Corresponding author: Gideon Nave

Assistant Professor of Marketing, Wharton School, University of Pennsylvania

Phone: (215) 898-8248 | Fax: (215) 898-2534 | Email: gnave@wharton.upenn.edu

https://marketing.wharton.upenn.edu/profile/gnave/

Word count (including acknowledgments): 2,106

Authors: Amos Nadler[a], Colin F. Camerer[g], David T. Zava[c], Triana L. Ortiz[d], Neil V. Watson[e], Justin M. Carré[f], & Gideon Nave[b,1]

Affiliations:

[a] Department of Finance, Ivey Business School at Western University, 1255 Western Rd, London, ON, Canada N6G 0N1; anadler@ivey.ca

[b] Marketing Department, The Wharton School of the University of Pennsylvania, 3730 Walnut St, Philadelphia, PA, USA 19104; gnave@wharton.upenn.edu

[c] ZRT Laboratory, 8605 SW Creekside Place, Beaverton, OR, USA 97008; dzava@zrt.com

[d] Department of Psychology, Nipissing University, 100 College Drive, North Bay, ON, Canada, P1B 8L7; trianao@nipissingu.ca

[e] Department of Psychology, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6; nwatson@sfu.ca

[f] Department of Psychology, Nipissing University, 100 College Drive, North Bay, ON, Canada, P1B 8L7; justinca@nipissingu.ca

[f] Department of the Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, MC 228-77, Pasadena, CA, USA 91125; camerer@hss.caltech.edu

[1]To whom correspondence should be addressed. Email: gnave@wharton.upenn.edu

Title: Does testosterone impair men's cognitive empathy? Evidence from two large-scale randomized controlled trials

Short title/running head: Does testosterone impair men's cognitive empathy?

Abstract

The capacity to infer the mental states of others (known as "cognitive empathy") is essential for social interactions, and a well-known theory proposes that it is negatively affected by intrauterine testosterone exposure. Furthermore, previous studies reported that testosterone administration impaired cognitive empathy in healthy adults, and that a biomarker of prenatal testosterone exposure (finger digit ratios) moderated the effect. However, empirical support for the relationship has relied on small-sample studies with mixed evidence. We investigate the reliability and generalizability of the relationship in two large-scale double-blind placebo-controlled experiments in young men (N=243 and N=400), using two different testosterone administration protocols. We find no evidence that cognitive empathy is impaired by testosterone administration or associated with digit ratios. With an unprecedented combined sample size, these results counter current theories and previous high-profile reports, and demonstrate that previous investigations of this topic have been statistically underpowered.

Key words: testosterone, cognitive empathy, mind reading, prenatal priming, steroid hormones, pharmacology

There is a growing scientific focus on the biological basis of social aptitudes and investigating the causes of their deficits (Ferguson, Young, & Insel, 2002). One important element of social cognition is "cognitive empathy", that constitutes the capacity to infer the emotions, beliefs, and goals of others. Cognitive empathy impairments characterize a broad range of psychopathological conditions, and are a part of the clinical diagnostic criteria for autism spectrum disorders (ASDs) (American Psychiatric Association, 2013).[1]

Testosterone-based biological theory of social cognition

A popular biopsychological model known as the Extreme Male Brain (EMB) hypothesis (Baron-Cohen, 2002) proposes that two distinct cognitive styles, "systemizing" and "empathizing", typify males and females respectively. The stereotypically male systemizing domain has no social dimension, and in its extreme form social cognition is fully diminished. Guided by observations that ASDs emerge early in life and are substantially more prevalent among males[2], and that males typically score lower than females in tests of cognitive empathy (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), the EMB hypothesis proposes that elevated prenatal exposure to the sex steroid testosterone (T) causes impairments in cognitive empathy, through its masculinizing effect on the developing brain (Zuloaga, Puts, Jordan, & Breedlove, 2008).

The EMB hypothesis found evidential support in a study that reported a correlation between amniotic T levels and ASD traits ((Auyeung et al., 2009), though see (Kung et al., 2016)), and has remained popular yet controversial to date. Much of its research has relied on the assumption that the ratio between the hand's second (index) to fourth (ring) digit (known as 2D:4D) is a

[1] The DSM V criteria for ASDs include "Non-verbal communication problems, such as abnormal eye contact, posture, facial expressions, tone of voice and gestures, as well as an inability to understand these."
[2] ASD incidence rates vary widely by study, from 5.2 to 72.6 per 10,000 people and ratios range from 1.81 to 15.7 male:female.

developmental proxy for prenatal T exposure (Manning, 2002), which motivated examinations of correlations between 2D:4D and cognitive empathy and ASDs occurrences. While some studies provided supporting evidence (e.g., (Lutchmaya, Baron-Cohen, Raggatt, Knickmeyer, & Manning, 2004)), several others failed to detect a relationship (Guyatt, Heron, Le Cornu, Golding, & Rai, 2015; Hönekopp, 2012). Moreover, because it is not feasible to experimentally manipulate prenatal T exposure in humans (due to ethical considerations), findings along this line of research have been correlational, which cannot establish causal relations (Auyeung et al., 2009).

Testing testosterone's causal effect on cognitive empathy

A handful of experiments (see Table 1) attempted to address the above limitation by testing the effects of T administration on cognitive empathy in neurotypical adults, and investigating the dependency of these effects on the 2D:4D biomarker (Bos et al., 2016; Carré et al., 2015; Olsson, Kopsida, Sorjonen, & Savic, 2016; van Honk et al., 2011). This line of research critically relies on an assumption originating in animal research, that *in utero* androgen exposure moderates the activational effect of T (Zheng & Cohn, 2011). The seminal publication along this line of research reported a placebo-controlled within-subject experiment of 16 healthy females, in which exogenously administered T strongly impaired cognitive empathy measured using the "Reading the Mind in the Eyes Test" (RMET), a 36-item battery testing the ability to infer others' emotional states and intentions from pictures of their eye regions (Baron-Cohen et al., 2001) (see Fig. S2 in Supplementary Material available online for example item). More than 50% of the individual differences in the effect of exogenous T on the RMET were explained by the participants' variation in the right-hand 2D:4D, implying involvement of prenatal T exposure in the causal effect (van Honk et al., 2011).

