Towards better hypothesis tests in oxytocin research: Evaluating the validity of auxiliary assumptions

Daniel S. Quintana

**ARTICLE INFO**

Keywords:
- Oxytocin
- Social cognition
- Social behavior
- Methodology
- Theory
- Reproducibility

**ABSTRACT**

Various factors have been attributed to the inconsistent reproducibility of human oxytocin research in the cognitive and behavioral sciences. These factors include small sample sizes, a lack of pre-registered studies, and the absence of overarching theoretical frameworks that can account for oxytocin’s effects over a broad range of contexts. While there have been efforts to remedy these issues, there has been very little systematic scrutiny of the role of auxiliary assumptions, which are claims that are not central for testing a hypothesis but nonetheless critical for testing theories. For instance, the hypothesis that oxytocin increases the salience of social cues is predicated on the assumption that intranasally administered oxytocin increases oxytocin levels in the brain. Without robust auxiliary assumptions, it is unclear whether a hypothesis testing failure is due to an incorrect hypothesis or poorly supported auxiliary assumptions. Consequently, poorly supported auxiliary assumptions can be blamed for hypothesis failure, thereby safeguarding theories from falsification. In this article, I will evaluate the body of evidence for key auxiliary assumptions in human behavioral oxytocin research in terms of theory, experimental design, and statistical inference, and highlight assumptions that require stronger evidence. Strong auxiliary assumptions will leave hypotheses vulnerable for falsification, which will improve hypothesis testing and consequently advance our understanding of oxytocin’s role in cognition and behavior.

**1. Introduction**

Oxytocin is an evolutionarily ancient hormone and neuromodulator (Theofanopoulou et al., 2021) that is primarily synthesized in the brain for both central and peripheral release (Chini et al., 2017; Fig. 1). Oxytocin was first identified via its role in childbirth and lactation (Dale, 1906; Ott and Scott, 1910), but more recently has been the subject of considerable research interest for its involvement in human behavior and cognition (Kenkel, 2019; Leng and Leng, 2021) and its potential to treat psychiatric disorders characterized by social dysfunction (Grinevich and Neumann, 2020). However, it has proven difficult to reliably replicate the early reports of oxytocin’s effect on social cognition and behavior (Alvares et al., 2017; Mirop et al., 2020; Nave et al., 2015; Sikich et al., 2021). Moreover, several studies (e.g., Declerck et al., 2020; Klackl et al., 2013; Yao et al., 2014) have failed to replicate the first landmark report in the field published in 2005, that intranasal oxytocin administration increases trust behaviors (Kosfeld et al., 2005).

Various factors have contributed to the lack of reproducibility in oxytocin research, including small sample sizes (Walum et al., 2016), the inconsistent use of pre-registration (Leng and Ludwig, 2016), and the absence of overarching theories that advance beyond oxytocin’s role in social behavior (Winterton et al., 2021b). However, one issue that has received little systematic attention in the field is the impact of poorly-supported auxiliary assumptions (Lakatos, 1978a; Meehl, 1990a; Scheel et al., 2020; Trafimow, 2012), which are also known as auxiliary hypotheses (thus these terms will be used interchangeably throughout this paper). For example, the core hypothesis that oxytocin administration influences social cognition is predicated on the auxiliary assumption that intranasally administered oxytocin elevates levels of oxytocin in the brain. Research that uses p-values to make claims, which is the majority of oxytocin research, operates under the Popperian hypothetico-deductive framework (Fidler et al., 2018), which is designed to falsify hypotheses (although in practice, most researchers use this framework to confirm hypotheses). The hallmark of robust
Intranasal oxytocin delivery pathways and endogenous oxytocin release. Endogenous oxytocin is primarily synthesized by magnocellular and parvocellular nuclei in the hypothalamus. Oxytocin is delivered and stored in the posterior pituitary for release into peripheral circulation. Oxytocin is also secreted to other regions of the brain via axonal and dendritic mechanisms. Intranasally administered oxytocin is thought to enter the brain via olfactory and trigeminal nerve pathways. This administration route bypasses the blood brain barrier, which limits the entry of large molecules like oxytocin from peripheral blood circulation. Intranasal oxytocin also increases peripherally circulating oxytocin levels.

Figure adapted from Quintana et al. (2021) and created with Biorender.com.
complement other types of research with narrow phenotypes (Lyon et al., 2019; Pingault et al., 2018). Indeed, recent work using a polygenic approach in a sample of just under half a million participants with genetic data and broad health phenotypes demonstrated oxytocin’s role in energy regulation (Winterton et al., 2021a), which has been previously shown in small oxytocin administration studies with narrow phenotypes (e.g., Ott et al., 2013). Triangulation can also play a role when selecting data models by performing different types of analysis that rely on different assumptions (e.g., Beevers et al., 2019), such as frequentist and Bayesian hypothesis testing.

The objective of this article is to carry out an integrative evaluation of the research supporting critical auxiliary hypotheses in human oxytocin research. These assumptions will be organized via ascending levels of the derivation chain described above: theoretical models, experimental models, and data models (Tunc et al., 2021). While resolving weak (i.e., poorly supported) links in the derivation chain and uncertain auxiliary assumptions is an important first step for generating better hypothesis tests and improving the replicability of oxytocin research by exposing theories to falsification opportunities (Scheel et al., 2020), triangulating multiple experimental models and data models will facilitate stronger auxiliary assumptions than using individual models alone.

