Psychoneuroendocrinology 87 (2018) 127–130

Contents lists available at ScienceDirect

Psychoneuroendocrinology

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

Short Communication

Revisiting non-significant effects of intranasal oxytocin using equivalence testing

Daniel S. Quintana*

NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, University of Oslo, Oslo University Hospital, Oslo, Norway

ARTICLE INFO

Keywords:
Oxytocin
Social cognition
Equivalence test

ABSTRACT

The effects of intranasal oxytocin on social cognition are mixed, with several non-significant reports casting some doubts on its efficacy. Nevertheless, drawing inferences from non-significant values is problematic as non-significant results can be indicative of either statistical equivalence or insensitive data. Equivalence tests can be used to assess evidence for statistical equivalence, which can consequently facilitate theory falsification. To improve the inference of non-significant NHST *p*-values, this paper reports a set of equivalence tests performed on data from a recent meta-analysis synthesizing 32 intranasal oxytocin studies. Data from 26.1% of non-significant meta-analytic effects were indicative of data insensitivity, rather than statistical equivalence. Equivalence tests were also performed on a set of previously unpublished data from one laboratory, to examine whether unpublished data yields similar outcomes. Of the 34 non-significant effects, 73.5% were due to data insensitivity. As these analyses illustrate how non-significant intranasal oxytocin results may not necessarily support the absence of an effect, researchers are encouraged to implement equivalence tests in the design of their studies. By facilitating theory falsification, the adoption of equivalence tests can advance the field by redirecting resources to more promising avenues of research.

1. Introduction

The neuropeptide oxytocin is critically involved in mammalian social behavior (Donaldson and Young, 2008), with preclinical work contributing to the burgeoning interest in the use of oxytocin administration to address social dysfunction in a range of psychiatric disorders. In spite of considerable early promise, a recent meta-analysis of intranasal oxytocin's effect on performance in a range of social cognition tasks yielded varied effects, with several non-significant summary effects, among a few significant effects. (Leppanen et al., 2017). While these non-significant effects may be discouraging for the field, drawing inferences from non-significant null hypothesis significance test (NHST) *p*-values can be problematic.

A NHST *p*-value is used as a criterion to either reject or not reject the null hypothesis. Rejecting the null hypothesis can be a useful approach to assess whether two groups vary on a given variable, however, researchers cannot make any inferences regarding the null hypothesis, no matter how large the *p*-value. Consequently, it is uncertain if non-significant *p*-values in the intranasal oxytocin literature are indicative of either the absence of a meaningful effect, or that the data were simply too insensitive to detect an effect. A related issue is that intranasal oxytocin studies contain tests that are statistically underpowered, in general (Walum et al., 2016). The aggregation of effect sizes via meta-analysis can improve statistical power (Cohn and Becker, 2003), but it is unknown if sufficient studies are available to perform appropriately powered meta-analyses on the social-cognitive effects of intranasal oxytocin. Publication bias is an intertwined issue, as the exclusion of unpublished non-significant results may bias meta-analyses that only aggregate published studies, which are more likely to be statistically significant. However, the extent to which unpublished non-significant studies support the absence of an effect or simply have insensitive data is also uncertain.

This article has two primary objectives. First, it aims to promote equivalence testing, which is an underused tool that can assess the evidence for the absence of meaningful effects and aid in sample size selection. Second, to improve the inference of previously reported non-significant intranasal oxytocin results, equivalence testing was applied to recent meta-analytic outcomes (Leppanen et al., 2017). Equivalence testing was also applied to a set of unpublished intranasal oxytocin outcomes (Lane et al., 2016) to examine whether unpublished data provides similar results to published meta-analytic data.