A similar experiment with roughly twice the sample size (*N*=33) found a much smaller[3] effect (*P*=0.048, one-tailed), and no moderation by 2D:4D (Olsson et al., 2016). A third experiment of 16 females found neither a main effect nor a moderation by 2D:4D (Bos et al., 2016). Last, one experiment investigated the effect of T administration on the RMET in 30 healthy males and found neither a main effect nor a moderation by the right-hand 2D:4D; however, analysis revealed that T administration reduced cognitive empathy in participants with relatively low (but not high) left-hand 2D:4D (Carré et al., 2015).

Do the data support the hypothesis?

While these initial findings seem promising, they are subjected to important limitations. First, only one of the four studies rejected the null hypothesis with a 95% confidence for a main effect of T on the RMET (van Honk et al., 2011). Moreover, the seminal publication's report of a strong moderating role of the right-hand 2D:4D was not replicated in any of the other studies (see Table 1).[4]

A second concern is statistical power. Although the RMET is a noisy psychological instrument,[5] and 2D:4D is, at best, a noisy proxy of prenatal T exposure (Berenbaum, Bryk, Nowak, Quigley, & Moffat, 2009), all samples ranged between 16 and 33 participants, which might have been too meager to credibly estimate a true effect size. It is therefore impossible to know whether the inconsistencies in the literature are due to an absence of a true association, or the result of false

[3] The primary publication (van Honk et al. 2011) had a statistical power of only 0.26 to detect the effect size found in the similar study with twice the sample size (Olsson et al. 2016).

[4] One experiment (in males) reported a statistically significant moderating effect, but only for the left-hand 2D:4D; the two other experiments reported no moderation of the 2D:4D (Bos et al., 2016).

[5] The RMET has a test-retest reliability of 0.7 (Baron-Cohen et al., 2001).

negative findings due to low statistical power. Thus, these inconsistent results necessitate clarification through additional studies.

To this end, we conducted a powerful direct test of the activational and developmental effects of T on cognitive empathy in two studies of healthy young men with greater than 15 and 25 times the sample size conducted in females (van Honk et al. 2011) and greater than 7 and 12 times the largest experiment in males (Carré et al., 2015), respectively, constituting the two largest behavioral T administration experiments conducted to date. In both studies we used a computer-based version of the RMET to test the hypothesis that T administration and its purported developmental biomarkers affects cognitive empathy.

Methods

Experiment 1

Participants and experimental procedure. Two hundred forty-three males aged 18 to 55 (M=23.63, SD=7.22) participated in the study and were mostly private Southern California consortium students from diverse ethnic backgrounds (see *Participants* section and Table S1a in Supplementary Material). The institutional review boards of Caltech and Claremont Graduate University approved the study, all participants gave informed consent, and no adverse events occurred. All data and materials are available on the Open Science Framework (https://osf.io/hztfe/).

Participants registered by their preferred session dates and were added to cohorts of 13 to 16. They arrived at the lab at 9 a.m., signed informed consent forms, and had both of their hands scanned before being randomly assigned to private cubicles where they completed demographic and mood questionnaires (see Supplementary Material for all independent variables) and

provided an initial saliva sample by passive drool. Next, participants proceeded to gel application (further details below), after which they were given printed material containing precautions and instructions prior to dismissal (experimental timeline shown in Fig. 1). All participants returned to the lab at 2 p.m., provided a second saliva sample, and began a battery of tasks which lasted approximately two hours. We did not randomize the order of the behavioral tasks, in similar fashion to previous studies (Zethraeus et al., 2009), to standardize hormonal measurements (which have diurnal cycles) among participants. Following the experiment, participants completed an exit survey, where they indicated their beliefs about the treatment they had received using a five-point scale.

Treatment administration. Following initial intake, participants were escorted in groups of two to six to a semiprivate room. There they were provided *en masse* small plastic cups containing either 10 g of topical T that is a widely prescribed transdermal T gel with clearly mapped pharmacokinetics (Eisenegger, von Eckardstein, Fehr, & von Eckardstein, 2013) (100 mg, Vogelxo™, *N*=123) or volume equivalent of inert placebo of similar viscosity and texture placebo (80% alcogel, 20% Versagel®, *N*=118) under a double-blind protocol (see Fig. S1a in *Supplementary Material* for randomization protocol). Participants were instructed to apply the entirety of the gel container following manufacturer instructions.

Saliva samples. Each participant provided four passive drool saliva samples (upon arrival prior to treatment administration, shortly after returning for afternoon session, another closely following the RMET, and a final sample prior to exit survey) for subsequent assay (see Supplemental Material for precise timing). To allow robust manipulation checks and obtain statistical control for hormonal markers of participants' biological states, we used liquid chromatography tandem mass spectrometry (LC-MS/MS, detection levels and precision are

available in Supplementary Material Table S2) to measure the following salivary steroids: estrone, estradiol, estriol, T, androstenedione, DHEA, 5-alpha DHT, progesterone, 17OH-progesterone, 11-deoxycortisol, cortisol, cortisone, and corticosterone (see Supplemental Material Table S7 for measurements).

Digit Ratio Measurements. Participants' hand scans acquired at intake were measured by two independent raters using and digital calipers to quantify 2D:4D; inter-rater correlation was 0.96 and their scores were averaged.