1.1. Theoretical models

The foundation of the derivation chain is its theoretical core. The reproducibility crisis in the biobehavioral sciences has been primarily attributed to questionable research practices and a lack of pre-registration (Munafo et al., 2017), but an underrecognized factor has been a lack of overarching theories that can help generate hypotheses across multiple contexts (Muthukrishna and Henrich, 2019). Hypotheses generated via past results in specific contexts, instead of via broad theoretical frameworks, are less likely to receive consistent empirical support as they tend to scrutinize auxiliary assumptions less carefully. For example, it was once thought that oxytocin was a pro-social hormone (Zak, 2012), which was partly based on reports that oxytocin administration increases trust (Rosfeld et al., 2005) and generosity (Zak et al., 2007), and linked to oxytocin’s role in maternal care. However, several failed conceptual replication attempts (for a review of these studies, see Nave et al., 2015), and more recently, a large unsuccessful direct replication (Deecker et al., 2020), suggested otherwise. A failure for subsequent research to find effects of oxytocin on pro-social behavior in other contexts is somewhat unsurprising considering the pro-social theory was not compatible with some, but certainly not all, prior animal work demonstrating oxytocin’s effects on aggression (Bales and Carter, 2003; Ferris et al., 1992; Young et al., 1998) and putatively non-social processes, such as energy regulation (Uvnäs-Moberg, 1994), bone remodeling (Elabd et al., 2007), and appetite (Arletti et al., 1989). These results highlight the complexity and important role of context on oxytocin’s effects. While such incompatibilities are easier to identify in hindsight, this example illustrates the perils of basing hypotheses on a specific set of past research without considering broader theoretical frameworks.

Two theories emerged to replace the pro-social theory that accounted for oxytocin’s effects on both pro-social and anti-social behavior: the social salience theory (Shamay-Tsoory et al., 2009; Shamay-Tsoory and Abu-Akel, 2016) and social approach/withdrawal theory (Kemp and Guastella, 2011, 2010). While these theories account for both pro- and anti-social behaviors, they are still largely based on prior results within similar domains, in which the predictive ability of hypotheses is more limited, instead of a broader theoretical framework that is applicable across a variety of domains (Muthukrishna and Henrich, 2019). A general approach-avoidance theory that accounts for social and non-social behavior was later proposed (Harari-Dahan and Bornstein, 2014). While this theory recognizes oxytocin’s cardiovascular effects, it does not make explicit predictions for oxytocin’s effects across broad somatic domains and how this can influence behavior. In response to this theory gap, we recently proposed an Allostatic Theory of Oxytocin that suggests oxytocin helps maintain the stability of several physiological processes (e.g., energy regulation) under changing conditions (Quintana and Guastella, 2020). Unlike a homeostatic system, which responds to changes from static physiological set points post hoc, an allostatic system adjusts physiological set points based on current environmental demands and anticipatorily shifts physiological parameters according to the prediction of future environmental changes based on prior learning (Ramsay and Woods, 2014). Rather than being informed primarily from past research, this theory was developed using an integrative evolutionary and proximate explanation framework by answering Nikolaas Tinbergen’s “Four Questions” (Tinbergen, 1963): how does oxytocin work (mechanism); how does the role of oxytocin change during development (ontogeny); how does oxytocin enhance survival (purpose); and how did the oxytocin system evolve (phylogeny)?

The answers to these questions converged towards oxytocin’s role in the four key elements of an allostatic system, as shown in Fig. 3: sensing (Beets et al., 2012; Rash et al., 2014), responding (Elia et al., 2016; Scheele et al., 2013), learning (Beets et al., 2012; Quintana et al., 2019a), and prediction (Zheng et al., 2014). A theory does not exist in isolation as it relies on theoretical auxiliary hypotheses (Tunc and Tunc, 2020). For instance, there are two primary auxiliary hypotheses for the Allostatic Theory of Oxytocin. First, this theory depends on the
assumptions of the overall concept of allostatics (Ramsay and Woods, 2014; Sterling, 2012), which is that organisms adjust physiological set points (e.g., fluid balance) based on current environmental demands (e.g., ambient temperature) and can shift these parameters in the anticipation of future environmental changes. Second, this theory was partly derived by surveying the evolution of the oxytocin signaling system by using genomic comparative methods, thus it is also dependent on evolutionary theory (Hofmann et al., 2014). However, our current understanding of oxytocin signaling’s evolution is based on a non-systematic comparison of genomes from species across the evolutionary timeline (Feldman et al., 2016), which has many gaps in coverage (Theofanopoulou et al., 2021). Our understanding of oxytocin signaling’s evolution may change as more high-resolution genetic datasets across more species become available. In sum, accepting the assumption, suggesting that very little intranasally administered oxytocin actually reaches the brain (Leng and Ludwig, 2016). More specifically, by evaluating prior research measuring oxytocin levels in cerebrospinal fluid (CSF) before and after intranasal oxytocin administration, Leng and Ludwig (2016) calculated that less than 1% of intranasally administered oxytocin reaches the central nervous system. Assuming that the half maximal effective oxytocin concentration required for human oxytocin receptor binding is 10 micromolars (Pasoni et al., 2016), and that oxytocin concentrations in human CSF after intranasal administration is less than 1 micromolar (Striepens et al., 2013), it has been suggested that these CSF increases are not likely to be physiologically meaningful (Bowen, 2019). However, research has found that increased oxytocin levels in microdialysates from the amygdala and hippocampus of rodents after intranasal administration is not matched by increased CSF oxytocin levels (Neumann et al., 2013). Consequently, CSF measures of oxytocin may not represent oxytocin’s action in brain parenchyma, where oxytocin receptors are located. But since it is not feasible to collect microdialysates in humans, except for convenience samples from individuals undergoing neurocritical care (Hutchinson et al., 2015), it is difficult to evaluate whether these rodent results are applicable to humans.

