How can intranasal oxytocin research be trusted? A systematic review of the interactive effects of intranasal oxytocin on psychosocial outcomes

Article *in* Perspectives on Psychological Science · February 2020

CITATIONS
0

READS
418

7 authors, including:

Adrien Mierop
Université Catholique de Louvain - UCLouvain
15 PUBLICATIONS **117** CITATIONS
SEE PROFILE

Moïra Mikolajczak
Université Catholique de Louvain - UCLouvain
141 PUBLICATIONS **5,059** CITATIONS
SEE PROFILE

Jérémy Béna
Université Toulouse II - Jean Jaurès
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

How can intranasal oxytocin research be trusted?

A systematic review of the interactive effects of intranasal oxytocin on psychosocial outcomes

Mierop, A.[1], Mikolajczak, M.[1], Stahl, C.[2], Béna, J.[3], Luminet, O.[1, 4], Lane, A.[1], & Corneille, O[1].

This is the preprint version.

Manuscript to be published in "Perspectives on Psychological Science"

Affiliations:

1. UCLouvain, Louvain-la-Neuve, Belgium
2. Universität zu Köln, Köln, Germany
3. CLLE, Université de Toulouse, CNRS, Toulouse, France
4. Belgian Fund for Scientific Research (FRS-FNRS)

Keywords: neuroendocrinology, oxytocin, social cognition, positive predictive value

Running title: Intranasal oxytocin research.

Abstract

Over the past two decades, research about the role of Oxytocin (i.e., OT) in human behavior has grown exponentially. However, research efforts have not yet allowed developing a unified theory of OT effects. Relatedly, growing concerns about the robustness of conclusions drawn in the field have been raised. The present article contributes to this debate by reporting and discussing key conclusions from a systematic review of published studies addressing interactive effects of intranasal-OT (i.e., IN-OT) administration on psycho-social outcomes in a healthy population. The review indicated that (1) tested interactive IN-OT effects were highly heterogeneous; (2) for a majority of published interactions, no replication was attempted; (3) when attempted, replications were largely unsuccessful; (4) significance was unrelated to sample size; (5) statistical power was critically low and unrelated to the rate of significant results; (6) research practices were characteristic of an exploratory approach. This concerning state of affairs makes it virtually impossible to tease apart true from false interactive IN-OT effects. Based on this observation and Positive Predictive Value simulations, we provide constructive directions for future research that should help extract true effects from noise and move the IN-OT field forward.

How can intranasal oxytocin research be trusted?

A systematic review of the interactive effects of intranasal oxytocin on psychosocial outcomes

Over the last two decades or so, a growing body of studies has adopted a mechanistic perspective on human behavior and looked at the endocrinological roots of social behavior in humans. Prominent in this research stream, evidence has accumulated suggesting that Oxytocin (OT) may play a critical role in humans' emotional and social lives. Recently, however, both the reliability of intranasal OT (i.e., IN-OT) effects on psychological outcomes (Walum, Waldman, & Young, 2016) and the very physiological plausibility of such effects (Leng & Ludwig, 2016) have been questioned. The aim of the present contribution is to foster advances in OT research by reporting and discussing key conclusions of a systematic review of published IN-OT studies on psycho-social outcomes. To introduce the research questions examined in this article and their relevance for the field, we first provide the reader with a brief state of the art. We then report and discuss the main outcomes of the review, essentially focused on interactive effects of IN-OT. In the general discussion, we suggest constructive avenues for strengthening the reliability of research published in the field.

OT as a social hormone

A large number of empirical studies have suggested that IN-OT administration promotes a wide range of pro-social behavioral tendencies, such as trusting strangers (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005), boosting perceived trustworthiness and attractiveness (Theodoridou, Rowe, Penton-Voak, & Rogers, 2009), and promoting self-confidence (Cardoso, Ellenbogen, & Linnen, 2012). IN-OT is also hypothesized to foster relationship maintenance and relational efficiency by improving emotion recognition (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007), enhancing the recognition of familiar faces

(Rimmele, Hediger, Heinrichs, & Klaver, 2009), improving constructive communication in couples (Ditzen et al., 2009) and facilitating parent-infant attachment (Naber, van IJzendoorn, Deschamps, van Engeland, & Bakermans-Kranenburg, 2010). These findings helped build OT's reputation as the "prosocial hormone". Several meta-analyses have examined the aggregated effects of oxytocin on emotion recognition (Shahrestani et al., 2013; Leppanen et al., 2017) and threat processing (Leppanen et al., 2018) in healthy subjects, as well as social cognition in patients suffering from neurodevelopmental disorders (Keech et al., 2018) and schizophrenia (Bürkner et al., 2017). These meta-analyses have reported small-to-medium effect sizes.

Recently, however, several findings have tempered down this positive view of IN-OT by suggesting that OT may also facilitate anti-social behaviors, such as defensive aggression (De Dreu et al., 2010), ethnocentrism (De Dreu, Greer, Van Kleef, Shalvi, & Handgraaf, 2011) and gloating (Shamay-Tsoory et al., 2009). This apparent contradiction was accommodated by new theories about the role of OT stating that, rather than promoting affiliation or prosociality *per se*, as originally postulated (Insel, 1992), OT may either enhance the salience of social stimuli (Shamay-Tsoory & Abu-Akel, 2016) or facilitate social approach and inhibit social withdrawal (Kemp & Guastella, 2011). As the field grew, a number of findings that could not be accommodated by existing theories emerged (see e.g., Lane et al., 2015).

Growing Concerns in IN-OT Research

Although research about the role of OT in human behavior has grown exponentially over the past two decades, research efforts and findings have not yet allowed developing a unified theory of IN-OT effects (Bartz, Zaki, Bolger, & Ochsner, 2011). Relatedly, a set of

troublesome findings has elicited growing concerns about the robustness of conclusions drawn in the field.

A first set of concerns has to do with the pharmacodynamics of OT. OT research has relied on the pharmacokinetic properties of arginine vasopressin (AVP) administration (Born et al., 2002). The structure of this peptide hormone is structurally similar, yet not identical to OT (Meyer-Lindenberg, Domes, Kirsch, & Heinrichs, 2011). Small molecular differences, however, may imply large functional differences. For instance, while OT is typically related to prosocial behavior, AVP has primarily been related to social behaviors that are stereotypically related to masculinity (see Heinrichs, von Dawans, & Domes, 2009). Moreover, the efficiency of intranasal administration on the elevation of OT levels in the cerebrospinal fluid (CSF) has been questioned (Lee et al., 2018). Typically, the intranasal administration is probed by measuring OT levels on peripheral samples (e.g., blood, urinary, salivary) because they are far easier to obtain than more invasive CSF samples. Some researchers have argued, however, that peripheral levels of OT do not necessarily reflect cerebral levels (Amico, Challinor, & Cameron, 1990; Kagerbauer et al., 2013), and some have argued that OT might not even cross the blood-brain amount a substantial degree (Landgraf & Neumann, 2004). As a result, IN-OT studies might lack of sufficient signal-to-noise ratio for observing reliable effects.

Carson and colleagues (2015) observed that blood and CSF levels of OT are positively related. Recently, however, Leng and Ludwig (2016) estimated that 0.005% of the intranasal OT solution, at most, makes its way to the CSF. Huge amounts of IN-OT may therefore be required to induce detectable behavioral effects (see for instance, Kirkpatrick, Francis, Lee, De Wit, & Jacob, 2014). Of further concern, Born et al.'s study showed that it takes not less than 60 minutes for 40 International Units (or IU) of AVP to reach significant concentrations in the CSF. It is therefore unclear whether time windows generally used in IN-OT research

(the median time window between INOT administration and the first task is 45 minutes in the studies that are included in this review) are adequate, especially as one study (Striepens et al., 2013) showed that IN-OT yields elevated cerebrospinal fluid OT levels after 75 minutes (not 45) after administration. A recent research suggests that IN-OT effects are most effective from 45 to 70 minutes after administration of 24 IU (Spengler et al., 2017). More research is definitely needed that advance our understanding on the pharmacokinetics of OT.

A second major concern relates to the statistical power of OT studies. Walum et al. (2016) estimated that the average statistical power of human IN-OT studies is only 16%. Based on this finding, the authors concluded that many published IN-OT effects might actually reflect type I errors. Low power may also have hindered the detection of true effects and, in some instances, may have contributed to replication failures. Several attempts to replicate seminal results of the IN-OT literature (e.g., Domes et al., 2007; Kosfeld et al., 2005) have failed, suggesting that effects may not be as robust as one might have thought (Lane et al., 2015; Nave, Camerer, & McCullough, 2015; Radke & de Bruijn, 2012; Tabak et al., 2019). Furthermore, when underpowered studies detect true effects, effect sizes are typically overestimated. As a result, aggregated (small-to-medium) effect sizes that are reported in meta-analyses are likely to be overestimated as well (Pereira & Ioannidis, 2011).