Behavioral task. We administered the adult version of the RMET developed by Baron-Cohen and colleagues (Baron-Cohen et al., 2001) which shows the eye region of an actor's face, and a list of four words that describe emotional states and cognitive processes among which participants select the one that best described the person in the image (see Fig. S2 in the Supplemental Material for example item). The task was divided into two segments, baseline (morning) and post-treatment (afternoon), in a repeated measures design (see Fig. 1 below): each participant completed a half of the RMET in the morning (either part "A", items 1-18, or part "B", items 19-36) prior to receiving treatment, and the other half following treatment when, according to published pharmacokinetics, androgen levels are elevated and stable following exogenous application.

Psychological Questionnaires. We measured mood using the PANAS-X scale (Watson & Clark, 1999), both pre- and post-treatment (see Table S1a in Supplementary Material for aggregated responses).

Experiment 2

Participants. Experiment 2 included both students and general public for a total of 400

participants (M=22.80, SD=4.68). The sample was composed of males with low ethnic heterogeneity representative of the region (see *Participants* section and Table S1b in Supplementary Material). All accepted participants completed the task and were included in the analysis (For pre-screening criteria see Supplementary Material). The Nipissing University Research Ethics Board approved this study and no harms or adverse events occurred.

Experimental procedure. Participants arrived at one of three testing session times (10:00 a.m., 12:30 p.m., or 2:30 p.m.) in cohorts of six and brought individually into private testing rooms to read and sign an informed consent form, receive a participant number, and complete questionnaires (see Supplementary Material for all independent variables). Afterwards, participants provided a 1-2 ml saliva sample before treatment administration, after which they had their photos taken and hands scanned. Approximately two hours after arrival to the lab and one-and-a-half hours after drug administration participants completed the RMET then provided their final saliva sample. Upon session completion, participants received compensation and completed an exit survey asking which treatment they believed they had received (see Figure S1b in Supplementary Material).

Treatment administration. Following initial saliva sample collection, a researcher provided two syringes pre-filled by a pharmacist following a double-blind protocol each containing 5.5 mg of either placebo or T gel (for a total of 11 mg). This is a newly-approved nasal gel used for the treatment of hypogonadism. Pharmacokinetic data indicate that serum T concentrations rise sharply within 15 mins of testosterone gel application and remain elevated (relative to placebo) up to 180 mins post-application (Geniole SN, Procyhyn TL, Marley N, Or...). The gel was either Natesto® or the volume equivalent of an inert placebo of similar viscosity and texture. Random assignment was determined such that half the participants in every group received T and half

received placebo such that total participants base was bisected with *N*=200 for both T and placebo groups (see Fig. S1b in Supplementary Material for randomization protocol).

Saliva samples. Each participant provided two saliva samples, with the first sample collection time 30 minutes following arrival (and prior to gel application), and the second 120 minutes after arrival. Participants provided passive drool into a 5 ml polystyrene tube while situated in their individual testing rooms as instructed by a research assistant and samples were analyzed for pre- and post-treatment T and baseline cortisol using commercially available enzyme immunoassay kits (DRG International) (see Table S2b in Supplementary Material for hormone measures).

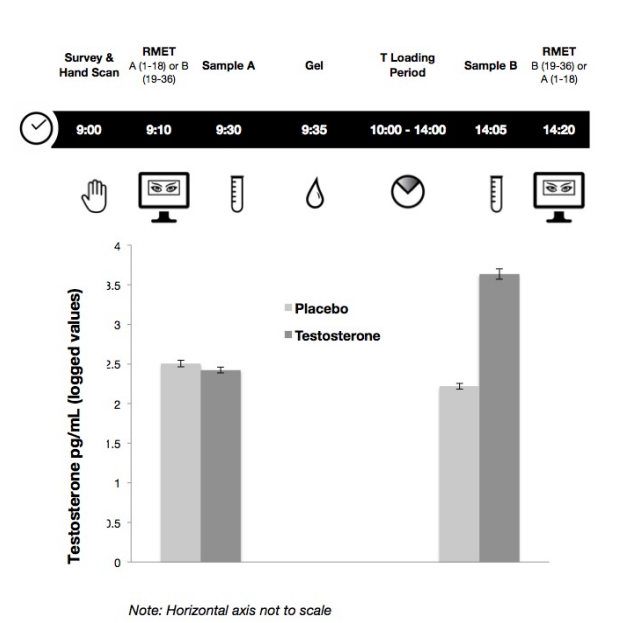
Digit Ratio Measurements. Participants' 2D:4D were measured by two independent raters using hand scans and digital calipers with an inter-rater correlation of 0.86.

Behavioral task. Participants evaluated all 36 items of the RMET (Baron-Cohen et al., 2001) as a single task.

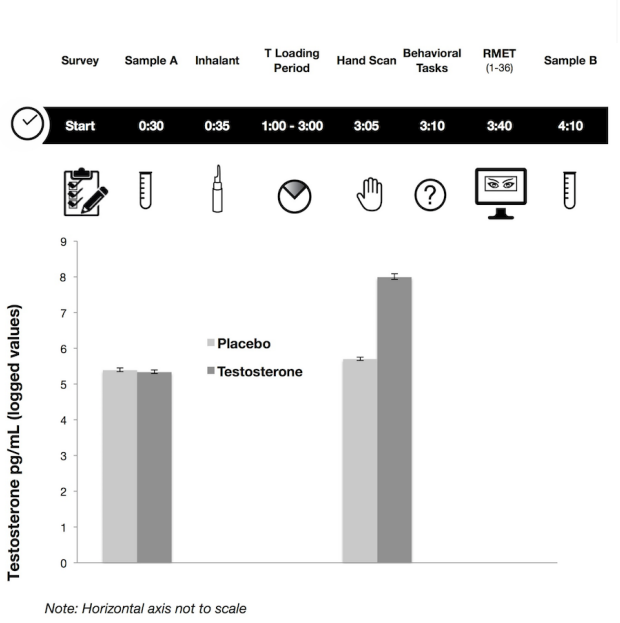
Sample size determination. Power analysis based on a statistical power of 0.80 stipulates that detecting the results reported in van Honk et al. (2011) would require, at minimum, a sample size of approximately 126 participants as this study reported an estimated change in scores from 80.8% (*SD* = 7.3) in placebo condition to 76.8 (*SD* = 8.6) (pooled SD = 7.98) in T condition (Cohen's *d* = 0.50). The corresponding power is 0.29; replicating an experiment with 0.29 power yields a 0.085 probability of detecting a veritable effect.