Peripheral circulating oxytocin is thought to not easily cross the blood brain barrier (BBB) due to its structure (but see Yamamoto et al., 2019), so researchers have turned to intranasal administration as an alternative approach to deliver oxytocin to the brain (Fig. 1). However, because of the peculiar anatomy of the human nasal cavity, intranasal oxytocin delivery can also be a challenging procedure (Djupesland et al., 2013; Quintana et al., 2018a). It is thought that oxytocin molecules deposited in the upper and posterior regions of the nasal cavity can be directly transported to the brain via extracellular mechanisms along olfactory and trigeminal nerve fibers, which innervate these hard-to-reach nasal cavity regions (Lochhead and Thorne, 2012; Quintana et al., 2018a). One line of indirect evidence that intranasally administered oxytocin reaches the brain in humans is research reporting oxytocin’s significant influence on neural activity compared to placebo, particularly in the left superior temporal gyrus (Grace et al., 2018) and amygdala (Wang et al., 2017). There is also a high degree of crossover between oxytocin receptor gene expression patterns in the brain (Quintana et al., 2019a) and meta-analytically derived neural activity patterns after oxytocin administration (Habets et al., 2021), as identified by Grace and colleagues (2018). However, changes in neural activity patterns do not provide causal evidence that intranasal oxytocin elevates oxytocin levels in the brain. For example, it is possible that oxytocin’s effects on peripheral receptors, which are located throughout the body (Jurek and Neumann, 2018), are indirectly influencing brain activity via feedback mechanisms. Therefore, other approaches are required to directly assess whether intranasal oxytocin increases oxytocin levels in the brain.

Animal researchers have used direct measures of oxytocin concentrations in the central nervous system via the collection of brain tissue or microdialysates to demonstrate that intranasal administration can
elevate oxytocin levels in the brain (e.g., Lee et al., 2020; Smith et al., 2019). Importantly, by assessing oxytocin concentrations after intranasal administration in oxytocin gene knockout mice, who cannot produce oxytocin, Smith and co-workers (2019) demonstrated that observed oxytocin concentration increases in brain parenchyma was exogenous. In other words, central increases in oxytocin do not seem to be due to peripheral oxytocin receptor binding providing a feedback signal to the brain to produce additional endogenous oxytocin. As the collection of brain tissue or microdialysates to index central levels after intranasal administration is not practical for human research, cerebrospinal fluid (CSF) collection can be potentially used instead as a measure of oxytocin concentrations after intranasal administration. This CSF approach has been used in non-human primate studies, which have reported that intranasal administration increases oxytocin levels in the central nervous system (Dal Monte et al., 2014; Freeman et al., 2016; Lee et al., 2018). However, this is not practical for widespread use in humans given the invasiveness of CSF collection. Indeed, most studies that have collected CSF for the measurement of oxytocin concentrations have used convenience samples (e.g., Carson et al., 2014). There has been one human study that describes the calculation of oxytocin levels in CSF levels after intranasal oxytocin administration, reporting increased oxytocin levels (Striepens et al., 2013), but further verification is required due to the study’s small sample size.

As well as the question of whether intranasal oxytocin administration meaningfully increases oxytocin levels in the brain, there is also uncertainty regarding how oxytocin is transported from the nose to the brain. Almost every oxytocin administration study uses intranasal administration, despite the challenges associated with this route (Insel, 2016). There have been several different approaches to determine whether intranasally administered oxytocin reaches the brain via a direct nose-to-brain pathway, or whether oxytocin’s cognitive and behavioral effects are derived from peripheral actions (Quintana et al., 2015a). One such approach is the direct comparison of the cognitive and neural effects of intranasal and peripheral oxytocin administration, which to date been investigated in relatively small samples (Martins et al., 2020b; Quintana et al., 2015b, 2019b). If the BBB limits the transport of intravenously administered oxytocin to the brain, then intravenous oxytocin administration should not have any appreciable effects on cognition and neural activity. However, if effects on cognition and neural activity are observed after intranasal administration, and oxytocin concentrations in peripheral circulation are similar between intranasal and intravenous administration conditions, then one could conclude that intranasal administration did not enter the brain by crossing the BBB via the circulatory system (Fig. 1). Comparing intranasal and intravenous administration only revealed significant differences compared to placebo after intranasal administration (Quintana et al., 2019b, 2016, 2015b). But these effects are probably nuanced, as Martins and colleagues (2020b) reported comparable effects in some regions after both intranasal and intravenous administration, whereas neural activity in other regions was only observed after intranasal administration.