The above-mentioned issues may account for inconsistencies in IN-OT research. Another way to account for them is to consider that the effects of oxytocin may be constrained by features of situations and/or individuals (Bartz and collagues, 2011). This interactionist approach may explain why IN-OT main effects are not systematically observed (see for instance in emotion recognition studies: Bartz et al., 2010; Domes et al., 2007; Guastella et al., 2010; Hurlemann et al., 2010). Consistent with this idea, Bartz et al. (2011) noted that a majority of IN-OT studies reported significant moderations of IN-OT effects.

Accordingly, these authors recommended to shift from the question 'Does oxytocin improve social cognition?' to the question 'Under what circumstances does oxytocin improve social cognition?' This approach suggests that interactive rather than main effects of IN-OT administration are the effects of interest for behavioral researchers.

The Present Review

Consistent with Bartz et al. (2011)'s recommendation, the present review examined interactive effects of IN-OT on psycho-social outcomes. Our objectives were manifold: (1) gathering knowledge about the diversity of tested moderations, (2) providing information on the general pattern of reported results and interpreting them in light of different theoretical scenarios, (3) examining current research practices in the field. Based on this analysis, our overarching goal was to provide constructive recommendations for moving research on IN-OT forward.

Method

Literature Search

The present review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA; Moher, 2009). Studies were sourced from PsychInfo[1] and PubMed[2]. We collected articles in May 2018, and our database includes studies of the past 15 years of research, published between 2003 and April 2018.

[1] With the following syntax : (MAINSUBJECT.EXACT("Oxytocin")) NOT (AB("plasma") OR AB("release") OR AB("gene*") OR AB("receptor*") OR AB("vole*") OR AB("uterine") OR AB("labor") OR AB("parturition") OR AB("autis*"))

[2] With the following syntax : (((((("Oxytocin"[Mesh] OR "Oxytocics"[Mesh]) AND oxytocin[Title/Abstract] AND Humans[Mesh])) NOT (plasma[Title/Abstract] OR haemorrhage[Title/Abstract] OR blood[Title/Abstract] OR release[Title/Abstract] OR gene[Title/Abstract] OR genetic*[Title/Abstract] OR receptor*[Title/Abstract] OR vole*[Title/Abstract] OR uterine[Title/Abstract] OR labor[Title/Abstract] OR parturition[Title/Abstract] OR autis*[Title/Abstract] OR rat[Title/Abstract] OR rats[Title/Abstract] OR mouse*[Title/Abstract] OR mice*[Title/Abstract])))) AND ("2005/01/01"[PDat] : "2019/12/31"[PDat]))

Inclusion Criteria

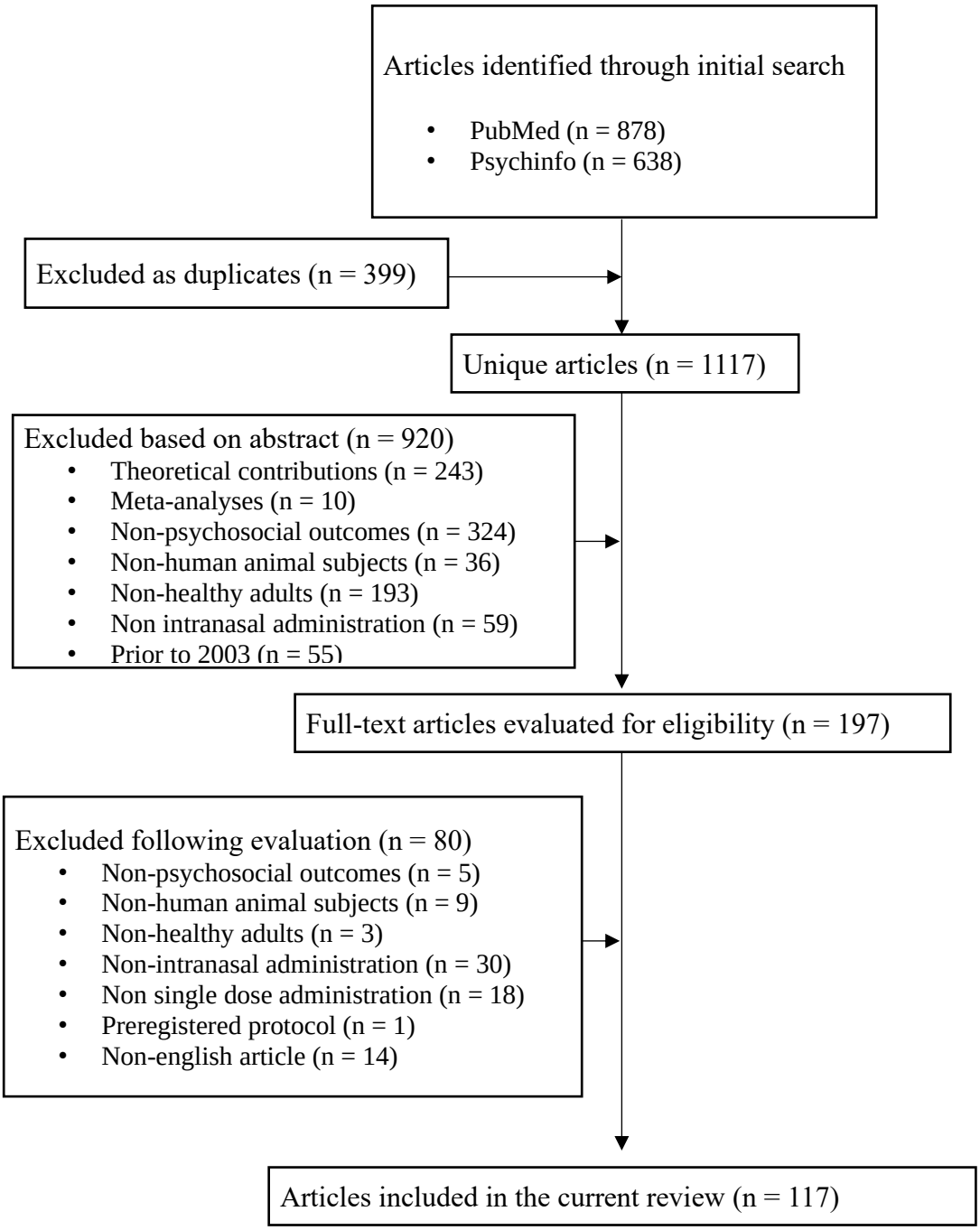
In this review, we focus on the effect of *intranasal* administration of OT on *psycho-social* outcomes measured on *healthy adults*, which represents the most frequently addressed case (see Tabak et al., 2019). Findings and conclusions should therefore not be overgeneralized to other categories of outcomes or to outcomes observed on clinical populations (we will come back to this point in the discussion).

We included studies that met all of the following criteria (see Flowchart):

1) Consisted of an experimental investigation
2) Involved a placebo (PL) controlled IN-OT administration
3) Conducted on healthy volunteers
4) Focused on psychosocial (affective, behavioral or cognitive) variables externally measured (i.e., not via fMRI, EEG, EMG, ECG).
5) Were written in English.

We excluded studies based on the following criteria:

1) Theoretical contributions (e.g. reviews), meta-analysis, case study, qualitative study, animal study.
2) Involved intravenous administration of OT or peripheral (plasmatic/salivary) OT assessments
3) Conducted on a clinical sample (e.g., individuals with schizophrenia)
4) Focused on parturition/breastfeeding (for a review see Wei et al., 2009), physical (e.g. pain) or neuro/physiological variables.
5) Published before 2003.



Flowchart illustrating the literature search process in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Coding Protocol

For each article, we encoded three types of information: (1) Study identifiers (author[s], year of publication, journal, keywords); (2) Study characteristics (quantity in

International Units [I.U.] of the OT dose, OT solution brand, time interval [in minutes] between OT administration and data collection, total sample size, gender of the sample [female, male, or mixed]); (3) Transparent research practices (a priori power analysis, pre-registration, disclosure of full report, open data). Research practices were probed by searching for relevant words (e.g., pre-registration) in the search function of the pdf reader application. In addition, a visual inspection search was realized on the abstract, supplementary materials and method sections.

We encoded all reported interactions that included the IN-OT factor. More specifically, for each interaction, we encoded: the dependent variable assessed, the interaction including the IN-OT administration and the moderator(s), the interaction type (the number of factors presents in the interaction, the quantity of within-participants factors present in the interaction), and the significance of the interaction. The coding protocol and the data are publicly available on OSF (osf.io/zr9yq).