Figure 1. Experimental Timelines

a. Experiment 1	b. Experiment 2



Time-coded experimental timeline shows that, following intake (morning), participants completed half of the RMET (portion A or B), provided a saliva sample, and received gel prior to being dismissed. Upon their return to the lab (afternoon), another saliva sample was collected prior to taking the second portion of the RMET (B or A) (Standard errors shown).



Timeline generalizes experimental sequence for all three start times for Experiment 2. Following arrival and completion of consent and self-report questionnaires, participants provided a saliva sample approximately 30 minutes prior to drug administration. The second sample was collected at the end of a two-hour protocol approximately one-and-a-half hours after drug administration and fifteen minutes after the RMET.

Comparison of experimental features to van Honk et al. (2011)

We note several differences between our study and the primary positive report of T administration on the RMET in 16 females (van Honk et al., 2011).

Participant Sex. The EMB theory does not make any sex-specific predictions regarding the developmental and activational effects of T on cognitive empathy. However, van Honk et al. (2011) conducted their study in a female-only sample because the pharmacokinetics of a single dose of T had only been studied in females at the time (van Honk et al., 2011): *"We exclusively recruited women because the parameters … for inducing neurophysiological effects … are known in women but not in men."* (p. 3450). The recent pharmacokinetic mappings for short-term single dose T administrations (Carré et al., 2015; Eisenegger et al., 2013) and availability of

two unique administration modalities of FDA-approved exogenous T provided us with a reliable foundation for testing the EMB hypothesis in men.

Drug dosage and delivery. van Honk et al. (2011) used a sublingual T administration procedure, which causes a sharp increase in serum T of 10-fold or more within 15 minutes, with a rapid decline to normal levels within 90 minutes in women (Tuiten et al., 2000). It is important to note that the pharmacokinetic data for sublingual administration (published in Fig. 1 of Tuiten et al. (2000)) show that at the time the task was performed—4 hours after sublingual administration—participants' T levels were the same across the T and placebo groups. Moreover, the study that served as a justification for using a 4 hour delay had only 8 participants, and reported a statistically weak treatment effect (p=0.04, uncorrected for multiple comparisons).

In Experiment 1 we chose to administer T using topical gel as this was the only T administration method for which the pharmacokinetics of a single-dose had been investigated at the time (Eisenegger et al., 2013). That study demonstrated that plasma T levels peaked 3 hours after single-dose exogenous topical administration, and that T measurements stabilized at high levels during the time window between 4 and 7 hours following administration. Therefore, we had all participants return to the lab 4.5 hours after receiving gel, when androgen levels were elevated and stable. We used a 100 mg transdermal dose, which quickly elevates then holds T levels high and stable for approximately 24 hours (Swerdloff et al., 2000) and was shown to generate effects on cognition, decision making, and other behaviors (Nadler et al. 2017; Nave et al. 2017; Nave et al. 2018).

In Experiment 2 we used nasal delivery, following a recent study indicating that serum T concentrations rise sharply within 15 minutes after Natesto® gel application and remain elevated

for approximately three hours post application among hypogonadal males (Geniole SN, Procyhyn TL, Marley N, Or...) (see Fig. S3 in Supplementary Material). This method conforms to our experimental paradigm's pharmacokinetic structure as serum T approached its zenith in treated participants as they completed the RMET (see Supplementary Material Tables S2a and b for T levels).

The doses in both experiments are commonly prescribed daily to men with low circulating T levels and serve as two distinct physical transport channels (transdermal and intranasal, respectively). Although it is less than the 150 mg used in one study (Carré et al., 2015), T levels followed a similar pharmacokinetic arc (see Fig. 1). Various studies show significant heterogeneity in change in T levels depending on delivery method, location of application in the body, and biofluid measured.

Experimental designs. van Honk et al. (2011) used the same questions in pre- and post-treatment testing. As T treatment might affect participants' capacity to recall answers (Cherrier, Craft, & Matsumoto, 2003), this design choice might have introduced memory confounds,. In Experiment 1 we divided the RMET into two portions, and administered each portion as either pre- or post-treatment measure in, allowing us to capture baseline abilities while ruling out such confounds. In Experiment 2 we conducted a between-subjects experiment that removes all effects of practice and recall from the data.

Results

Manipulation check

T treatment resulted in higher saliva T values in the treatment groups compared with the placebo

groups in both experiments (see Fig. 1 and Tables S2a and S2b in Supplementary Material). Consistent with previous reports, we found no treatment effects on mood and treatment expectancy (e.g., (Eisenegger, Naef, Snozzi, Heinrichs, & Fehr, 2010)), or levels of other hormones unaffected by exogenous T, as measured by LC-MS/MS in Experiment 1 or enzyme immunoassay in Experiment 2.