Larger effects after intranasal oxytocin administration, compared to peripheral oxytocin administration, have also been observed in mice models of autism (Peñagarikano et al., 2015). Remarkably, Peñagarikano and colleagues (2015) also found that stimulating central endogenous oxytocin release via a selective melanocortin 4 receptor agonist rescued social deficits, demonstrating that exogenous oxytocin can modulate central neural processes. However, this wider body of results is not consistent with work from Kou and colleagues (2021), which indicates that oral oxytocin administration, but not intranasal administration, modulates social cognition. This study also found differential effects on brain activity depending on the oxytocin administration method (i.e., oral vs. intranasal), which suggests that different routes might target different brain regions, like Martins and colleagues reported (2020b).

Table 1 Advantages and assumptions of methods to determine intranasally administered oxytocin reaches the brain.

| Method | Advantages | Assumptions |
|--------|------------|-------------|
| Measuring CSF levels | Relatively feasible to collect in humans | Oxytocin does not cross the BBB |
| Measuring microdialysis levels | Direct measure of brain parenchyma levels | Changes reflect endogenous oxytocin activity |
| Comparing delivery response route | Relatively easy to assess in humans | Oxytocin delivery pathways |
| Administering radiolabeled oxytocin | Can track oxytocin delivery pathways | Radiolabelled oxytocin behaves like oxytocin |
| Assessing neural activity | Relatively easy to assess in humans | Activity reflects receptor binding |

Note: BBB = blood brain barrier; CSF = cerebrospinal fluid.
nose-to-brain delivery (Djupesland and Skretting, 2012; Guastella et al., 2013). To address this, nasal spray devices have been developed that are specifically designed for nose-to-brain delivery, which have been used in oxytocin research (Martins et al., 2020b; Quintana et al., 2015b). In terms of experimentally validating that intranasal oxytocin administration increases concentrations in the central nervous system, meta-analysis has demonstrated that blood concentrations can provide a satisfactory proxy of central concentrations (Valstad et al., 2017).

2.2. The uniformity of effects

As oxytocin is often described as a “social” hormone, this can imply that oxytocin plays a role in all types of social-cognitive processing, which is a considerably broad category (Frith and Frith, 2012). This wide categorization can also imply that intranasal oxytocin will inevitably benefit any target population with deficits in social behavior, which is reflected by the fact that intranasal oxytocin administration has been evaluated across almost all psychiatric disorder categories, including schizophrenia spectrum disorders, trauma related disorders, affective disorders, neurodevelopmental disorders (Peled-Avron et al., 2020), anxiety disorders (De Cagna et al., 2019), personality disorders, and neurocognitive disorders (Leppanen et al., 2017). While there is mixed evidence for the effectiveness of intranasal oxytocin treatment across these disorders, in some cases intranasal oxytocin treatment has performed worse than placebo (e.g., Borderline personality disorder; Bartz et al., 2010). However, it is worth noting that this study design (a between participants design with six participants assigned oxytocin and eight participants assigned placebo) and test (independent samples t-test) could only reliably detect (i.e., 80% power) an effect size of $\delta = 1.6$ or larger (alpha = 0.05 with a two-sided criterion for detection). Instead of assuming that oxytocin will be beneficial for any illness, or that it will always benefit social-cognitive processes, researchers should carefully consider this auxiliary assumption and direct resources to studies that are more likely to yield positive and replicable results. Moreover, given the breadth of social cognition there are no universally accepted markers of what oxytocin treatment is supposed to be effective for, which highlights the importance of researchers partnering with individuals diagnosed with psychiatric conditions and their allies to determine meaningful measures of interest (Fletcher-Watson et al., 2019). Another important consideration related to the effects of oxytocin is safety, especially in terms of long-term administration. To date, safety has been demonstrated in autism for at least 6 months of oxytocin treatment (DeMayo et al., 2017), but this needs to be continually monitored and demonstrated for longer periods, as well as in other treatment groups.

A related issue for assuming the uniformity of oxytocin’s effects is that results from studies with specific psychiatric populations or investigating a particular social cognitive process—both positive and negative—are often extended to all psychiatric populations or social cognitive processes. Although a positive result can provide a guide for hypothesis generation for oxytocin’s effects in other domains and a negative result can be used to help inform the abandonment of a theory among a larger body of work, researchers should avoid making broad inferences regarding their study’s observed effects, or lack thereof (Yarkoni, 2020).

2.3. The most efficacious dose is 24 International Units

Most intranasal oxytocin studies administer a 24-international unit (IU) dose. However, the choice of this particular dose is mostly due to precedent rather than rigorous dose-response research. The most efficacious oxytocin dose is another critical assumption to assess, as a failure of oxytocin’s effects can easily be attributed to the incorrect dosage rather than oxytocin’s effects (Insel, 2016). If the intranasal dose is too low this will not make an appreciable difference in oxytocin receptor binding. If the dose is too high this might lead to vasopressin receptor binding, which can elicit the opposite effects of oxytocin (Neumann and Landgraf, 2012). In research comparing a traditional 24IU oxytocin dose with a lower 8IU oxytocin dose in terms of brain activity (Quintana et al., 2016), social cognition (Quintana et al., 2015b), and pupillometry (Quintana et al., 2019b), effects were only observed relative to placebo after the 8IU administration, although these results were derived from a small sample. Another neuroimaging study also reported the largest changes in amygdala activity after a lower intranasal oxytocin dose (Martins et al., 2021), however, others have reported that a 24IU dose is associated with a stronger decrease in amygdala activation, compared to 12IU or 48 IU (Spengler et al., 2017).