Reliability Check

To assess coding reliability, a randomly selected subset of 20% of the studies was coded by a second independent coder. This allows sufficient precision to estimate a kappa coefficient as small as k = .67 at the effect-level (Gwet, 2014). Selected articles were identified by the author names and publication year. The coding reliability was estimated with the percentage of agreement and the Cohen's kappa coefficient, which is suited to assess accuracy of categorical coding while controlling for chance agreement.

Study-level variables. First, we assessed coding reliability of the study-level variables (i.e., journal, keywords, OT dose, OT manufacturer, interval, sample size, gender, presence of power analysis, preregistration, and disclosure, and data availability). There were 24 matching cases for which study-level agreement was computed. Interrater agreement was good overall

(i.e., the median value of Cohen's kappa was 1). The largest number of disagreements (4 cases; 83 % agreement) were found in sample size; this reflected coding of full sample vs. retained sample (after exclusions); the coding protocol was clarified, and the retained sample size was coded for all studies. Agreement was perfect for 6 out of 11 variables; a maximum of one or two disagreements were found for the remaining 4 variables (agreement between 90% and 96%). Cases of disagreement were solved by discussion.

Effect-level variables. The effect sizes were aligned via an ID variable that was jointly assigned by both coders. Agreement was very good overall (98-99%, median kappa = .95), but with the exceptions that (1) the number of within-person factors had a substantial proportion of disagreements, and (2) there was a substantial proportion of cases in which a given interaction was included by the second coder but not the first (main) coder. This was due to a systematic difference in strategies used to identify statistical tests of interactions (i.e., one coder used a more conservative approach and coded an interaction only when a test statistic was reported; the other coder inferred that an interaction was tested when results were reported in the text). The main coding relied on the more conservative approach, integrating only the reported test statistics. Therefore, our results represent a lower bound estimate of the number of statistical tests that were conducted by researchers.

Analyses and Results

In the following analytic sections, we estimate and report (1) the heterogeneity of the tested interactive effects, (2) the replications attempts and successes, (3) the relation between sample size and rate of significant findings, (4) the statistical power, and (5) the adherence to open-science practices. We then discuss three extreme scenarios of the observed results pattern on the statistical power, the proportion of true effects and the proportion of false

positives as an interpretative background and turn to simulations to provide more nuanced interpretations as well as informed recommendations.

A Large Number of Interactive IN-OT Effects Have Been Tested

A first analysis reports the number and rate of significant interactions, for IN-OT interactions whose statistical outcomes were reported in the reviewed studies. The 117 published articles included in the present review reported 127 studies that tested a total of 829 interactive IN-OT effects (for instance, the IN-OT × gender interaction). This list of interactive IN-OT effects does not contain interactions between IN-OT and Time (i.e., pre vs post IN-OT administration). Conceptually, these interactions reflect main IN-OT effects controlled for baseline and they will be addressed separately. Hence, on average, an IN-OT study reported 6.5 interactive effects involving IN-OT. An average of 26% of the interactive effects were reported as significant, and 91% of the studies reported at least one significant interactive effect involving IN-OT. These proportions are difficult to interpret in and by themselves or to relate to other domains of research. We discuss extreme and more nuanced interpretative scenarios in a later section (i.e., PVV simulations).

Tested Interactions Are Highly Heterogeneous, And The Majority of Interactions Has Been Tested Only Once

A second analysis examined the heterogeneity and rate of conceptual replication for the tested interactive effects of IN-OT. More specifically, we were interested in the diversity of outcomes studied (i.e., the dependent variable, coded at a conceptual level), as well as the combination of variables included in a given interaction effect (i.e., IN-OT x gender, IN-OT x gender x age). On the 127 studies addressed in this review, we identified a large heterogeneity

of interactive effects including the IN-OT factor and a given conceptual dependent variable[3]. A total of 350 different interactions were identified (e.g., IN-OT × gender), and 95 distinct conceptual variables were identified (e.g., "altruism", "lie detection", "empathy", "threat perception", "emotion recognition"). The most frequent categories of response involved were subjective ratings and accuracy scores as well as response times collected in speeded tasks. Combined, 467 unique combinations were tested in the published literature, only less than half of which were conceptually tested more than once (i.e., 210). The tested IN-OT interactions are therefore characterized by a very large heterogeneity and a modest rate of replication attempts (see Figure 1A).

For Repeatedly Tested Interactions, Replication Attempts Are Largely Unsuccessful

Next, we examined the success of conceptual replications in interactive IN-OT effects. As can be seen on Figure 1B, IN-OT research is characterized by a low rate of successful conceptual replications. It should be noted that Figure 1B represents an optimistic estimate of successful replication. This is because (1) significant effects are reported here independent of their direction (such that two significant effects may reflect effects of opposite direction), and (2) the largest value concerns data coming from a single study that replicated the same interaction six times across six conceptually related DVs.

[3] For each dependent variable, the most relevant keyword reported in the article was used as the 'conceptual dependent variable'. When no keyword was provided, the DV was categorized according to the most relevant existing categories. A table containing the different interactions and conceptual dependent variables is available on: osf.io/8jgan

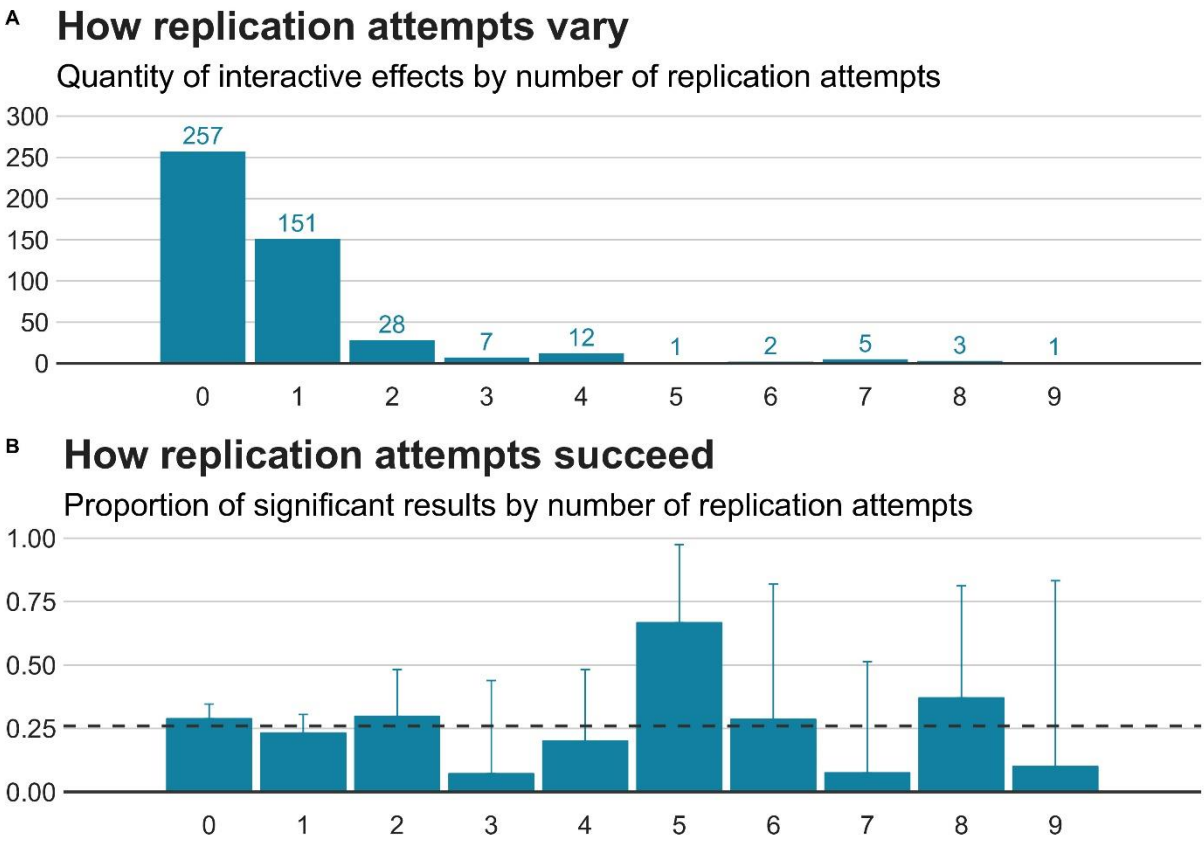


Figure 1. A: Illustration of the frequency of replication attempts. B: Proportion of significant results as a function of the frequency with which a specific moderation was tested. The error bars represent the (upper bound) binomial proportion 95% confidence interval. A majority of moderations were tested only once (k=257) or twice (k=151); only 59 cases (13%) were tested more than twice. The dashed line represents the average proportion of significant results.