Influence of T on RMET scores

Overall RMET scores in our samples were comparable with previous studies of similar populations (see Fig. 2A below). Figure 2B shows baseline and post-treatment RMET scores in Experiment 1, separated by treatment group and order. As expected, baseline (morning) RMET performance were reliably correlated with afternoon scores (r(241)=0.40, P<0.001). In addition, participants' scores were, on average, slightly higher in the B portion of the test (A portion average=13.54, SD=2.43; B portion average=13.95, SD=2.19, t(241)=2.53, P=0.01). Figure 2C shows Experiment 2 scores by treatment groups.

Figure 2. RMET distributions and scores across treatment groups

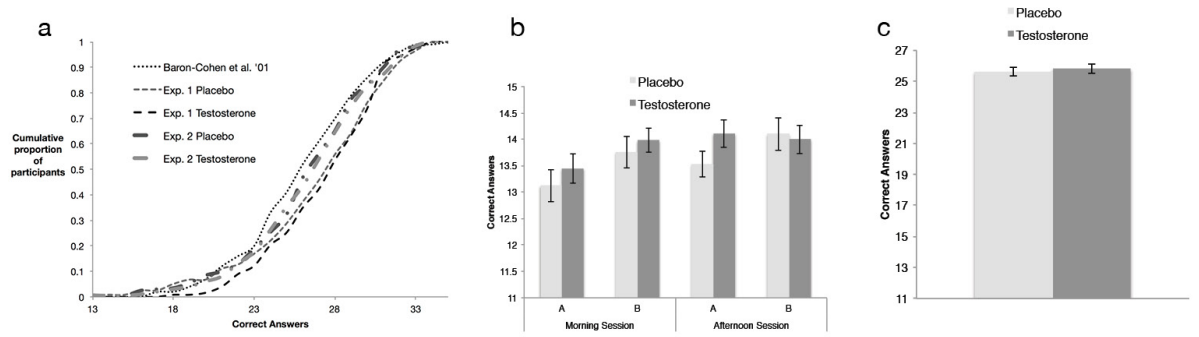


Fig. 2. (*A*) Cumulative distributions of scores from our experiments juxtaposed with the original paper (Baron-Cohen et al., 2001); scores for Experiment 1 portions A and B (both of which are unaffected by treatment) are

combined into a total score: RMET scores in Experiment 1 was 27.5 (*SD*=3.9), and 25.6 (*SD*=3.9) in Experiment 2, which are similar to male students in (Baron-Cohen et al., 2001) showing average score of 27.3 (*SD*=3.7). *(B)* Experiment 1 pre- and post-treatment RMET scores. No pre- or post-treatment differences were found between the two groups, regardless of the order in which portions of the tests were taken (standard errors shown). (*C*) Experiment 2 RMET scores. No differences were found between treatment groups.

To test for the main effect of T administration on cognitive empathy for Experiment 1, we estimated linear regression models with the post-treatment (afternoon) RMET score as the dependent variable, a binary treatment indicator (T=1, placebo=0) as the independent variable of primary interest, controlling for baseline performance, the order of the two portions of the RMET (A and B) and additional control variables[6] (the results remain unchanged when these control variables are excluded from the models; see Tables S3a and S4a in Supplementary Material). Analogously, Experiment 2 data were analyzed using linear models with total RMET score as the dependent variable, a binary treatment indicator as the chief independent variable of interest, and control variables[7] (results remain unchanged with their exclusion from the models; see Supplementary Material Tables S3b and S4b).

We found no reliable effect of T administration on the RMET in Experiment 1 (beta=0.11, 95% CI=[−0.45, 0.68]; $t(237)=0.37$, $P=0.71$; Cohen's $d=0.04$, 95% CI=[−0.19, 0.28]), and the effect's point estimate was positive. These results rule out a *negative* effect size of T administration on the RMET that is greater in magnitude than $d=0.19$ with 97.5% confidence. A sample of at least 870 participants (in a between-subject design), or 435 subjects (in a within-subject design), which is over 26 to 54 (within-subject 13 to 26) times greater than previous investigations, would

[6] These include RMET baseline scores, portion A or B, 2D:4D and treatment interactions, Cognitive Reflection Task (CRT) scores, math abilities, mood and affective measures, treatment expectancy, age, marital status, sexual preference, and all other measured hormones that were not influenced by T treatment. The CRT control was added because performance is impaired by exogenous T (Gideon Nave, Nadler, Zava, & Camerer, 2017) and people with ASD outperform non-ASD age-matched controls (Brosnan, Lewton, & Ashwin, 2016).

[7] These include Cognitive Reflection Task (CRT) scores, factor 1 and 2 psychopathy measures, treatment expectancy, age, marital status, sexual preference, and all other measured hormones that were not influenced by T treatment.

be required to reliably detect even this "optimistic" negative effect size estimate with statistical power of 0.8. Regression analyses with comprehensive controls corroborate the absence of a main treatment effect, and the absence of moderation by 2D:4D (right hand, left hand, and their average), as implied by insignificant interaction coefficients (see Supplementary Material, Table S3a). Furthermore, in an analysis analogous to the previous positive report (van Honk et al., 2011), we found no correlation between the treatment effect on the RMET and the right-hand 2D:4D in the T group ($r(123)=0.04$, $P=0.66$, 95% CI=(−0.138, 0.215)) (see Fig. S4 in the Supplementary Material).

Experiment 2, which had 400 participants, could also not reject the null hypothesis (beta=0.27, 95% CI=[−0.49, 1.02]; $t(398)=0.69$, $P=0.49$; Cohen's $d=0.04$, 95% CI=[−0.15, 0.24]) and there was no significant treatment effect in any regression model (Supplementary Material Tables S3b and S4b). Similarly, the point estimate of the effect in Experiment 2 was positive, and our results rule out a negative effect size of T administration on the RMET that is greater in magnitude than 0.15 with 97.5% confidence. A sample of at least 1,394 participants (in a between-subject design), or 697 subjects (in a within-subject, cross-over design), which is 42 to 87 times greater than previous investigations would be required to reliably detect the lower negative bound of the stated effect size.

