

Does Testosterone Affect Cognitive Reflection? Evidence From a Double-Blind, Randomized Controlled Study of 1,000 Participants

Psychological Science
1–14

© The Author(s) 2026

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976261458325
www.psychologicalscience.org/PS

Erik L. Knight¹, Gideon Nave², Steven D. Shaw², Coren Apicella³,
Pierre L. Bonin⁴, Anna Dreber⁵, Shawn N. Geniole⁶,
Magnus Johannesson⁵, Dylan Manfredi², Pranjal Mehta⁷,
Valentina Proietti⁶, Steven J. Stanton⁸, Francesca R. Luberti⁹,
Triana Ortiz⁹, and Justin M. Carré⁹

¹Department of Psychology and Neuroscience, University of Colorado Boulder; ²Marketing Department, The Wharton School, University of Pennsylvania; ³Department of Psychology, University of Pennsylvania;

⁴Northern Ontario School of Medicine; ⁵Department of Economics, Stockholm School of Economics;

⁶Department of Psychology, University of the Fraser Valley; ⁷Department of Experimental Psychology,

University College London; ⁸Department of Management and Marketing, School of Business Administration,

Oakland University; and ⁹Department of Psychology, Nipissing University

Abstract

The cognitive reflection test (CRT) measures reliance on intuitive thinking versus deliberate reasoning and predicts important real-world outcomes. Prior research has suggested that testosterone administration impaired CRT performance, but follow-up studies produced null results. To provide a rigorous test, we conducted a large, preregistered, double-blind, placebo-controlled experiment, unprecedented in size, with 1,000 adult men, as part of an adversarial collaboration. Participants received a single dose of intranasal testosterone or placebo, completed the CRT, and rated their confidence level. We found an insignificant treatment effect on the CRT, with the point estimate in the opposite direction of the original hypothesis ($\beta_{\text{LOGIT}} = 0.118$, 95% confidence interval (CI) = $[-0.099, 0.335]$). In a second primary test, we found a significant negative treatment effect on confidence ($\beta_{\text{LOGIT}} = -0.329$, 95% CI = $[-0.558, -0.100]$), which is also the opposite of our prediction. Our findings challenge earlier claims about testosterone's cognitive effects and highlight the importance of high-powered replications. Long-term or developmental testosterone effects remain potentially important but difficult to study.

Keywords

testosterone, hormones, cognitive reflection, overconfidence, randomized trial, experiment

Received 5/1/25; revision accepted 3/30/26

Introduction

The cognitive reflection test (CRT; Frederick, 2005), is widely used throughout the social sciences to measure individuals' tendency to rely on intuitive versus deliberate judgments and decision-making. Better test performance corresponds to greater reliance on deliberate decision-making, whereas poorer performance corresponds to greater reliance on intuition. Test performance predicts numerous outcomes, from susceptibility

to misinformation to believing in and spreading misinformation (Martire et al., 2023) to behavioral biases (Hoppe & Kusterer, 2011). It also often serves as a simple proxy for cognitive performance (Otero et al., 2022).

Corresponding Author:

Gideon Nave, University of Pennsylvania, The Wharton School,
Marketing Department

Email: gnave@wharton.upenn.edu

The CRT has become a valuable tool for examining how contextual, psychological, and biological factors affect decision-making (e.g., Margittai et al., 2016). One such factor is the sex steroid testosterone. Testosterone levels in animals and humans respond dynamically to social and environmental stimuli, rising, for example, in competitive or mating contexts (Geniole et al., 2017) to facilitate instinctive, fast-acting behaviors, such as physical aggression (e.g., Geniole et al., 2019). This situational reactivity is central to theories such as the *challenge hypothesis*, which posits that testosterone modulates behavior in a context-sensitive way, especially under social threat or competition for status. Translating this idea to human cognition led to the hypothesis that transient testosterone elevations shift judgments and decisions toward intuitive responses, as measured by the CRT, by reducing the likelihood of engaging in the slow, difficult reflection needed to override initial impulses.

In a 2017 *Psychological Science* article, Nave et al. (2017) reported that a single dose of testosterone reduced CRT performance in a randomized, double-blind, placebo-controlled study. Drawing on theories and evidence linking testosterone to status-seeking behavior (Eisenegger et al., 2011) and status-seeking to overconfidence (Anderson et al., 2012), the authors speculated that testosterone-induced overconfidence may underlie the effect. This study included a notably large sample size of 243 men, five to ten times larger than that of prior published experimental studies at the time on testosterone's effects on cognition and decision-making (e.g., Dreher et al., 2016; Eisenegger et al., 2010; van Honk et al., 2011).