The Rate of Significant Results Does Not Increase With The Sample Size

As a further test of the overall reliability of published IN-OT interactive effects, we examined the relation between sample sizes and the proportion of significant results. It is possible that IN-OT effects are true but remain undetected (false negatives) because of a lack of statistical power. If this is the case, studies with larger samples should be associated with a more frequent detection of (true) effects. That is, because larger sample sizes are associated with enhanced statistical power, sample size and rate of significant results should be positively

associated if a substantial portion of non-significant effects are false rather than true negatives.

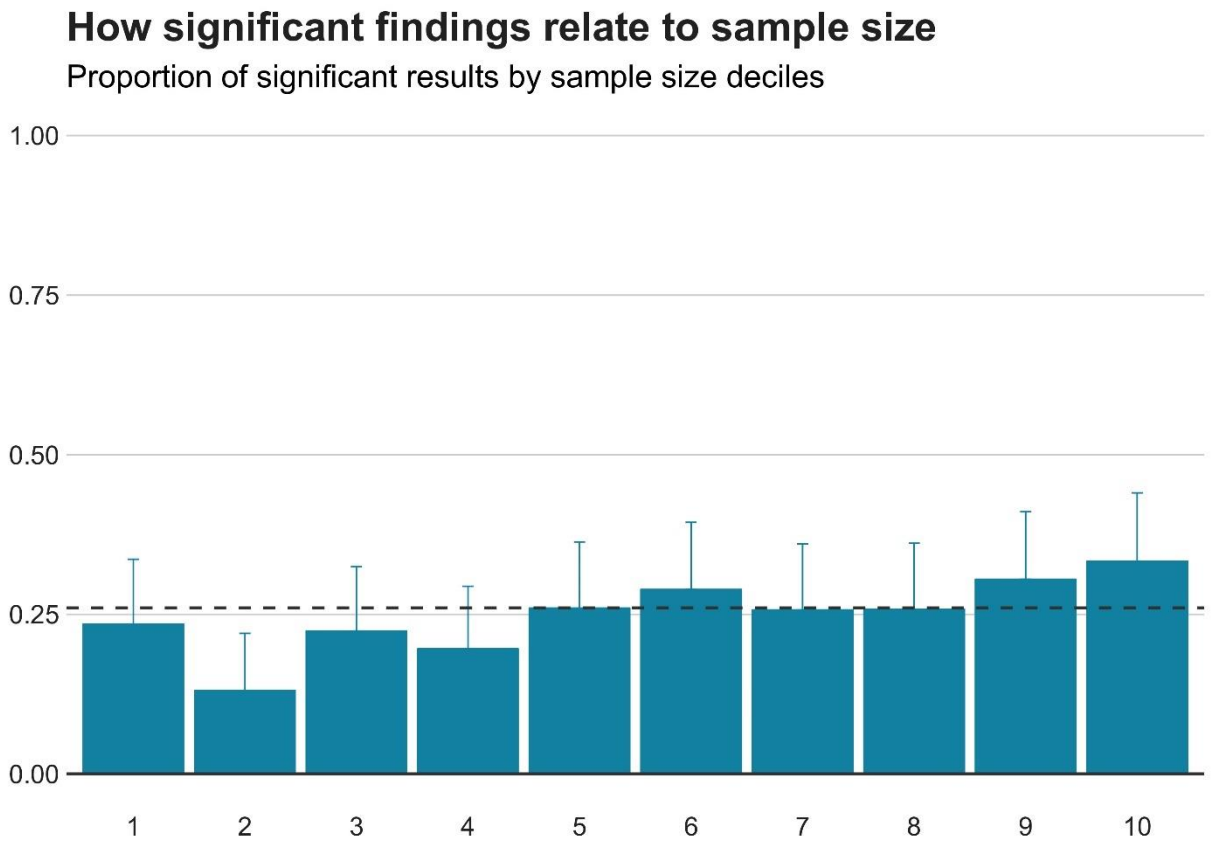


Figure 2. Proportion of significant interactions by deciles of sample size. About 83 interactive effects are represented by decile. The error bars represent the binomial proportion 95% confidence interval. The dashed line represents the average proportion of significant results.

As can be seen on Figure 2, the relation between sample size deciles and proportion of significant results is flat. The fact that, on average, significant effects are not more frequently observed with larger sample sizes could either mean that (1) weaker effects were examined in studies with larger sample sizes (which seems unlikely, considering the general lack of power analyses – see below), or (2) that reported interactive effects generally do not reflect true effects, or (3) that the statistical power of IN-OT studies should be considerably larger than it currently is in order to reliably detect true interactive effects.

Statistical Power Is Low

Even though the present review was primarily concerned with interactive IN-OT effects, we based our considerations about statistical power on the assumption that the study goal was to investigate main effects, in order to obtain an (optimistic) estimate of statistical power. The median sample size was 57 participants, a relatively small sample size that corroborates Walum et al.'s (2016) concerns regarding the low statistical power to detect main IN-OT effects. As displayed in Figure 3B, the sample sizes are generally too small for reliably detecting the main IN-OT effect size estimated by Walum et al. (2016) to be of Cohen's d = .28. And this is true both when considering the IN-OT main effect in between-subject and in within-subject designs. The minimal sample size to detect an effect size of Cohen's d = .28 in a between-subject design with a statistical power of .8 and a two-tailed type I error of .05 is 404. Figure 3C (blue bars) illustrates the statistical power for detecting an effect of d=.28 with alpha=.05 and the distribution of sample sizes, assuming a between-subject design. We can see that none of the studies analyzed in the present review reached sufficient statistical power. Assuming a within-subject design, one would require 103 pairs of observations to achieve such statistical power (assuming d = dz = .28. i.e., a correlation between measures of r = .5). Even if the present studies had all used within-subject designs with the given sample sizes, only 12% of the studies would be sufficiently powered to detect an effect of d=.28. In line with previous findings, the power to detect an effect of d = .28 with the observed median N = 57 was approximately .18 (between-subject) or .55 (within-subject).

The reported IN-OT studies were not sensitive enough to investigate small-to-medium effects, for either main or interactive IN-OT effects. On average, they were capable of detecting only medium-to-large effects. For within-participant studies, the median sample size

yields sufficient power (.8, with alpha = .05) to detect an effect size of dz = .38; for a between-participant study, power suffices for detecting only large effects (d = .76).