Sample size estimations to reach ES boundary

Experiment 1. Requisite sample size for the negative effect size boundary of -0.19 found in our results is 870 for a between-subject design and 435 for within-subject design.

Experiment 2. Requisite sample size for the negative effect size boundary of -0.14 found in our results is 1,396 for a between-subject design and 698 for within-subject design.

Further analyses of each question in isolation using Chi-squared tests revealed no systematic differences between treatment conditions in any of the RMET items in both experimental datasets (see Supplementary Material Tables S5a and b).

Testing for effect of 2D:4D

Prenatal T proxies (the 2D:4D, either right-hand, left-hand, or their average) did not correlate with RMET scores in both experiments or moderate the effect of T administration, echoing other recent findings (Bos et al., 2016; Carré et al., 2015; Olsson et al., 2016) (see Supplementary Material Tables S3a-S4b). These results are in line with previous reports showing no correlation between the 2D:4D and RMET scores (Hönekopp, 2012; Masuya et al., 2015; Voracek & Dressler, 2006), and in contrast to the two papers reporting an interaction between 2D:4D and the exogenous T's effect on the RMET (Carré et al., 2015; van Honk et al., 2011). As noted above, the two latter reports are inconsistent with each other, with one reporting a moderation of the right-hand 2D:4D and the other reporting a moderation of only the left-hand 2D:4D.

Discussion

Our experiments used two notably large samples to test the effects of pharmacological T manipulation on cognitive empathy. Despite experimental differences between them, their collected data exhibit the same results with robust statistical consistency, to demonstrate a lack of effects of T administration and 2D:4D on cognitive empathy. These findings, and the literature as a whole, cast serious doubts on the proposal that T causally impairs cognitive empathy, for several reasons.

First, the low statistical power of previous investigations undermines their reliability in capturing true effects. Even if we assume that a purported size of T's negative effect on cognitive empathy

is the overly "optimistic" negative bound of our confidence intervals, we find that all previous investigations of the topic were statistically underpowered (< 0.3 power) (see Table 1 and Methods section).

Second, the results of the previous small sample studies' are discrepant. Our large samples draw on drastically more data than all previous investigations combined, and generalize across geographically, economically, and culturally distinct populations (see Participants section of Supplementary Material). Our use of two different experimental designs and T administration protocols across these populations further mitigates the concern that the outcomes were due to a particular experimental factor. Of note, there are some design differences between our studies and previous investigations (see Table 1; differences from van Honk et al. (2011) discussed above). However, even if those design differences led to a complete abolishment of a "real" effect of T on cognitive empathy, our results demonstrate beyond a reasonable doubt that such an effect is not generalizable.

Table 1. Summary of literature linking T administration and the RMET

Design	Sex	*N*	Design	Dose	Main effect	Effect Size	SE of ES	95 % CI for ES		2D:4D moderation
Nadler et al. Experiment 2	M	400	Within subject; no repetition	11 mg intranasal	No effect; *P* = 0.66	-0.04	0.26	-0.24	0.15	No
Nadler et al. Experiment 1	M	241	Between subject; no repetition	100 mg transdermal	No effect; *P* = 0.83	-0.03	0.36	-0.28	0.22	No
Olsson et al. (2016)	F	33	Within subject; repeated task	50 mg transdermal	One-tailed *P* = 0.048; *d* = 0.32	0.33	0.13	-0.15	0.82	No
Carré et al. (2015)	M	30	Within subject	150 mg transdermal	No effect; *P* = 0.25	-0.20	0.35	-0.70	0.31	Left hand
van Honk et al. (2011)	F	16	Within subject; repeated task	0.5 mg sublingual	One-tailed Wilcoxon *P* = 0.01; *d* = *0.54*	0.49	0.25	-0.21	1.19	Right hand
Bos et al. (2016)	F	16	Within subject; no repetition	0.5 mg sublingual	No effect; *P* = 0.78	0.10	0.10	-0.60	0.79	–

A third reason concerns the validity of the 2D:4D biomarker. The initial findings that prenatal T exposure correlates with 2D:4D are supported in non-clinical and clinical human populations (Lutchmaya et al., 2004), as well as in preliminary causal evidence in relative phalanx/tibia lengths in mice (Zheng & Cohn, 2011). However, recent work highlights concerns regarding the reliability of 2D:4D as a biomarker (Valla & Ceci, 2011). The 2D:4D of complete androgen insensitivity syndrome patients were found to be only somewhat feminized, and had the same variance as in healthy controls, demonstrating that the preponderance of individual differences in the measure is not attributable to the influence of T exposure (Berenbaum et al., 2009). There is also longitudinal evidence that 2D:4D systematically changes during childhood (McIntyre, Ellison, Lieberman, Demerath, & Towne, 2005), which is unconformable with the preposition that it accurately quantifies prenatal influences *per se*. Moreover, a study of the hunter-gatherer Hadza people reported an absence of 2D:4D distinction between males and females, suggesting that sex differences in the measure are not universal (Apicella, Tobolsky, Marlowe, & Miller, 2016).

Many reports of correlations between 2D:4D and behavioral traits hold only for subsets of the population (e.g., particular sex or race). Correlation sometimes holds only for the right hand 2D:4D but in other times only for the left hand, or for the average of both hands (Teatero & Netley, 2013). Overall, significant results are seldom replicated, and few survive correction for multiple comparisons or meta-analytic aggregations (e.g., (Turanovic, Pratt, & Piquero, 2017)). These concerns belie the validity of the measure as a biomarker and its capacity to detect reliable correlations with noisy psychological constructs in studies of small samples.