Reaching scientific consensus, however, requires more than a single study. A subsequent *Psychological Science* commentary reported inconsistent effects in three additional randomized, placebo-controlled studies on the impact of testosterone administration on CRT performance (Knight et al., 2020). All three studies ($N = 116$, $N = 396$, and $N = 116$) reported weak, statistically insignificant effects, calling into question the robustness of the original findings. However, there were critical differences between the additional studies and the original work. Most notably, the two smaller studies among the three included additional manipulations of acute stress and treatment blindness, and a significant fraction of participants (40%) in the third (larger study) were at floor CRT performance, making it difficult to detect a treatment effect in the negative direction (Nave et al., 2020). Finally, although finding no main effects of testosterone administration on cognitive reflection, Knight et al. (2020) reported a significant interaction between trait impulsivity and testosterone treatment: Testosterone negatively affected CRT performance at low levels of impulsivity,

but positively affected CRT performance at high levels of impulsivity.

Preregistered studies with greater statistical power are essential for generating more reliable evidence. We provide such evidence through a very large study, unprecedented in size, conducted as an adversarial collaboration between teams that had previously reported conflicting findings (Knight et al., 2020; Nave et al., 2017, 2020). In this double-blind, placebo-controlled, randomized experiment, participants ($N = 1,000$ men) received either a single dose of intranasal testosterone gel (11 mg of Natesto) or a placebo. After a 30-min soak period, participants completed several tasks, including the CRT (completed approximately 1 hour after drug administration). To avoid potential floor effects and to mitigate participants' familiarity with the test's questions, we relied on a validated 7-item modified version of the CRT (Manfredi & Nave, 2019). We incentivized participants by offering \$1 per correct answer in the CRT, as in Nave et al.'s (2017) study. Following the task, participants self-reported their confidence in their performance and their expectations for others' performance. We also collected survey data on trait impulsivity at baseline. The testosterone-administration paradigm used increases testosterone concentrations to approximately 60% to 80% above baseline levels, bringing them to the high-normal physiological range (Luberti et al., 2021). Notably, this paradigm has been used in several recent studies reporting behavioral effects of the treatment (Geniole et al., 2019; Luberti et al., 2024). The study was preregistered with a preanalysis plan (PAP) at the Open Science Framework (OSF; <https://osf.io/x6g3v>). Consistent with our preregistration, we used the conservative 0.5% threshold (i.e., $p < .005$) for statistical significance, as suggested by Benjamin et al. (2018).

Our primary hypothesis is that a single dose of testosterone administration negatively affects CRT performance (see Table 1 for a description of the primary hypotheses). We find no compelling evidence for such an effect, with a point estimate in the opposite direction of the hypothesis and the 99.5% confidence interval (CI) indicating that effects larger than $d = 0.11$ in the hypothesized direction are highly unlikely. In two additional primary tests, we investigated whether testosterone positively affects confidence relative to one's own performance or others' performance. We found no evidence supporting the hypothesized effects, but we unexpectedly observed a significant decrease in confidence ($p < .005$) relative to one's performance following testosterone administration. Despite the high-powered preregistered design and the strong statistical evidence, we interpreted this finding very cautiously, given the low prior probability of finding an effect in this direction, which contrasts current theories of testosterone's role in human cognition and behavior.

Table 1. Primary Hypotheses Tested in the Study

Hypothesis	Outcome variable	Definition of outcome variable	Hypothesized sign
Testosterone decreases cognitive reflection	CRT score	Number of correct answers in the CRT (0–7 correct responses).	Negative
Testosterone increases confidence	Confidence score	The guessed number of one's own correct responses (0–7) in the CRT, controlling for actual performance (a measure of overconfidence in relation to one's own performance).	Positive
Testosterone increases relative confidence	Relative confidence score	The guessed number of correct responses on the CRT minus the guessed average CRT score of the other participants (a measure of overconfidence in relation to the performance of others)	Positive

Note: CRT = cognitive reflection test.