Considering the evidence to date, there are some indications that a lower dose might be more efficacious, perhaps due to higher doses occupying vasopressin receptors. However, the use of different nasal spray devices between studies makes the direct comparison of doses between studies difficult, as a low dose using an optimized nasal spray device (Martins et al., 2021; Quintana et al., 2019b, 2016, 2015b) might be equivalent to a conventional dose delivered using a conventional pump-actuated nasal spray device (Spengler et al., 2017). Another implicit auxiliary assumption when evaluating the dose response is that dosage has a uniform effect across brain regions. However, as this may not necessarily be the case (Martins et al., 2021) it is important to better establish the effects of different dosages on activity in different brain regions. Similarly, it may also be assumed that dosage will have a uniform effect on cognitive processes but there is little evidence to support this. Altogether, these myriad issues highlight the benefits of administering different doses within the same study, which negates the need to rely on dose-response information from other studies. The development of an oxytocin receptor ligand for PET studies would also assist the discovery of the optimum intranasal oxytocin dose. There has been some progress regarding a PET ligand in animal models (Beard et al., 2018), but a PET ligand has yet to be tested in humans.

2.4. Oxytocin concentrations in peripheral fluids are a reliable measure of oxytocin system activity

The calculation of oxytocin concentrations from peripheral fluids (e.g., blood, saliva, and urine) is a common approach for indexing oxytocin activity, but this has been the subject of critique from some quarters (e.g., Leng and Sabatier, 2016). Basal oxytocin levels are often correlated with psychological phenotypes (e.g., Fuji et al., 2016) or compared between psychiatric and neurotypical groups (e.g., Bakker-Huvenaars et al., 2020). There are several auxiliary hypotheses underlying this approach. For example, one critical assumption for using peripheral oxytocin concentrations as a biomarker is the test-retest reliability of neuropeptide measures. However, evidence suggests that there is poor week-to-week reliability of oxytocin concentrations in extracted plasma samples and unextracted saliva samples (Martins et al., 2020a). As this study used a male sample and a single baseline assessment, additional research is required in females and using an average of pooled samples to improve generalizability and reliability. Evidence also points to poor week-to-week reliability of plasma vasopressin (a closely related neuropeptide) concentrations in males (Quintana et al., 2017) and females (Stachenfeld et al., 1999) collected via single baseline assessments. Moreover, both participant age and time of day seems to influence oxytocin concentrations (Engel et al., 2019), highlighting the need for the careful control of these covariates. Similarly, the measurement of salivary oxytocin levels after intranasal oxytocin is inaccurate, as detected oxytocin simply reflects endogenous oxytocin that has dripped down from the nasal cavity into the oral cavity (Martins et al., 2020a; Quintana et al., 2018b), not oxytocin concentrations in the periphery, at least in males.

Another common assumption, often made implicitly, is that basal levels of peripheral oxytocin are related to levels of oxytocin in the central nervous system. However, a meta-analysis has demonstrated that peripheral oxytocin measures are not strongly related to central
concentrations under resting state conditions (i.e., baseline oxytocin levels) suggesting that peripheral and central oxytocin secretion is not necessarily coordinated (Valstad et al., 2017). Altogether, hypotheses associating oxytocin levels in peripheral fluids with phenotypes of interest typically rely on weak auxiliary assumptions due to a lack of evidence supporting both the reliability of these measures and a link between peripheral and central levels of oxytocin under resting state conditions. However, more evidence is required across heterogeneous populations before these assumptions can be dismissed. In the meantime, if researchers would like to meaningfully link peripheral oxytocin concentrations to phenotypes of interest or use this measure as an illness biomarker, they should also demonstrate the reliability of peripheral oxytocin concentration measures with repeated assays from the same participants as well as also reporting detailed validation data (e.g., parallelism, accuracy, intra- and inter-assay variation; MacLean et al., 2019).

3. Data models

3.1. Statistical hypothesis testing

There are various assumptions one must examine when using statistical inference for hypothesis testing, which are seldom explicitly considered. Frequentist null hypothesis significance (NHST) testing is the dominant approach in the biobehavioral sciences. Frequentist NHST yields a\( p \)-value, which is the probability of observing the sample data, or data that is more extreme, assuming the null hypothesis is true. While \( p \)-values have been the subject of criticism (e.g., Cumming, 2008), they can be useful when two assumptions are satisfied: (1) the null hypothesis is plausible and (2) they are used to make ordinal claims (e.g., there is a difference between two groups) while keeping error control of false\( p \)-values. As previously demonstrated using frequentist NHST, it is certainly plausible that intranasal oxytocin administration has no effect on typical variables of interest or that oxytocin concentrations are not associated with a phenotype of interest, thus satisfying the first assumption. Whether researchers satisfy the second assumption of making an ordinal claim about an effect depends on how researchers interpret \( p \)-values. Frequentist NHST \( p \)-values cannot be used to quantify the probability that a claim or hypothesis is correct (Anderson, 2020). As researchers often misinterpret \( p \)-values as the probability that a hypothesis is correct (e.g., Badenes-Ribera et al., 2016), this suggests that this is something that many researchers want to know. Fortunately, Bayesian null hypothesis testing using Bayes factors offers a solution as this approach can quantify the relative evidence for two competing hypotheses (or models) given the observed data (Wagenmakers et al., 2018). Therefore, if researchers would like to quantify the evidence for a hypothesis and make claims like, “The alternative hypothesis is 7.4 more favored than the null hypothesis”, while satisfying statistical inference assumptions, they can use Bayes factors (For a Bayesian hypothesis test tutorial with oxytocin research examples, see Quintana and Williams, 2017).

