.1007/BF02888369.