Thus, further replication studies are needed to confirm the robustness and generalizability of this result. Finally, in preregistered exploratory analyses, we failed to conceptually replicate the interaction between testosterone treatment and trait impulsivity reported by Knight et al. (2020). Still, we found significant evidence of a negative association between basal testosterone and CRT score and suggestive evidence of a positive interaction between basal cortisol and testosterone treatment on CRT performance. To carry much weight, however, these exploratory findings will require confirmation in future studies.

In conclusion, our findings fail to conceptually replicate the negative effect of single-dose testosterone on cognitive reflection reported in Nave et al. (2017), suggesting instead that short-term testosterone fluctuations do not play a meaningful role in cognitive reflection. The potential of developmental or long-term testosterone effects, which are ethically challenging to test experimentally in humans, remains an open question for future research.

Research Transparency Statement

General disclosures

Conflicts of interest: The authors declare no conflicts of interest. **Funding:** This research was supported by the Jan Wallander and Tom Hedelius Foundation (Grant Nos. P21-0091 and P23-0098 to A.D.), Knut and Alice Wallenberg Foundation and Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to A. Dreber), and Riksbankens Jubileumsfond (Grants NHS14-1719:1 and P21-0168 to M. Johannesson). This work was also supported by a National Science Foundation Faculty Early Career Development Program grant (Grant No. 1942917) to G. Nave and a Discovery Grant from the Natural Sciences and Engineering Research Council of

Canada to J. M. Carré. (RGPIN-2020-05120). **Artificial intelligence:** No artificial-intelligence-assisted technologies were used in this research or the creation of this article. **Ethics:** The study was approved by the Nipissing University Research Ethics Board (Protocol No. 101652).

Study 1 disclosures

Preregistration: The research aims and hypotheses, along with the materials, methods, and analysis plan, were preregistered (<https://osf.io/x6g3v>) and posted on the Open Science Framework (OSF) in March 2018, shortly after data collection had started in February 2018 and before data were accessed or analyzed (data collection was completed in December 2023). There were minor deviations from the preregistration in the analyses and in the collection and coding of the independent variables (for details, see Table S1 in the Supplemental Material available online). **Materials:** All study materials are publicly available (<https://osf.io/37ycj>). **Data:** All primary data are publicly available (<https://osf.io/37ycj>). **Analysis scripts:** All analysis scripts are publicly available (<https://osf.io/37ycj>). All materials, data, and analysis scripts are under a Creative Commons CC0 license. **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team. **Open Science Framework (OSF):** To ensure long-term preservation, we registered all OSF files at <https://osf.io/37ycj>.

Method

Study location and sample

Location. We collected data in three locations: (a) Nipissing University's Social Neuroendocrinology Laboratory in

North Bay, Ontario; (b) Georgian College in Barrie, Ontario; and (c) medical offices in Sudbury, Ontario. Participants completed the study in sessions of up to seven people in isolated computer stations.

Sample. Participants were men ranging in age from 18 to 45 years old. They included university students, community college students, and community members. Participants were recruited through Facebook, Kijiji, campus posters and booths, classroom announcements, a research participant database, community posters, local mall booths, newspaper ads, radio ads, and word of mouth. Data collection began on February 13, 2018, and concluded on December 13, 2023, when we reached our target sample size of 1,000. The data collection took much longer than initially planned because of the COVID-19 pandemic.

Participant payment. Participants received \$40 for their participation, which required 2 hours of their time, plus a variable bonus based on their decisions and chance.

Experimental procedure

Below, we describe the experimental procedure in chronological order.

Informed consent and rules. Participants read and signed an informed-consent form that outlined the study rules; they then changed into a new T-shirt to wear during the study (as part of another study, not reported here, that examined the potential effects of testosterone on olfactory cues).

Trait measures. Participants filled out the Ten-Item Personality Inventory (TIPI), the Self-Report Dominance and Prestige Scale, the Self-Control Scale (Tangney et al., 2004), the Barratt Impulsivity Scale (BIS; Patton et al., 1995), and the Thirty-Item Self-Construct Scale (Singelis, 1994).

First saliva sample. Participants provided their first saliva sample by chewing on a synthetic swab for 30s. Saliva samples were stored at -20°C until the time of hormone determination, when they were thawed and centrifuged. The supernatant was then extracted and analyzed in duplicate. Salivary testosterone and cortisol were assayed using commercial enzyme immunoassay kits from DRG International, Inc. (Springfield Township, New Jersey).