Bayesian hypothesis testing is also not without its challenges, especially surrounding the specification of a prior distribution, which requires careful consideration (Wagenmakers et al., 2018). This is especially the case in research areas where prior data is scarce, such as the non-social effects of oxytocin. A sensitivity analysis, in which the impact of different prior distributions are assessed, is an important step in evaluating the impact of prior distribution specifications (Kruschke, 2021; Wagenmakers et al., 2018). As previously demonstrated using oxytocin administration data, the prior distribution can influence Bayes factor calculations (Quintana and Williams, 2017), with the degree of this influence depending on several factors, such as sample size (Kruschke, 2021). However, if prior data is scarce then a wider prior distribution—but not too wide (Tendeiro and Kiers, 2019)—is recommended to reflect this uncertainty, for which the posterior is generally less sensitive to the prior (Kruschke, 2021).

3.2. The ability to falsify hypotheses

A key premise of the Popperian hypothetico-deductive framework is that hypotheses have the opportunity to be falsified (Fidler et al., 2018). As outlined above, a solid derivation chain and strong assumptions are required to falsify hypotheses (Scheel et al., 2020). Despite the dominance of the Popperian hypothetico-deductive framework, the conventional application of frequentist NHST cannot falsify hypotheses by providing evidence for the absence of an effect (Lakens et al., 2020). Rather, conventional null hypothesis significance testing can only reject a null hypothesis, with the decision to reject a null hypothesis based on a specified alpha threshold (typically \( p = 0.05 \)). A non-significant \( p \)-value is not very informative, as this can be attributed to either data insensitivity or support for a null hypothesis (Dienes, 2014).

There are two statistical inference approaches that can be used to provide evidence for the absence of an effect that can facilitate hypothesis falsification that are growing in popularity in the biobehavioral sciences: Equivalence testing (Lakens, 2017; Schuirmann, 1987) and Bayesian hypothesis testing (Jeffreys, 1961; Wagenmakers et al., 2018), which was introduced above. When performing equivalence testing, a set of equivalence bounds are specified, which represent the smallest effect that is considered meaningful, and then a two one-sided tests procedure (Schuirmann, 1987) can be used to reject the presence of the smallest meaningful effect (also known as the smallest effect size of interest). Hence, this tool can help facilitate the decision of whether the data were consistent with the absence of an effect or whether the data were too insensitive. Applying equivalence testing to a representative body of published oxytocin research revealed that most studies reporting a non-significant result cannot provide evidence for the absence of effects (assuming that oxytocin’s effects are small), suggesting that the sample sizes in these studies were not large enough to generate informative conclusions (Quintana, 2018; Tabak et al., 2019). However, it is important to note that setting equivalence bounds is not necessarily straightforward as the smallest effect size of interest may not be immediately clear. For suggestions on determining a smallest effect size of interest, see Lakens et al. (2018). As mentioned previously, Bayesian hypothesis testing using Bayes factors can used to quantify the evidence for an alternative hypothesis, relative to a null hypothesis. Using the same logic, researchers can also quantify the evidence for a null hypothesis, relative to an alternative hypothesis, thus providing the opportunity to falsify hypotheses (Wagenmakers et al., 2018). Research has demonstrated how Bayesian hypothesis testing can be applied to oxytocin research to quantify evidence for a null hypothesis and used to complement traditional frequentist inference (Quintana and Williams, 2017; Tabak et al., 2019).

Frequentist NHST and Bayesian hypothesis testing rely on different assumptions, which limit how they can be applied to make statistical inferences. To help triangulate data models and broaden statistical inference, researchers can present both frequentist and Bayesian approaches (Harms and Lakens, 2018; Quintana and Williams, 2017). Moreover, both approaches can be used to falsify hypothesis. Including statistical tools that can reject hypotheses will help strengthen the evaluation of current theories, as conventional statistical inference tools cannot provide evidence for the absence of an effect, which is a key tenet of the Popperian hypothetico-deductive framework. Whereas equivalence testing and Bayesian hypothesis testing have been historically inaccessible to most users as they have been excluded in popular statistical software packages, these tools have been included in more recently released point-and-click packages that are free to download and
facilitate the sharing of analysis scripts (e.g., JASP: https://jasp-stats.org/ and JAMOVI: https://www.jamovi.org/).