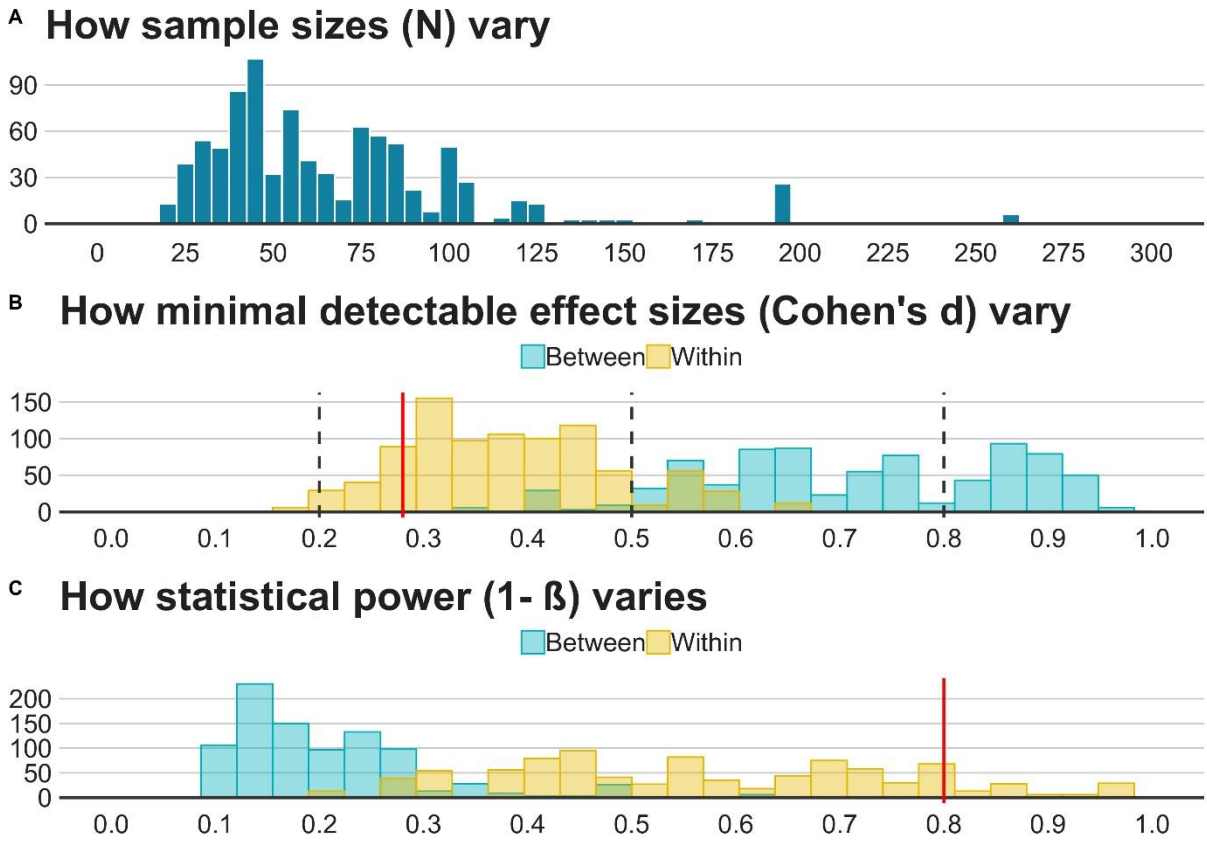


Figure 3. Histograms of the frequency of sample sizes (A), minimal detectable effect sizes (B) and statistical power (C) of the studies. In Figure 3B, the red line depicts the effect size estimated by Walum et al. (2016) (i.e., Cohen's d = .28) and the black dashed line represents the thresholds of small, medium and high effect sizes. In Figure 3C, the red line represents the threshold of satisfactory statistical power of 80%.

Main IN-OT Effects Probed by an IN-OT × Time Interaction

IN-OT × Time interactions can be seen as main IN-OT effects controlled for baseline measures. Hence, this interaction test provides a more controlled estimate of main IN-OT effects. We repeated the analyses reported above for the interactive IN-OT effects with this special interactive case. There were 72 IN-OT × Time interactions on 14 different conceptual

dependent variables. Only 8% of these interactions are reported as significant. The median sample size was N = 40, meaning that the statistical power to detect a main IN-OT effect of d = .28 was only of $(1 - \beta) = .41$ assuming a within-participants design and of $(1 - \beta) = .24$ assuming a between-participants design.

Adherence to Transparent Research Practices Is Uncommon

Finally, we analyzed the reported use of transparent research practices. Our search returned three articles reporting a pre-registration of the experimental design and hypotheses (one article reported a registration on European Clinical Trial Registry, and two on clinicaltrials.gov), and no article containing a disclosure of full reporting. Only 1% of the studies provided public access to the data. Only 10% of the studies reported power analyses. These numbers, possibly underestimating the true state of affairs due to a few identification misses, are consistent with the notion that IN-OT research at this point is largely exploratory, as also suggested by the large set of possible combinations of moderators and conceptual DVs and low rate of replication attempts. Finally, only 4% of the articles reported more than one study using IN-OT administration. Hence, single-paper replications of one's own findings are extremely uncommon in IN-OT research (as compared to the more common practice of reporting multiple replications in a single paper as observed in other fields of research).

How Should We Interpret These Findings?

To sum up, we found that (1) effects tested were highly heterogeneous; (2) for a majority of published interactions, no replication was attempted; (3) when attempted, replications were largely unsuccessful; (4) significance was unrelated to sample size; (5) statistical power was critically low (and unrelated to rate of significant results); (6) research practices were characteristic of an exploratory approach.

In this section, we first discuss three extreme scenarios for interpreting the key outcomes of this review. These should be considered extreme case scenarios. As a consequence, none of these scenarios is likely to perfectly depict the current state of affairs. Instead, they offer useful conceptual backgrounds against which interactive IN-OT effects reported in the published literature can be interpreted. We then turn to Positive Predictive Value simulations allowing (1) for a more nuanced interpretations of the current state of affairs as well as (2) guidance for moving IN-OT research forward. This approach presents the combined benefits of (1) being able to approximate "how can we trust research on interactive effects of oxytocin" by quantifying the proportion of significant results that reflect true effects, (2) relatedly, to highlight if a bias is necessary to reach the pattern of results observed, and (3) to quantify the benefits of adhering to open and cumulative research practices.

Three extreme scenarios as an interpretative background.

The first scenario specifies that 100% of the tested hypotheses were true and that the average statistical power of the studies was 26% (which reflects the rate of significant results in the present data). According to this first scenario, the average effect size should therefore be of Cohen's d = 0.35 (with a between-subject design), which corresponds to a small-to-medium effect size according Cohen's (1988) norms. This first extreme stands in contrast to estimates that only 10% of the tested hypotheses in experimental psychology are likely to be true, especially in exploratory studies (Pashler & Harris, 2012). Furthermore, it is inconsistent with the fact that the sample size was not or only weakly related to the proportion of significant moderations. As discussed above, one should expect more significant effects with larger sample sizes.

The second scenario specifies that 26% of the tested hypotheses were true and that published studies that tested them were maximally (i.e., 100%) powered for detecting a true

effect. According to this second scenario, the average effect size should be of Cohen's d=1.15 (using 99.9% power for computation and with a between-subject design), which is considerably larger than Cohen's (1988) "large effect" norm (d=.8). This implausible scenario stands in contrast with the very low proportion of significant results among the relatively few studies that were conceptually replicated.

If we take the observations at face value and ignore possible biases, then IN-OT effects sizes must be either very small (first scenario, assuming low power), or very large (second scenario, assuming low a-priori probability). Considering the possibility of publication bias, it is more likely that true effects are small, and that mean effect size estimates as well as the rate of significant findings in the literature were overestimated. In this vein, a third extreme scenario attributes the findings entirely to publication bias and specifies that the 26% significant interactions are false positives. This third scenario is consistent with the fact that almost every article reports at least one significant interaction but that those are rarely replicated when a replication attempt is made (see Franco, Malhotra, & Simonovits, 2014; Simmons, Nelson, & Simonsohn, 2011).

Simulations for more nuanced interpretations and ways forward.

In an attempt to outline more realistic scenarios, we used Ioannidis' (2005) approach for calculating the positive predictive value (PPV), which is the proportion of significant findings that actually reflect true effects. The PPV depends on the alpha and beta levels, as well as the rate of true relationships in a field (R) and a bias parameter (u) that subsumes all other influences leading to publication of a significant finding (e.g., flexibility in data collection, analysis, and reporting). These influences may for instance stem from unbalanced assignment of participants across conditions in between-participants designs (i.e., unsuccessful

randomization, likely to apply to low sample sizes), lack of double-blindness, coding errors, and, more generally, all influences that may contribute to increasing the rate of false positives.

Within this approach, the three extreme scenarios we have just discussed can be conceptualized as follows. In the first scenario, it was assumed that power = .26, R = 1, and u = 0 (these assumptions would yield a perfect PPV value of 1, given that there are no false hypothesis). In the second scenario, it was assumed that statistical power (i.e., $1 - \beta$) = 1, the rate of true relationships in a field (R) = .26, and bias (u) = 0 (this scenario would yield a PPV of 88%, with approximately 30% of results turning out statistically significant). Both would suggest high rates of successful replication (inconsistent with observations). More consistent with the observed low rates of replication success, the third scenario of 26% false positives in the absence of true effects (R=0) can be accommodated in this approach either via alpha-level inflation (i.e., alpha = .26 and u=0), or via the bias parameter (i.e., alpha=.05 and u = .22); given that there are no true hypotheses, the PPV of the third scenario would be 0 (regardless of statistical power).

Turning to more realistic scenarios, we simulated the PPV for a two-sided independent-samples t-test. As a starting point, we relied on the observation that the median sample size is N=57 and we set the alpha level to .05. Based on the median sample size, we considered very small, medium, and very large effect sizes (i.e., Cohen's d = .1, .5, .9) for computing statistical power (i.e., $1 - \beta$; the resulting power values were .07. .46, and .92). For the probability of true hypotheses, we considered the values 1 in 100 (.01), 1 in 10 (.1), or 1 in 3 (.33). Regarding bias, we considered values between .2 and .8 (see Ioannidis, 2005), as well as a zero-bias baseline.

We also considered the rate of significant results[4]: First, we found that 26% of all reported results were significant. On the other hand, 91% of all articles reported at least one significant result. If we assume that all reported results are equally theoretically relevant, we should consider scenarios more relevant that yield approximately 26% of significant findings. If, on the other hand, one of the reported results is of theoretical interest, and the other results are reported for completeness, then we should mainly consider scenarios with 91% of significant results. We will compute both variants, but because typically a single psychological hypothesis is in the focus of an investigation, we will focus our discussion on the latter scenario.