Self-administration of testosterone or placebo. Under the supervision of a research assistant, participants

self-administered 11 mg of testosterone or placebo (5.5 mg per nostril) using 1-mm amber syringes. Both the testosterone gel and the placebo gel were prepared by Pharmasave; a third-party pharmacy technician filled the syringes before the study. The syringes were filled with either 0.12 mL of Natesto or Secaris. The volume of 0.12 mL was determined to deliver 5.5 mg of testosterone and account for the void space in each syringe. This volume was determined by using the density (weight per volume) of testosterone and then weighing the delivered amounts until a consistent volume was released.

Soak period (30 min). During the soak period, participants answered a series of demographic questions, completed the MacArthur Scale of Subjective Social Status, reported their income, answered two questions regarding sleep, had a picture of their faces taken, and had their hands scanned (for studies not reported here).

Second saliva sample. Participants provided their second saliva sample using the same procedure as before.

Mouthwash sample. Participants provided a mouthwash sample used for genetic analyses (not reported here).

Primary tasks. The experimental battery included the following tasks: Brief Mood Introspection Scale (Mayer & Gaschke, 1988), risky decision-making task, ultimatum game, trust game, dictator game, charity game, CRT/confidence task, competition task, sexual motivation task, and grip-strength task. The order of these tasks was the same for all participants.

Cognitive reflection test (CRT). The CRT task included seven math problems designed to elicit compelling but incorrect intuitive responses. We used the CRT version developed by Manfredi and Nave (2019), a reworded adaptation of the extended seven-item test by Toplak et al. (2014), to reduce the risk of floor and ceiling effects and prior familiarity with the items. Each question was presented with a prompt and an open-response text box. Participants' answers were coded as either correct or incorrect, resulting in a total score ranging from 0 to 7. Participants received \$1 for each correct answer, similar to the approach used in Nave et al.'s (2017) study.

Confidence. Participants responded to two questions in which they guessed their own and others' performance on the CRT using open-response text boxes: "How many of the previous seven questions do you think you answered correctly?" and "How many of the previous seven questions do you think the average participant answered correctly?" Participants were asked to round

their responses to the nearest integer. Any text response that could not be easily interpreted as an integer number was excluded from analysis, as were numbers larger than 7. Nonnumeric responses that could be readily interpreted as integer numbers were included. For example, “most,” “half,” and “90%” would be excluded because they could not be coded as an integer, whereas “100%” and “I think I missed one” were recoded as 7 and 6, respectively. This interpretation process was not preregistered. We defined relative confidence as the guessed number of correct responses for oneself minus the (guessed) number of correct responses attributed to the average participant. We did not incentivize the two confidence questions, and we cannot rule out that this could have affected the responses.

Closing tasks. Participants completed two study-satisfaction questions, answered a few questions about the study’s purpose and whether they had participated in a similar study before, and answered two questions about treatment expectancy.

Preregistration and statistical testing

We posted a preanalysis plan (PAP) at OSF in March 2018 shortly after the data collection had started and before data were accessed or analyzed (<https://osf.io/x6g3v>). The preregistered analyses and tests were divided into hypotheses (referred to here as the three primary hypotheses), robustness tests, and exploratory hypotheses. We followed the PAP precisely, except for the deviations listed in the Supplemental Material available online (see Table S1 for details). These deviations and clarifications are due to differences and ambiguities in how some variables were collected and coded. Consistent with our preregistration, we based all tests on two-sided p values and interpreted results with $p < .005$ as “statistically significant evidence” and results with $.005 \leq p < .05$ as “suggestive evidence,” as recommended by Benjamin et al. (2018). One motivation for using the more conservative $p < .005$ threshold for statistical significance is that multiple comparisons increase the risk of false positives. Our study has 80% statistical power to detect a Cohen’s d effect size as small as 0.23 at the 0.5% threshold for statistical significance used in the study, and 80% power to detect a Cohen’s d effect size as small as 0.18 at the suggestive 5% level. (These power estimates are based on using an independent-samples t test or ordinary least squares [OLS] regression to interpret power in Cohen’s d units, rather than the ordered logistic regressions used in the primary hypotheses—but as seen in Table S4 in the Supplemental Material, the test statistics are very similar for an OLS model and an ordered logistic regression.)