### 3.3. The range of effect sizes that can be reliably detected

There are various ways to justify required sample sizes (Lakens, 2021b). No matter which approach is chosen, researchers are making an assumption, even if implicit, regarding the range of effect sizes that they are interested in reliably detecting. For example, a within-participants intranasal oxytocin study with 100 participants could reliability detect (i.e., 80% power) an effect size of at least 0.28 (with an alpha of 0.05). Put another way, this experimental design assumes that an effect size of 0.28, or larger, is of practical or theoretical interest, and that smaller effects sizes are not of interest as they cannot be reliably detected. A highly-cited study published in 2016 reported that oxytocin research tends to be statistically underpowered (Walum et al., 2016), which means that most studies cannot reliably detect a large range of effects. A 2020 study concluded that there has not been much improvement in the field in terms of appropriately powering studies (Quintana, 2020), although there have been some more recent exceptions (Decker et al., 2020; Kapetaniou et al., 2021; Zhang et al., 2019).

An underappreciated consequence of studies that cannot detect a wide range of effect sizes, is that these studies also cannot reject a wide range of effects, using the tools described above. Researchers reporting their “file-drawn” studies provide a critical contribution to the literature (e.g., Lane et al., 2016). However, appropriate statistical power is just as important for non-significant studies as significant studies (Lakens, 2017; Quintana, 2018). Non-significant studies with low statistical power cannot reject a wide range of effect sizes. For example, consider a study using an independent samples t-test to compare the effects of intranasal oxytocin or placebo, with 20 participants per group and an alpha of 0.05. This test design could only reject effects as large, or larger, than 0.93 using an equivalence test. In fact, a recent analysis reported that only 15% of a sample of non-significant oxytocin findings (i.e., 4 out of 26) could reliably reject findings ≥ 0.2, which is higher than the median reported effect size of 0.14 (Quintana, 2020) without even accounting for the effect size inflation of published studies that have not been pre-registered (Sächer and Schwarz, 2019). Using effect size benchmarks (e.g., Cohen, 1988) for setting the smallest effect of interest can be problematic as an effect size of 0.2 is not universally “small” across all fields. While such benchmarks can be used as a starting point when the expected effect sizes for a given field are unclear, researchers should justify their smallest effect size of interest using more informed approaches if possible (e.g., prior research) or simply state that this was dictated by resource limitations (Lakens et al., 2018). Regardless, reporting the smallest effect size that could be reliably detected for a given study design will help readers better determine the evidential value of reported results.

### 4. Conclusions and future directions

Human oxytocin research has been characterized by results that have not consistently aligned with expectations (Alvares et al., 2017). The promise of early animal work made way to disappointment when many of these results did not successfully translate to human participants. Pre-registration has been put forward as a remedy for poor reproducibility (Munafò et al., 2017), but despite its benefits preregistration is not necessarily straightforward (Nosek et al., 2019) as specifying precise predictions for a study requires the careful consideration of methodology and theory (Winterton et al., 2021b). If researchers find it difficult to specify hypotheses, then they are probably not ready for hypothesis testing (Scheel et al., 2020) and should instead focus on strengthening the evidence for auxiliary assumptions, starting with the weakest link of a derivation chain (Scheel et al., 2020). While the field would certainly benefit with a greater focus on the establishment of auxiliary assumptions before primary hypothesis testing, some auxiliary assumptions can be tested in parallel with direct hypothesis testing. For example, a study could evaluate a range of auxiliary hypotheses (e.g., administering a range of doses, demonstrating the reliability of peripheral oxytocin concentrations with repeated measures) while also assessing the primary hypothesis.

Scientists have two choices when faced with a negative result: maintaining that a theory is correct and placing the blame on faulty auxiliary assumptions or rejecting the theory (Duhem, 1991). In other words, weak auxiliary assumptions can be used to shield theories from falsification. As impregnable theories cannot advance fields, it is critical to examine the validity of auxiliary assumptions. This review has not been an exhaustive overview of auxiliary assumptions as many are either unique to specific studies or exceptionally broad assumptions such as the validity of measurement tools (Flake and Fried, 2020) and if statistical inferences can reliably generalize to unformalized verbal hypotheses (Yarkoni, 2020), that have received treatment elsewhere. Moreover, a comprehensive assessment for the validity of each assumption has not been provided as this was beyond the scope of the article. Rather, it is hoped that this discussion will motivate researchers in the oxytocin field to consider the auxiliary hypotheses for their own studies, which will make hypothesis tests more informative (Scheel et al., 2020) and ultimately provide a better understanding oxytocin’s role in health and wellbeing.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

This work was supported by the Research Council of Norway (301767) and the Novo Nordisk Foundation (NNF16OC0019856). A preprint of an earlier version of this manuscript has been posted to the PsyArXiv preprint server (https://psyarxiv.com/48ayg/).

### References

Alvares, G.A., Quintana, D.S., Whitehouse, A.J., 2017. Beyond the hype and hope: critical considerations for intranasal oxytocin research in autism spectrum disorder. Autism Res. 10, 25–41. https://doi.org/10.1002/aur.1992.

Anderson, S.F., 2020. Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. Psychol. Methods 25, 596–609. https://doi.org/10.1037/met0000248.

Arletti, R., Benelli, A., Bertolini, A., 1989. Influence of oxytocin on feeding behavior in the rat. Peptides 10, 89–93. https://doi.org/10.1016/0196-9781(89)90082-X.