* Corresponding author.
E-mail address: daniel.quintana@medisin.uio.no.

http://dx.doi.org/10.1016/j.psyneuen.2017.10.010
Received 30 August 2017; Received in revised form 11 October 2017; Accepted 11 October 2017

2. Material and methods

Analyses were conducted using the R statistical environment (version 3.3.2), with data and scripts to reproduce the results available at https://osf.io/2h6j9. Although it is possible to assess support for a null hypothesis within a Bayesian framework (e.g., Bürkner et al., 2017), equivalence tests allow for the rejection of effects as large, or larger, than a smallest effect size of interest using a frequentist framework. Thus, the two one-sided test (TOST) procedure for equivalence testing was implemented (Hauck and Anderson, 1984; Lakens, 2017; Schuirmann, 1987), in which a significant *p*-value is indicative of statistical equivalence. A NHST *p*-value alone can only be used to provide support for a difference, whereas combining a NHST with the TOST procedure yields four possible outcomes for an effect: i) not equivalent and not different, ii) statistically equivalent and different, iii) Not equivalent and statistically different, and iv) statistically equivalent and not different (Fig. S1). The TOST procedure was performed for all primary and moderator analyses reported in Leppanen and colleagues' meta-analysis (Leppanen et al., 2017). There is no clear theoretical lower boundary for oxytocin's social-cognitive effects to determine the smallest effect size of interest (SESOI) for setting equivalence bounds. In lieu of this, the smallest effect size that can be detected with sufficient statistical power for a given a meta-analysis was used for setting equivalence test bounds (Lakens, 2017). For this analysis, 80% power was deemed to be sufficient. For the meta-analytic effects, statistical power was calculated using formulas from Valentine et al. (2010). Equivalence tests were performed using the TOSTER R package (version 0.2.3; Lakens, 2017). Instead of using the default one-tailed assumption for the NHST meta-analysis, two-tailed tests were used for consistency with the original analyses.

For Lane and colleagues' paper containing previously unpublished research (2016), the smallest effect sizes that could be detected with 80% statistical power were calculated using the "pwr" R package (version 1.2-1; Champely, 2016). Equivalence tests were then performed using these equivalence bounds with the TOSTER R package. To control the family-wise error rate, significant α levels were Bonferroni corrected for the number of effect sizes reported from each study.

3. Results

The smallest effect sizes that each meta-analysis had 80% power to detect are presented in Table 1. On average, primary meta-analyses (n = 10) had 80% power to detect an effect size of at least $d = 0.26$ (range: $d = 0.15$ to $d = 0.4$) and moderator meta-analyses (n = 18) had 80% power to detect an effect size of at least $d = 0.43$ (range: $d = 0.17$ to $d = 0.94$). Across all 28 meta-analytic tests (i.e., both primary and moderator analyses), 23 were non-significant (Leppanen et al., 2017). Of these 23 non-significant tests, 6 were not statistically equivalent (26.1%), which is indicative of data insensitivity (Table 1; Fig. 1A). More specifically, eight out of ten of the primary meta-analyses were not statistically significant (Table 1; Fig. 1A). Of these eight non-significant primary meta-analyses, only the emotion sensitivity analysis was not statistically equivalent ($Z = 0.2$, $p = 0.42$) given a SESOI of $d = 0.16$. Of the 18 moderator analyses, 15 were not statistically significant. Of these 15 non-significant moderator analyses, five were not statistically equivalent: the clinical ($Z = 0.85$, $p = 0.2$; SESOI of $d = 0.35$) and healthy ($Z = -1.31$, $p = 0.1$; SESOI of $d = 0.28$) subgroup analyses for theory of mind performance, emotion recognition in the clinical group ($Z = -0.13$, $p = 0.45$; SESOI of $d = 0.29$), and the sensitivity to detect sadness ($Z = 1.13$, $p = 0.13$; SESOI of $d = 0.32$) and anger ($Z = 1, p = 0.16$; SESOI of $d = 0.32$). The original meta-analysis did not perform multiple test corrections for moderator analyses, which is common practice for meta-analyses in the biobehavioral sciences. Regardless, all equivalence tests were still statistically significant after Bonferroni correction of critical α values.