Results

Manipulation check. We recruited 1,000 participants who completed the study, with 500 in the placebo group and 500 in the testosterone group (i.e., a list of 1,000 was randomized so that half were placebo and half were testosterone). Of these, 999 participants (500 in the testosterone group) completed the CRT task. Tables S2 and S3 in the Supplemental Material provide descriptive statistics of the testosterone and placebo groups. In addition to the baseline measurement, testosterone levels were measured in saliva samples from both groups 30 min after drug administration. As a manipulation check, we confirmed that the treatment increased testosterone levels compared with the placebo group (independent-samples t test on the log of testosterone: $t(993) = 37.07$, $p < .0001$; $N = 995$). This manipulation check was not preregistered, and we note that it is difficult to assess the actual magnitude of the testosterone increase in saliva samples after administration. Thus, the test only confirms that the treatment increased testosterone levels, but not by how much. Based on previous work measuring serum testosterone, the testosterone-administration paradigm is expected to increase serum levels by 60% to 80% (Luberti et al., 2021).

Confidence responses. Of the 999 self-confidence responses, 957 contained appropriate numeric responses, and 20 text responses were converted to numerical values; 7 contained text that could not be converted, and 15 contained numbers that were too large. Of the responses related to the average participant’s performance, 985 were appropriate, 13 contained numbers that were too large, and one was left blank.

Testosterone’s effects on cognitive reflection and confidence. We conducted three primary hypothesis tests, as described in Table 1, with the hypothesized directions indicated. These tests’ three primary outcome measures are CRT performance, confidence, and relative confidence scores. The CRT score is the number of correct answers on the seven-item cognitive reflection test used in the study (Manfredi & Nave, 2019); the Confidence Score is the participant’s belief about the number of correct answers on the cognitive reflection test (between 0 and 7), and the relative confidence score is the confidence score minus the belief about the average number of correct responses of other participants (between -7 and 7).

Consistent with our preregistration, we used ordered logistic regression [OLR] to test these three hypotheses. All models were run using the *polr* function from the MASS package (Venables & Ripley, 2002) in R (Version 4.4.1). We calculated p values for OLRs by comparing the

t value to the standard normal distribution. The t values and normal distributions are equivalent in the limit of infinite degrees of freedom and can be used as reasonable approximations for sufficiently large sample sizes. To ease the interpretation of the estimated effects' magnitudes, we have also included the coefficient and confidence intervals from an ordinary least squares regression with robust standard errors in Table S4 in the Supplemental Material. (These ordinary least squares regressions were not preregistered and were not used for hypothesis testing, but were reported only for descriptive purposes.) The regression models estimated for the three primary tests include a binary treatment indicator (1 = *testosterone*, 0 = *placebo*) as the independent variable. In the tests of the confidence score and the relative confidence score, we also, as preregistered, controlled for each participant's CRT score (i.e., their actual performance).

Figure 1a plots the CRT score in the testosterone and placebo groups, and Table 2 reports the results of testing the three primary hypotheses. The average number of correct answers on the CRT is 3.01 ($SD = 2.18$) in the testosterone group and 2.86 ($SD = 2.12$) in the placebo group. We find no statistical evidence of an effect of testosterone on the CRT score, with the point estimate in the opposite direction of the hypothesis ($\beta_{\text{LOGIT}} = 0.118$, 95% CI = $[-0.099, 0.335]$, $z = 1.065$, $p = .287$, $N = 999$). Figure 2a illustrates the predicted effect of testosterone administration on the CRT score.

Figure 1b displays participants' confidence (i.e., their estimates of how many CRT questions they answered correctly) in the testosterone and placebo groups. The mean guess is 5.37 ($SD = 1.42$) in the testosterone group and 5.55 ($SD = 1.35$) in the placebo group, indicating that participants are overconfident. Figure 2b illustrates the predicted effect of testosterone treatment on the confidence score. Our regression models testing our second primary hypothesis failed to find evidence supporting the hypothesis that testosterone has a positive effect on the confidence score; however, we found a statistically significant effect of testosterone on the confidence score in the opposite direction of the hypothesis ($\beta_{\text{LOGIT}} = -0.329$, 95% CI = $[-0.558, -0.100]$, $z = -2.821$, $p = .0048$; $N = 977$). As this is an unexpected finding with a low prior probability, we interpret it cautiously until it is confirmed by additional studies. We did not adjust our hypothesis tests for multiple comparisons, but instead used the conservative $p < .005$ threshold for statistical significance. We tested three primary hypotheses in this study and nine primary hypotheses in a previously published study on economic preferences (Dreber et al., 2025). A Bonferroni correction based on these 12 tests implies that the confidence-score result would have an adjusted p value greater than .05, which is another reason to interpret this result cautiously.