In an initial step, we narrowed down the possible scenarios by identifying those scenarios that are consistent with either the 26%-significant-results mark or the 91%-significant-results mark. It turned out that low values of bias (.2 or lower) were necessary to approximate 26% significant results, while high rates of bias (.8 or higher) were required to approximate the 91% observation. Therefore, we discuss here these two cases separately, assuming bias levels of .2 and .8 respectively.

Under the above set of assumptions, a high rate of significant results (.8 or higher) comes about only with large values of bias (i.e., u=.8). Within these cases, the PPV (i.e., the proportion of significant findings that represent a true effect) ranges from 1% to 55%,

[4] We also considered scenarios that yield a low proportion (approximately 20-30%) of significant results. This observation can be explained as the result of either (a) a low a-priori probability of true hypotheses as well as low power, combined with a contribution of bias (u=.2), or (b), in the absence of bias, by a high a-priori probability (.5) of true hypotheses with medium or high levels of power (>=.46; i.e., large effects). The second case appears implausible: In light of the findings reported here (i.e., low replication rate, independence of sample size and proportion of significant findings, exploratory approach), it is unlikely that the IN-OT field is free from bias and dominated by large effects with high a-priori probability. The first case, which features an a-priori probability for true hypotheses of (at most) .1 and some amount of bias, is therefore a more plausible account of the observed 26% of significant results.

depending largely on the a-priori proportion of true effects (and to a lesser degree on statistical power):

- If 1 in 100 tested hypotheses are true, then approximately 1 in 100 significant findings (1-1.2%) represents a true effect, regardless of power.

- If 1 in 10 tested hypotheses are true, we end up with 10-12% of true findings among the significant results (again, power has little effect).

- If 1 in 2 tested hypotheses are true, the scenario yields between 50 and 55% of true findings among significant results.

Assuming an a-priori proportion of true hypotheses considered typical for psychology (i.e., .1; see Pashler & Harris, 2012), this implies that only approximately 829 (total number of interactive effects) * .26 (overall proportion of significant results) * .12 (most likely PPV = 26 of the reported significant findings in the IN-OT literature (that is, only 3 % of the tested interactive IN-OT effects) represent true effects. This unsatisfactory state of affairs can and should be improved in future IN-OT research. We now discuss how.

Simulations for concrete guidance.

In order to provide more concrete guidance for increasing the reliability of IN-OT research, we simulated the effects on PPV of increasing power, reducing bias, and conducting replications. These are familiar recommendations (see for instance Walum et al., 2016). We quantify here how adherence to these recommendations may increase the reliability of IN-OT research.

The first and most straightforward way to improve the PPV is to ***reduce bias***. This can be achieved for instance by shifting to a confirmatory approach and using a Registered Report article format, in which a journal reviews and accepts a study prior to data collection and

commits to publishing the results. Another way to reduce bias is by systematically implementing double-blindness. In the absence of bias, the PPV increases from 10-12% (depending on power) to 13% (for low power), 50% (for medium power), or 67% (for high power). In other words, while eliminating bias does not help much given low power, if a study reaches at least medium levels of power (here, $1\text{-}\beta \geq .46$), reducing bias yields a massive increase in PPV.

Second, from this it immediately follows that ***increasing power*** can increase the PPV, but this increase depends on the level of bias: With high levels of bias, an increase in power (from low to medium, or from medium to high) increases the PPV by only 1 percentage point. Without bias, in contrast, increasing power from low to medium improves PPV by 37 points, while an increase from medium to high yields an additional improvement of 17 points.

The third way to improve the PPV is by conducting (unbiased) ***replications*** of the significant results. Successful replication improves the PPV of a set of findings, but the magnitude of the benefit again depends on power (of both initial and replication study): As compared to a PPV of 10-12% among initial findings, with low power, the rate of true findings among the significant replications is only slightly increased to 13%. With medium power, already more than half of significant replications (53%) represent true findings. With high power, 71% of successful replications represent true findings.

Take-home messages from simulations.

If the goal is to increase the PPV, it is helpful to combine at least two recommendations: Reducing bias yields little improvement in PPV unless statistical power is medium or high. Vice versa, increasing power is of little effect unless bias is reduced. When the three are combined (i.e., low-bias, high-powered replications), a PPV of 97% can be achieved. Notably, however, reducing bias alone would have important benefits by substantially lowering the rate

of significant findings, and therefore helping to focus researchers' resources on a smaller set of potential true effects.

Increasing power typically implies using larger samples: The present simulation is based on median N = 57, which yields high power only for unusually large effects (d =.9). To achieve acceptable power (1-beta=.8) for medium effects (d=.5), at least N=128 is required; medium-to-small effects (d=.28) require N=404; for (very) small effects (d=.14), required sample size increases to N=1604. Such sample sizes appear difficult to achieve for a typical researcher alone. Please note that we set rather than estimated effect sizes. Estimating the effect size would be highly questionable in the present context. This is because of the very high heterogeneity of effects tested along with the suspected presence of a large number of false positives. This issue precisely motivated our current approach (as compared to e.g., a meta-analytic one). Therefore, it appears necessary to pool resources across labs and coordinate efforts. It should be kept in mind, however, that increased power can also be achieved, for instance by (1) switching to within-subject designs, (2) maximizing the effect of IN-OT administration (e.g., by using to-be-determined optimal dosage and optimal time elapsed following administration), and (3) improving the reliability of dependent measures.

General Discussion

Across the last decade, IN-OT has built a reputation as the social hormone in the literature. Press coverage polarized this view by referring to OT as a “love hormone”[5], a “trust potion”[6] or a “social glue”[7]. This prompted companies to sell OT sprays for improving social

[5] See, for example: http://www.livescience.com/42198-what-is-oxytocin.html
[6] See, for example: http://www.livescience.com/3821-scientists-discover-trust-potion.html
[7] See, for example: http://www.fastcompany.com/1659062/social-networking-affects-brains-falling-love

competences, connectedness, and emotional well-being[8]. This simplistic view was rightly questioned on both empirical and theoretical grounds: IN-OT effects should be addressed in interaction with contextual and inter-individual factors (Bartz et al., 2011). In this review, we examined published studies reporting interactive effects of IN-OT on psycho-social outcomes. This review informs us that an impressively large diversity of conceptual moderations was tested, but that those were rarely tested more than once or twice and, when they were, that they were rarely replicated successfully.

We discussed three extreme scenarios in which the 26% of significant moderations are accounted by (1) low statistical power, (2) low rate of a-priori correct hypotheses, or (3) type I errors. The first scenario relies on the very implausible assumption that 100% of the tested interaction were true, and it is inconsistent with the fact that the sample size is unrelated to the proportion of significant moderations. The second scenario relies on the implausible assumption that all studies that detected a true effect were maximally powered and is inconsistent with the observation of a low conceptual replication rate. The third scenario is consistent with the finding that 91% of the published article contained at least one significant moderation and the low conceptual replication rate. We additionally relied on simulations for allowing a more nuanced interpretation of the current state of research and for methodological guidance.

In all likelihood, the number of true interactive IN-OT effects is largely overstated. We estimated that about 90% of the reported significant findings on the interactive IN-OT effects on human social behaviors are likely to be false positives. PVV simulations clearly point to the *combined* need for larger samples sizes, independent replications, and better attempts at controlling biases.

[8] See, for example: https://www.verolabs.com

It should be stressed that some of the interactions reported in the IN-OT literature may correspond to true effects. Of importance too, other true effects may have gone undetected because of a lack of power, or because replicative attempts were most of the time conceptual rather than direct, or simply because dosage or elapsed time were not optimal. It is most unfortunate, however, that the very large heterogeneity of published moderations, combined with the low replication rate and the low reliance on open-research practices makes it virtually impossible to discover the theoretically relevant signal (i.e., true effects) in the empirical noise (i.e., Type I errors). As a result of this disappointing state of affairs, the development of strong theories is hardly possible, and IN-OT seems therapeutically unexploitable at present.

In one of the most influential IN-OT articles, Bartz and colleagues (2011) strongly recommended to switch from a main to interactive view of IN-OT effects, stressing the role of contextual and individual variables. The authors argued that adopting this interactionist approach is a way to overcome disappointing limitations associated with research probing main IN-OT effects. The authors highlighted that "inconsistencies across studies should not be seen as 'noise,' but rather as clues to the context- and person-dependent nature of the effects of oxytocin" (Bartz et al., 2011, p. 301), and stressed that "characterizing this context- and person-dependency could enable more refined theorizing on the social effects of oxytocin in humans" (Bartz et al., 2011, p. 301).