Despite our dissenting results, the absence of evidence is not necessarily evidence of absence. Specifically, the lack of an association between 2D:4D and cognitive empathy could be

attributable to the failure of the measure to serve as a reliable androgenic biomarker. We therefore agree with Baron-Cohen et al. (2011) (p. 6) that it is worthwhile to study the occurrence of impaired cognitive empathy and other ASD traits in developmentally unique populations. One such study reported mixed evidence of higher scores in some Autism Quotient self-rating subscales among women with congenital adrenal hyperlexia (CAH) and lower scores in other subscales, compared with their unaffected relatives (Knickmeyer et al., 2006), with no significant results in men. However, results along this line of research, too, are far from being conclusive. For example, (Kung et al., 2016) found that young females with and without CAH did not differ in autistic traits, and that amniotic T levels were not associated with scores from either sex individually or the entire sample. Other longitudinal studies also found no association between various measures of prenatal androgens measured in umbilical cord blood and amniotic fluid and autistic traits (Jamnadass et al., 2015; Whitehouse et al., 2012). Thus, further investigations, preferably using larger samples, are required for resolving the inconsistencies in this literature.

To conclude, we tested T's causal role in cognitive empathy across distinct administration methods using notably large samples from two distinct populations, and found no evidence of an effect of T administration on RMET in young adult neurotypical males. While our results do not exclude all possible relationships between T and interpreting others' emotions and states of mind, our large-scale study and evaluation of previous literature exhibit robust evidence of no causal relationship between activational and purported developmental T exposure and cognitive empathy.

Author contributions

A.N.: experimental design, manuscript, data analysis; G.N.: experimental design, manuscript, data analysis; C.C.: manuscript; D.Z.: manuscript, hormonal assay; T.L.O.: experimental design, data collection; hormonal assay; N.V.W.: manuscript; J.M.C.: experimental design, manuscript.

Acknowledgements. Funding for this work generously provided by Caltech, Ivey Business School, IFREE, Russell Sage Foundation, University of Southern California, INSEAD, Stockholm School of Economics, Wharton Neuroscience Initiative, the Natural Sciences and Engineering Council of Canada, and the Northern Ontario Heritage Fund Corporation. Special thank you to Jorge Barraza, Austin Henderson, Garrett Thoelen, Dylan Mandred, Kimberly Gilbert, Caelan Mathers, Emily Jeanneault, Nicole Marley, Kendra Maracle, Victoria Bass-Parcher, Nadia Desrosiers, Charlotte Miller, Brittney Robinson, Dalton Rogers, Megan Phillips, Brandon Reimer, Camille Gray, Christine Jessamine, and Brandon Reimer who assisted this study, and David Kimball for LC-MS/MS assay testing.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

Apicella, C. L., Tobolsky, V. A., Marlowe, F. W., & Miller, K. W. (2016). Hadza hunter-gatherer men do not have more masculine digit ratios (2D:4D). *American Journal of Physical Anthropology*, *159*(2), 223–232.

Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2009). Fetal testosterone and autistic traits. *British Journal of Psychology* , *100*(Pt 1), 1–22.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences*, *6*(6), 248–254.

Baron-Cohen, S., Lombardo, M. V., Auyeung, B., Ashwin, E., Chakrabarti, B., & Knickmeyer, R. (2011). Why are autism spectrum conditions more prevalent in males? *PLoS Biology*, *9*(6), e1001081.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251.

Berenbaum, S. A., Bryk, K. K., Nowak, N., Quigley, C. A., & Moffat, S. (2009). Fingers as a marker of prenatal androgen exposure. *Endocrinology*, *150*(11), 5119–5124.

Bos, P. A., Hofman, D., Hermans, E. J., Montoya, E. R., Baron-Cohen, S., & van Honk, J. (2016). Testosterone reduces functional connectivity during the "Reading the Mind in the Eyes" Test. *Psychoneuroendocrinology*, *68*, 194–201.

Brosnan, M., Lewton, M., & Ashwin, C. (2016). Reasoning on the Autism Spectrum: A Dual Process Theory Account. *Journal of Autism and Developmental Disorders*, *46*(6), 2115–2125.

Carré, J. M., Ortiz, T. L., Labine, B., Moreau, B. J. P., Viding, E., Neumann, C. S., & Goldfarb, B. (2015). Digit ratio (2D:4D) and psychopathic traits moderate the effect of exogenous testosterone on socio-cognitive processes in men. *Psychoneuroendocrinology*, *62*, 319–326.

Cherrier, M. M., Craft, S., & Matsumoto, A. H. (2003). Cognitive changes associated with supplementation of testosterone or dihydrotestosterone in mildly hypogonadal men: a preliminary report. *Journal of Andrology*, *24*(4), 568–576.

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature*, *463*(7279), 356–359.

Eisenegger, C., von Eckardstein, A., Fehr, E., & von Eckardstein, S. (2013). Pharmacokinetics of testosterone and estradiol gel preparations in healthy young men. *Psychoneuroendocrinology*, *38*(2), 171–178.

Ferguson, J. N., Young, L. J., & Insel, T. R. (2002). The neuroendocrine basis of social recognition. *Frontiers in Neuroendocrinology*, *23*(2), 200–224.

Guyatt, A. L., Heron, J., Le Cornu, K. B., Golding, J., & Rai, D. (2015). Digit ratio and autism spectrum disorders in the Avon Longitudinal Study of Parents & Children: a birth cohort study. *BMJ Open*.

Hönekopp, J. (2012). Digit ratio 2D:4D in relation to autism spectrum disorders, empathizing, and systemizing: a quantitative review. *Autism Research: Official Journal of the International Society for Autism Research*, *5*(4), 221–230.