Badenes-Ribera, L., Frías-Navarro, D., Iotti, B., Bonilla-Campos, A., Longobardi, C., 2016. Misconceptions of the p-value among Chilean and Italian Academic Psychologists. Front. Psychol. 7, 1247. https://doi.org/10.3389/fpsyg.2016.01247.

Bakker-Huvenaars, M.J., Greven, C.U., Herpers, P., Wiegers, E., Jansen, A., van der Steen, R., van Herwaarden, A.E., Baanders, A.N., Nijhof, K.S., Scheepers, F., Rommelse, N., Glennon, J.C., Buitelaar, J.K., 2020. Saliva oxytocin, cortisol, and testosterone levels in adolescent boys with autism spectrum disorder, oppositional defiant disorder/conduct disorder and typically developing individuals. Eur. Neuropsychopharmacol. 30, 87–101. https://doi.org/10.1016/j.euroneuro.2018.07.097.

Bales, K.L., Carter, C.S., 2003. Sex differences and developmental effects of oxytocin on aggression and social behavior in prairie voles (Microtus ochrogaster). Horm. Behav. 44, 178–184. https://doi.org/10.1016/S0018-506X(03)00154-5.

Bartz, J., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., Vicens, V., Hollander, E., 2010. Oxytocin can hinder trust and cooperation in borderline personality disorder. Soc. Cogn. Affect. Neurosci. https://doi.org/10.1093/scan/nsw085.

Beard, R., Singh, N., Grundsober, C., Gee, A.D., Tate, E.W., 2018. High-yielding 18 F radiosynthesis of a novel oxytocin receptor tracer, a probe for nose-to-brain oxytocin uptake in vivo. Chem. Commun. 54, 8120–8123. https://doi.org/10.1039/C8CC01400K.

Beets, I., Janssen, T., Meelkop, E., Temmerman, L., Suetens, N., Rademakers, S., Jansen, G., Schoofs, L., 2012. Vasopressin/oxytocin-related signaling regulates gustatory associative learning in C. elegans. Science 338, 543–545. https://doi.org/10.1126/science.1236806.
of oral and intranasal oxytocin on arousal and reward during social processing. Transl Psychiatry 11, 1–11. https://doi.org/10.1038/s41398-021-01241-w.

Knuutik, J. K., 2021. Report their research guidelines. Nat. Hum. Behav. 1–10. https://doi.org/10.1038/s41593-021-01177-7.

Lakatos, I., 1978a. Falsification and the methodology of scientific research programmes. In: Curie, G., Worrall, J. (Eds.), The Methodology of Scientific Research Programmes: Philosophical Papers. Cambridge University Press, Cambridge, pp. 8–101. https://doi.org/10.1017/CBO9780511621223.003.

Lakatos, I., Zahr, E., 1978b. Why did Copernicus’s research programme supersede Ptolemy? In: Curie, G., Worrall, J. (Eds.), The Methodology of Scientific Research Programmes: Philosophical Papers. Cambridge University Press, Cambridge, pp. 168–192. https://doi.org/10.1017/CBO9780511621223.006.

Lakens, D., 2021a. The practical alternative to the p value is the correctly used p value. Psychon Bull Rev. 28, 301–310. https://doi.org/10.3758/s13423-020-01924-9.

Lakens, D., 2021b. Sample size justification. PsyArXiv. https://doi.org/10.31234/osf.io/9dy3f.

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.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00692.

Lakens, D., McInnes, K., Isager, P.M., Scheel, A.M., Dienes, Z., 2020. Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. J. Gerontol. Ser. B 75, 45–57. https://doi.org/10.1093/geronb/gby065.

Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological research: a tutorial. Adv. Methods Pract. Psychol. Sci. 1, 259–269. https://doi.org/10.1177/2475288318754993.

Lee, M.R., Shnitko, T.A., Blue, S.W., Kaucher, A.V., Winchell, A.J., Erikson, D.W., Quintana, D.S., Westlye, L.T., Smerud, K.T., Mahmoud, R.A., Andreassen, O.A., 2016. Nasal administration of oxytocin results in detectable plasma levels: a dose–response mechanistic account of intranasal oxytocin pharmacodynamics in the brain. J. Neuroendocr. 28. https://doi.org/10.1111/jne.12384.

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. 79, 243–250. https://doi.org/10.1016/j.neubiorev.2017.08.008.

Lakens, D., Ludwig, M., Westlye, L.T., Mahmoud, R.A., Andreassen, O.A., 2016. Intranasal oxytocin: myosin mutations. Biol. Psychiatry 79, 243–250. https://doi.org/10.1016/j.biopsych.2015.05.003.

Lakens, D., Sabatier, N., 2016. Measuring oxytocin and vasopressin: bioassays, immunomodulators and random numbers. J. Neuroendocr. 28. https://doi.org/10.1111/jne.13245.

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. 79, 243–250. https://doi.org/10.1016/j.neubiorev.2017.08.008.

Loken, J.J., Thorne, R.G., 2012. Intranasal delivery of biologics to the central nervous system: overcoming the file drawer of one null result. Perspect. Psychol. Sci. 16, 639–648. https://doi.org/10.1177/1745691612455984.

Lochhead, J.J., Thorne, R.G., 2012. Intranasal delivery of biologics to the central nervous system: overcoming the file drawer of one null result. Perspect. Psychol. Sci. 16, 639–648. https://doi.org/10.1177/1745691612455984.