The results from the current systematic review and simulations suggest that, quite disappointingly, the thesis laid out by Bartz et al. (2011) - although conceptually relevant - is actualized in a broad heterogeneity of largely underpowered and inconsistent tests of interactive IN-OT effects. Although Bartz et al. (2011)'s argument is well grounded theoretically, it is awaiting a more reliable and coordinated implementation. We now discuss general recommendations to increase the trustworthiness of IN-OT research.

General Recommendations for Future Research

There is a positive side in the large number and diversity of studies conducted on IN-OT. Undoubtedly, OT research attracts multidisciplinary interest and addresses theoretically and practically important questions. Growing concerns regarding the reliability of IN-OT research, however, mitigate the interpretation of the effects reported and prevents meaningful advances in the field. The present review points to two necessary changes for promoting robust advances in the field (see below). These recommendations should be familiar to the informed reader. In the last decade, the replicability of research findings in psychology has been increasingly questioned (Ioannidis, 2005; Lehrer, 2010; Yong, 2012). This "credibility revolution" (Vazire, 2018) stimulated the adherence to open-science practices, which are now gaining momentum (Nosek & Lindsay, 2018).

First, the lack of coordinated efforts in identifying and testing interactive IN-OT effects is detrimental to both theoretical and practical advances. Although researchers should be left free to address the specific question they are interested in, the accumulation of hundreds of (arguably largely under-powered) studies that addressed different combinations between moderators and dependent variables ultimately prevents the establishment of robust knowledge and strong theories in this field. Complementarily, we strongly encourage researchers to report multiple experiments on a given research project. Within the 117 published articles that were identified here, only five presented more than one experiment. And, among those five, four presented only two experiments, and one presented a set of eight 'file-drawer' studies. As we discussed, a greater attention to statistical power is also highly recommended, and this likely involves multicentric research efforts.

Second, there should be much greater transparency in research practices, which is a precondition for collaboration and coordinated research efforts. In particular, pre-registration

of studies stating relevant theories, hypotheses, measures, analyses and power estimation, would greatly help limit the risk of false positives (Nosek et al., 2019). Relatedly, all moderators that are measured and tested should be reported as well as their effect (whether significant or not). Reviewers and editors should encourage such efforts and facilitate the publication of well-conducted studies, whichever their results are. One way to achieve this goal is to institutionalize such changes: as suggested by Leng and Ludwig (2016), journals may explicitly encourage researchers to preregister trials, to declare hypotheses and outcomes in advance, to specify statistical methods to be applied and to fully disclose materials, measures and data, including assessed moderators that did not influence the results. Even though effortful at first, this increased transparency of research practices seems necessary if we want to build a solid empirical basis for elaborating strong theories on the psychosocial effects of IN-OT.

These recommendations were framed in this review in terms of increasing "significant" findings that reflect true effects. The Null-Hypothesis Significance Testing (or NHST) approach is the most common one in this domain of research. The recommendations we suggest are not relevant only for the traditional NHST approach, though. One could argue that correcting alpha levels or turning to Bayesian statistics would be sufficient for improving the reliability of published results. Simmons and colleagues (2011), however, argued that such recommendations are clearly insufficient and could even make the problem worse by introducing more ambiguity and "researcher degrees of freedom". These authors argued that flexibility in data collection, analysis and reporting are the main drivers of low reliability research. Constraining or at least disclosing this flexibility requires adopting transparent research practices.

Limitations

In examining *conceptual* moderations, we made sure not to impose our own categorization strategy in establishing the conceptual DVs by relying on keywords selected by the authors themselves. It is possible and even likely that other categorization decisions would have brought a different picture. Because the present contribution is meant to advance IN-OT research, we share our data publicly and encourage the IN-OT research community to come up with their own theory-driven categorization of these DVs. We would like to insist, however, that this categorization should be theory-driven (i.e., the risk of falsely "discovering" an informative categorization increases with the number of attempts). Of note, we also coded for additional factors (e.g., gender, dosage, elapsed time) that may also be helpful for complementary analyses (with a similar note of caution).

The present review highlights that the interactive effects of IN-OT on psycho-social outcomes in a healthy population are highly heterogeneous. The broad focus of the review implies that a large number of conceptual dependent variables were entered in the analyses. Admittedly, it cannot reasonably be assumed that oxytocin has the same interaction effects on conceptual dependent variables as diverse as emotion recognition, aggression, memory and mood, as these might involve distinct physiological networks. By allowing public access to the data covered in this review, we encourage the test of more specific hypotheses. Again, however, we would like to stress points of caution. It is possible that confounding variables or further moderators should be taken into account. As an illustration, interactions between IN-OT and gender could be qualified by additional moderators (e.g. menstrual cycle phase, hormonal contraception, age) that were not considered in this review. In particular, recent meta-analytic evidence suggests that endogenous oxytocin levels may fluctuate during the menstrual cycle (Engel et al., 2019). So, for instance, the IN-OT by gender interactions might

very well depend on the menstrual cycle phase. Consistent with the conclusions reached in this review, however, inclusion of covariates and higher order moderators should be theoretically motivated and pre-registered. Otherwise, their consideration has the potential to make things even worse as they would add "researcher degrees of freedom" that could in turn increase further the rate of false positives. Second, some variables are unlikely to yield informative results (e.g., due to lack of variability). For instance, if one is interested in the effect of IN-OT dosage, one should accommodate the fact that a very large majority of studies used only a 24 I.U administration.

The quantitative analyses that we conducted in this review were mainly interested in the rate of significant findings as compared to other estimates such as effect sizes. In estimating the rate of significant findings, we essentially aimed to answer *how much can we trust the IN-OT research?* In contrast, estimating effect sizes would have addressed *how much IN-OT affects social behavior?* Estimating effect sizes is typically done in meta-analyses. As explained in the introduction, previous meta-analyses have pointed to small-to-medium effect sizes of main IN-OT effects. However, these meta-analyses only focused on *main* IN-OT effects, which are now thought to have little theoretical relevance as compared to interactive ones (Bartz et al., 2011). These meta-analyses also concluded in the existence of effects that could not be replicated in later studies (Tabak et al., 2019). In the current review, we also reported a very low rate of significant IN-OT effects in studies relying on a pre-post design; that is, in studies that included a baseline control. We also discussed that, if effect sizes estimated in previous meta-analyses reflected true effect sizes, these effect sizes are likely to be inflated due to various sources of bias and their reliance on underpowered studies (Pereira & Ioannidis, 2011; Walum et al., 2016). More generally, by focusing on the rate of significant findings, we were able to provide estimations of the quantity of bias as well as the proportion

of false positives reported. We also covered a much broader range of dependent variables than what could be achieved in typical meta-analyses.

The present analysis focuses on interactive IN-OT effects on non-clinical samples. One could argue that interactive effects are more reliable and stronger in studies that examined clinical samples. Several studies examining the effects of IN-OT on socio-emotional variables have found different effects in clinical than healthy samples. For instance, in a recent meta-analysis Bürkner, Williams, Simmons, and Woolley (2017) observed that IN-OT improves high-level social cognition in participants diagnosed with schizophrenia but not in the healthy controls. Yet, another meta-analysis reported a stronger sensitivity to IN-OT in healthy controls than clinical samples on emotional intelligence (Leppanen et al., 2017), and yet another meta-analysis did not observe an effect of IN-OT on emotion recognition and empathy, but found a small effect on theory of mind (Keech, Crowe, & Hocking, 2018).

A last limitation of the present work is inherent to the broad perspective in which it was conducted, i.e., a mechanistic perspective aiming to track down the biological foundations of human behavior. While it is certainly stimulating and important to understand how biology shapes social behaviors, mechanistic research bears the risk of over-simplifying the understanding of human functioning, especially if we overlook complex interactions between hormones and between different body systems more broadly, all of which are in constant interaction with dynamic environments.

Conclusion

This systematic review points to important concerns about the current state of IN-OT research. However, it also points to clear and helpful recommendations. Following a decade of what may be fairly considered an exploratory phase of research, time has now come to

establish a more empirically sound and theoretically grounded research on IN-OT in humans. First and foremost, this requires a better understanding of the pharmacokinetics of IN-OT in humans, to guarantee the use of biologically sensible doses and administration-to-effects time windows. Second, this requires engaging in collective, multicentric, research efforts allowing for the necessary coordination to systematically conquer the conceptual space. Finally, research practices should conform to contemporary research transparency criteria. Although we realize that these recommendations may be demanding, it is mandatory to meet them better if we want to move beyond the current - stimulating but, as the current review suggests, empirically and theoretically problematic – exploratory stage of research.