The mean guess about other participants' CRT scores is 4.65 ($SD = 1.14$) in the testosterone group and 4.75 ($SD = 1.07$) in the placebo group, indicating that participants also overestimate the performance of others, though to a lesser degree than their own performance. This pattern is consistent with previous results on overconfidence in the CRT (Ring et al., 2016). The third primary outcome variable, relative confidence score, is the difference between participants' estimate of their own performance and their estimate of others' performance. Figure 1c plots the distribution of the relative confidence score in both treatment groups. The hypothesis test on the relative confidence score yields an insignificant effect in the opposite direction of our hypothesis ($\beta_{\text{LOGIT}} = -0.205$, 95% CI = $[-0.432, 0.021]$, $z = -1.776$, $p = .076$, $N = 970$). Figure 2c illustrates the predicted treatment effect on the relative confidence score.

In a preregistered robustness test of the three primary hypotheses, we add the following control variables to the ordered logistic regressions: math ability, right-hand 2D:4D, pleasant-unpleasant mood, arousal-calm mood, time of day, and a categorical variable for the three sites (see the Supplemental Material for the exact variable definitions). Table 2 reports these results, showing that the statistical conclusion is the same for the CRT score. In contrast, the unexpected finding in the opposite direction of the hypothesis for the confidence score changes from statistically significant to suggestive evidence. For the relative confidence score, the effect of testosterone now reaches suggestive evidence of testosterone reducing relative confidence after we controlled for covariates ($\beta_{\text{LOGIT}} = -0.244$, 95% CI = $[-0.484, -0.004]$, $z = -1.989$, $p = .047$, $N = 875$). This effect runs counter to the hypothesis's prediction. Note that although the statistical conclusions differ somewhat in the robustness tests, the z values in the robustness tests are close to the z values in the main tests. The sample size is slightly smaller in the robustness tests because of missing values for some added control variables ($N = 875$ – 897 , depending on the model).

Preregistered exploratory hypotheses. We also conducted several preregistered exploratory analyses, and we report these tests in Table 3, which includes information about the hypothesized directions of these tests. Some of these tests involve interactions, and for those tests, we used an OLS regression model rather than an ordered logistic regression, because interpreting interactions in logistic regression is difficult (Ai & Norton, 2003).

First, we test whether participants' basal testosterone and cortisol levels (log-transformed and residualized for time of day and waking time) are associated with the CRT score. In line with the preregistered exploratory hypothesis, we found a significant negative association between basal testosterone and CRT score ($\beta = -0.158$,