For the eight unpublished studies reported by Lane et al. (2016), the average smallest effect size detectable with 80% power was $d = 0.75$ (range: $d = 0.58$ to $d = 0.89$; Table S1). Of the 34 out of 35 non-significant results, 25 of these tests (73.5%) were not statistically equivalent (Table S1; Fig. 1B).

Table 1
Meta-analyses and their corresponding equivalence tests.

Social-cognitive domain		SESOI	Meta-analysis		Equivalence test	
			Z	p	Z	p
ToM	Healthy	0.28	0.73	0.47	−1.31	0.1
	Clinical	0.35	−0.21	0.83	0.85	0.2
	Overall	0.22	1.18	0.24	−1.7	0.04
Recognition	Happiness (healthy)	0.34	1.23	0.22	−2.94	0.002
	Happiness (clinical)	0.94	−0.21	0.83	2.64	0.004
	Happiness (overall)	0.31	1.05	0.3	−3.01	0.001
	Fear (healthy)	0.22	2.69	0.007	0.22	0.59
	Fear (clinical)	0.52	1.05	0.3	−2.35	0.009
	Fear (overall)	0.2	3.05	0.002	0.15	0.56
	Anger (healthy)	0.2	0.78	0.43	−3.1	0.001
	Anger (clinical)	0.52	1.11	0.27	−2.29	0.01
	Anger (overall)	0.18	0.98	0.33	−2.55	0.005
	Surprise (overall)	0.4	−0.19	0.85	3.63	< 0.001
	Sadness (overall)	0.23	0.58	0.56	−2.76	0.003
	Disgust (overall)	0.4	1.72	0.09	−2.1	0.02
	Overall (healthy)	0.17	2.32	0.02	−0.71	0.24
	Overall (clinical)	0.29	1.79	0.07	−0.13	0.45
	Overall	0.15	3.07	0.002	0.51	0.7
Sensitivity	Anger	0.32	−1.67	0.1	1	0.16
	Fear	0.32	−0.36	0.72	3.44	< 0.001
	Sadness	0.32	−1.28	0.2	1.13	0.13
	Happiness	0.32	−1.2	0.23	2.29	0.01
	Overall	0.16	−1.41	0.16	0.2	0.42
Expression	Clinical negative	0.7	−0.36	0.72	2.76	0.003
	Clinical positive	0.7	0.17	0.87	−5.67	< 0.001
	Healthy negative	0.58	0.53	0.6	−2.54	0.005
	Healthy positive	0.58	2.28	0.02	−3.01	0.001
	Overall	0.31	0.76	0.45	−2.61	0.004

Note: ToM = Theory of mind; SESOI = Smallest effect size of interest.

4. Discussion

Equivalence testing is a straightforward procedure that improves the inference of non-significant NHST *p*-values. Here, equivalence testing suggested that 26.1% of non-significant meta-analytic findings for the interpretation and expression of emotions after intranasal oxytocin were due to data insensitivity, rather than statistical equivalence between groups. For the unpublished studies, 73.5% of non-significant findings were due to data insensitivity rather than statistical equivalence. Altogether, these results reinforce both the utility of equivalence tests and how insufficient statistical power can obscure the interpretation of non-significant findings, as only the presence of large effects can be rejected.

Of the 43 individual studies included in the original meta-analysis, only six (7.2%) reported tests that had at least 80% statistical power to detect an effect size of $d = 0.5$ (Leppanen et al., 2017). In contrast, all ten of the primary meta-analyses had at least 80% statistical power or more to detect a medium effect size ($d = 0.5$), demonstrating how meta-analysis can increase statistical power for the inference of both significant and non-significant effects. For the moderator meta-analyses, 61.1% had at least 80% statistical power or more to detect an effect size of $d = 0.5$. While this demonstrates that underpowered moderator analyses are not necessarily widespread, researchers should be wary of whether moderator analyses are sufficiently powered.

The analysis of unpublished intranasal oxytocin data revealed that almost three-quarters of these non-significant findings could be attributed to data insensitivity, rather than statistical equivalence. Although the release of previously unpublished results by Lane et al. should be