References

Amico, J. A., Challinor, S. M., & Cameron, J. L. (1990). Pattern of Oxytocin Concentrations in the Plasma and Cerebrospinal Fluid of Lactating Rhesus Monkeys (Macaca mulatto,) : Evidence for Functionally Independent Oxytocinergic Pathways in Primates*. *The Journal of Clinical Endocrinology & Metabolism*, *71*(6), 1531-1535. https://doi.org/10.1210/jcem-71-6-1531

Bartz, J. A., Zaki, J., Bolger, N., Hollander, E., Ludwig, N. N., Kolevzon, A., & Ochsner, K. N. (2010). Oxytocin selectively improves empathic accuracy. *Psychological science*, *21*(10), 1426-1428.

Bartz, J. A., Zaki, J., Bolger, N., & Ochsner, K. N. (2011). Social effects of oxytocin in humans : Context and person matter. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2011.05.002

Born, J., Lange, T., Kern, W., McGregor, G. P., Bickel, U., & Fehm, H. L. (2002). Sniffing neuropeptides : A transnasal approach to the human brain. *Nature Neuroscience*, *5*(6), 514. https://doi.org/10.1038/nn0602-849

Bürkner, P.-C., Williams, D. R., Simmons, T. C., & Woolley, J. D. (2017). Intranasal oxytocin may improve high-level social cognition in schizophrenia, but not social cognition or neurocognition in general : A multilevel Bayesian meta-analysis. *Schizophrenia Bulletin*, *43*(6), 1291-1303. https://doi.org/10.1093/schbul/sbx053

Cardoso, C., Ellenbogen, M. A., & Linnen, A.-M. (2012). Acute intranasal oxytocin improves positive self-perceptions of personality. *Psychopharmacology*, *220*(4), 741-749. https://doi.org/10.1007/s00213-011-2527-6

Carson, D. S., Berquist, S. W., Trujillo, T. H., Garner, J. P., Hannah, S. L., Hyde, S. A., … Strehlow, M. C. (2015). Cerebrospinal fluid and plasma oxytocin concentrations are

positively correlated and negatively predict anxiety in children. *Molecular psychiatry*, *20*(9), 1085.

De Dreu, C. K. W., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. J. (2011). Oxytocin promotes human ethnocentrism. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1262-1266. https://doi.org/10.1073/pnas.1015316108

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological psychiatry*, *65*(9), 728-731.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological psychiatry*, *61*(6), 731-733.

Engel, S., Laufer, S., Miller, R., Niemeyer, H., Knaevelsrud, C., & Schumacher, S. (2019). Demographic, sampling- and assay-related confounders of endogenous oxytocin concentrations : A systematic review and meta-analysis. *Frontiers in Neuroendocrinology*, *54*, 100775. https://doi.org/10.1016/j.yfrne.2019.100775

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences : Unlocking the file drawer. *Science*, *345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Guastella, A. J., Einfeld, S. L., Gray, K. M., Rinehart, N. J., Tonge, B. J., Lambert, T. J., & Hickie, I. B. (2010). Intranasal oxytocin improves emotion recognition for youth with autism spectrum disorders. *Biological psychiatry*, *67*(7), 692-694.

Heinrichs, M., von Dawans, B., & Domes, G. (2009). Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology*, *30*(4), 548-557. https://doi.org/10.1016/j.yfrne.2009.05.005

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., … Maier, W. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *Journal of Neuroscience, 30*(14), 4999-5007.

Insel, T. R. (1992). Oxytocin—a neuropeptide for affiliation : Evidence from behavioral, receptor autoradiographic, and comparative studies. *Psychoneuroendocrinology, 17*(1), 3-35.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kagerbauer, S. M., Martin, J., Schuster, T., Blobner, M., Kochs, E. F., & Landgraf, R. (2013). Plasma oxytocin and vasopressin do not predict neuropeptide concentrations in human cerebrospinal fluid. *Journal of neuroendocrinology, 25*(7), 668-673.

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

Kemp, A. H., & Guastella, A. J. (2011). The role of oxytocin in human affect : A novel hypothesis. *Current Directions in Psychological Science, 20*(4), 222-231.

Kirkpatrick, M. G., Francis, S. M., Lee, R., De Wit, H., & Jacob, S. (2014). Plasma oxytocin concentrations following MDMA or intranasal oxytocin in humans. *Psychoneuroendocrinology, 46*, 23-31.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673.

Landgraf, R., & Neumann, I. D. (2004). Vasopressin and oxytocin release within the brain : A dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology, 25*(3-4), 150-176. https://doi.org/10.1016/j.yfrne.2004.05.001

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*(9), e0137000. https://doi.org/10.1371/journal.pone.0137000

Lee, M. R., Scheidweiler, K. B., Diao, X. X., Akhlaghi, F., Cummins, A., Huestis, M. A., … Averbeck, B. B. (2018). Oxytocin by intranasal and intravenous routes reaches the cerebrospinal fluid in rhesus macaques : Determination using a novel oxytocin assay. *Molecular psychiatry*, *23*(1), 115.

Lehrer, J. (2010). The truth wears off. *The New Yorker*, *13*(52), 229.

Leng, G., & Ludwig, M. (2016). Intranasal oxytocin : Myths and delusions. *Biological psychiatry*, *79*(3), 243-250.

Leppanen, J., Cardi, V., Ng, K. W., Paloyelis, Y., Stein, D., Tchanturia, K., & Treasure, J. (2017). Effects of intranasal oxytocin on the interpretation and expression of emotions in anorexia nervosa. *Journal of Neuroendocrinology*, *29*(3), 1-13. https://doi.org/10.1111/jne.12458

Meyer-Lindenberg, A., Domes, G., Kirsch, P., & Heinrichs, M. (2011). Oxytocin and vasopressin in the human brain : Social neuropeptides for translational medicine. *Nature Reviews Neuroscience*, *12*(9), 524-538. https://doi.org/10.1038/nrn3044

Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses : The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Naber, F., van IJzendoorn, M. H., Deschamps, P., van Engeland, H., & Bakermans-Kranenburg, M. J. (2010). Intranasal oxytocin increases fathers' observed responsiveness during play with their children : A double-blind within-subject experiment. *Psychoneuroendocrinology*, *35*(10), 1583-1586. https://doi.org/10.1016/j.psyneuen.2010.04.007

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

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815-818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*(3).

Pereira, T. V., & Ioannidis, J. P. A. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, *64*(10), 1060-1069. https://doi.org/10.1016/j.jclinepi.2010.12.012

Radke, S., & de Bruijn, E. R. A. (2012). The other side of the coin : Oxytocin decreases the adherence to fairness norms. *Frontiers in Human Neuroscience*, *6*, 7. https://doi.org/10.3389/fnhum.2012.00193

Rimmele, U., Hediger, K., Heinrichs, M., & Klaver, P. (2009). Oxytocin makes a face in memory familiar. *The Journal of Neuroscience*, *29*(1), 38-42. https://doi.org/10.1523/JNEUROSCI.4260-08.2009

Shamay-Tsoory, S. G., & Abu-Akel, A. (2016). The social salience hypothesis of oxytocin. *Biological Psychiatry*, *79*(3), 194-202. https://doi.org/10.1016/j.biopsych.2015.07.020

Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., & Levkovitz, Y. (2009). Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biological psychiatry*, *66*(9), 864-870.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything

as Significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Spengler, F. B., Schultz, J., Scheele, D., Essel, M., Maier, W., Heinrichs, M., & Hurlemann, R. (2017). Kinetics and Dose Dependency of Intranasal Oxytocin Effects on Amygdala Reactivity. *Biological Psychiatry*, *82*(12), 885-894. https://doi.org/10.1016/j.biopsych.2017.04.015

Striepens, N., Kendrick, K. M., Hanking, V., Landgraf, R., Wüllner, U., Maier, W., & Hurlemann, R. (2013). Elevated cerebrospinal fluid and blood concentrations of oxytocin following its intranasal administration in humans. *Scientific Reports*, *3*(1). https://doi.org/10.1038/srep03440

Tabak, B. A., Teed, A. R., Castle, E., Dutcher, J. M., Meyer, M. L., Bryan, R., … Eisenberger, N. I. (2019). Null results of oxytocin and vasopressin administration across a range of social cognitive and behavioral paradigms : Evidence from a randomized controlled trial. *Psychoneuroendocrinology*. https://doi.org/10.1016/j.psyneuen.2019.04.019

Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., & Rogers, P. J. (2009). Oxytocin and social perception : Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and behavior*, *56*(1), 128-132.

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411-417. https://doi.org/10.1177/1745691617751884

Walum, H., Waldman, I. D., & Young, L. J. (2016). Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. *Biological psychiatry*, *79*(3), 251-257.

Yong, E. (2012). Replication studies : Bad copy. *Nature*, *485*(7398), 298-300. https://doi.org/10.1038/485298a