Jamnadass, E. S. L., Keelan, J. A., Hollier, L. P., Hickey, M., Maybery, M. T., & Whitehouse, A. J. O. (2015). The perinatal androgen

to estrogen ratio and autistic-like traits in the general population: a longitudinal pregnancy cohort study. *Journal of Neurodevelopmental Disorders*, *7*(1), 17.

Knickmeyer, R., Baron-Cohen, S., Fane, B. A., Wheelwright, S., Mathews, G. A., Conway, G. S., … Hines, M. (2006). Androgens and autistic traits: A study of individuals with congenital adrenal hyperplasia. *Hormones and Behavior*, *50*(1), 148–153.

Kung, K. T. F., Spencer, D., Pasterski, V., Neufeld, S., Glover, V., O'connor, T. G., … Hines, M. (2016). No relationship between prenatal androgen exposure and autistic traits: convergent evidence from studies of children with congenital adrenal hyperplasia and of amniotic testosterone concentrations in typically developing children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(12), 1455–1462.

Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., & Manning, J. T. (2004). 2nd to 4th digit ratios, fetal testosterone and estradiol. *Early Human Development*, *77*(1-2), 23–28.

Manning, J. T. (2002). *Digit Ratio: A Pointer to Fertility, Behavior, and Health*. Rutgers University Press.

Masuya, Y., Okamoto, Y., Inohara, K., Matsumura, Y., Fujioka, T., Wada, Y., & Kosaka, H. (2015). Sex-different abnormalities in the right second to fourth digit ratio in Japanese individuals with autism spectrum disorders. *Molecular Autism*, *6*, 34.

McIntyre, M. H., Ellison, P. T., Lieberman, D. E., Demerath, E., & Towne, B. (2005). The development of sex differences in digital formula from infancy in the Fels Longitudinal Study. *Proceedings. Biological Sciences / The Royal Society*, *272*(1571), 1473–1479.

Nadler, A., Jiao, P., Alexander, V., Johnson, C., & Zak, P. (2017). The Bull of Wall Street: Experimental Analysis of Testosterone and Asset Trading. *Management Science*. https://doi.org/10.1287/mnsc.2017.2836

Nave, G., Nadler, A., Dubois, D., Zava, D., Camerer, C., & Plassmann, H. (2018). Single-dose testosterone administration increases men's preference for status goods. *Nature Communications*, *9*(1), 2433.

Nave, G., Nadler, A., Zava, D., & Camerer, C. (2017). Single-Dose Testosterone Administration Impairs Cognitive Reflection in Men. *Psychological Science*, *28*(10), 1398–1407.

Olsson, A., Kopsida, E., Sorjonen, K., & Savic, I. (2016). Testosterone and estrogen impact social evaluations and vicarious emotions: A double-blind placebo-controlled study. *Emotion* , *16*(4), 515–523.

Rogol, A. D., Tkachenko, N., & Bryson, N. (2016). Natesto™ , a novel testosterone nasal gel, normalizes androgen levels in hypogonadal men. *Andrology*, *4*(1), 46–54.

Swerdloff, R. S., Wang, C., Cunningham, G., Dobs, A., Iranmanesh, A., Matsumoto, A. M., … Berman, N. (2000). Long-term pharmacokinetics of transdermal testosterone gel in hypogonadal men. *The Journal of Clinical Endocrinology and Metabolism*, *85*(12), 4500–4510.

Teatero, M. L., & Netley, C. (2013). A Critical Review of the Research on the Extreme Male Brain Theory and Digit Ratio (2D:4D). *Journal of Autism and Developmental Disorders*, *43*(11), 2664–2676.

Tuiten, A., Van Honk, J., Koppeschaar, H., Bernaards, C., Thijssen, J., & Verbaten, R. (2000). Time course of effects of testosterone administration on sexual arousal in women. *Archives of General Psychiatry*, *57*(2), 149–153; discussion 155–156.

Turanovic, J. J., Pratt, T. C., & Piquero, A. R. (2017). Exposure to fetal testosterone, aggression, and violent behavior: A meta-analysis of the 2D:4D digit ratio. *Aggression and Violent Behavior*, *33*, 51–61.

Valla, J. M., & Ceci, S. J. (2011). Can Sex Differences in Science Be Tied to the Long Reach of Prenatal Hormones?: Brain Organization Theory, Digit Ratio (2D/4D), and Sex Differences in Preferences and Cognition. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(2), 134–146.

van Honk, J., Schutter, D. J., Bos, P. A., Kruijt, A.-W., Lentjes, E. G., & Baron-Cohen, S. (2011). Testosterone administration impairs cognitive empathy in women depending on second-to-fourth digit ratio. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(8), 3448–3452.

Voracek, M., & Dressler, S. G. (2006). Lack of correlation between digit ratio (2D: 4D) and Baron-Cohen's "Reading the Mind in the Eyes" test, empathy, systemising, and autism-spectrum quotients in a general population sample. *Personality and Individual Differences*, *41*(8), 1481–1491.

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule-expanded form. Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs

Whitehouse, A. J., Mattes, E., Maybery, M. T., Dissanayake, C., Sawyer, M., Jones, R. M., … Hickey, M. (2012). Perinatal testosterone exposure and autistic-like traits in the general population: a longitudinal pregnancy-cohort study. *Journal of Neurodevelopmental Disorders*, *4*(1), 25.

Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., von Schoultz, B., Hirschberg, A. L., & Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6535–6538.

Zheng, Z., & Cohn, M. J. (2011). Developmental basis of sexually dimorphic digit ratios. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16289–16294.

Zuloaga, D. G., Puts, D. A., Jordan, C. L., & Breedlove, S. M. (2008). The role of androgen receptors in the masculinization of brain and behavior: what we've learned from the testicular feminization mutation. *Hormones and Behavior*, *53*(5), 613–626.