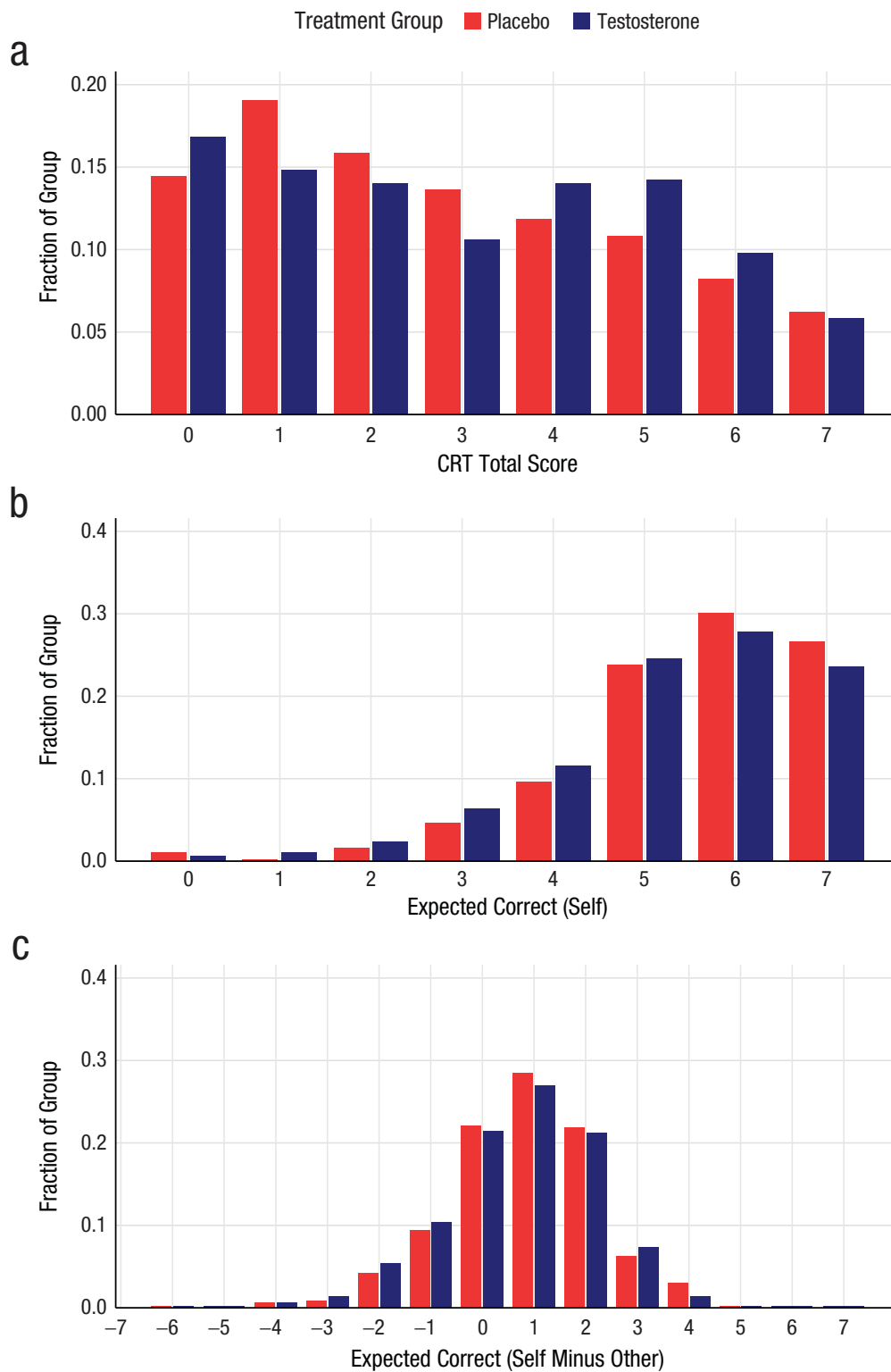


Fig. 1. Distribution of responses as the fraction of the treatment group for (a) CRT performance, (b) confidence (guessed number correct), and (c) relative confidence (guessed number correct minus guessed number correct for average participant). CRT = cognitive reflection test.

Table 2. Test Results and Robustness Tests of the Three Primary Hypotheses

Outcome variable	Hypothesized sign	N	Primary hypothesis tests (ordered logistic regression)			
			Treatment effect (SE)	95% CI (99.5% CI)	z value	p value
CRT	Negative	999	0.118 (0.111)	−0.099, 0.335 (−0.193, 0.429)	1.07	0.297
Confidence score	Positive	977	−0.329 (0.117)	−0.558, −0.100 (−0.657, −0.002)	−2.82	0.005
Relative confidence score	Positive	970	−0.205 (0.116)	−0.432, 0.021 (−0.530, 0.119)	−1.78	0.076
Outcome variable	Hypothesized sign	N	Robustness tests of primary hypothesis tests; ordered logistic regression			
			Treatment effect (SE)	95% CI (99.5% CI)	z value	p value
CRT	Negative	897	0.148 (0.118)	−0.084, 0.380 (−0.184, 0.480)	1.253	0.210
Confidence score	Positive	880	−0.330 (0.124)	−0.573, −0.087 (−0.678, 0.017)	−2.666	0.008
Relative confidence score	Positive	875	−0.244 (0.122)	−0.484, −0.004 (−0.587, 0.100)	−1.989	0.047

Note: CRT = cognitive reflection test.

95% CI = [−0.266, −0.049], 99.5% CI = [−0.314, −0.002], $z = -2.837$, $p = .0046$, $N = 988$). Cortisol was not associated with CRT performance. We also tested for an interaction between basal testosterone and basal cortisol, because some have argued that such an interaction affects behavior related to social status (Dekkers et al., 2019). We hypothesized a positive interaction effect, but we failed to confirm it. We further tested for an interaction between treatment and basal cortisol, finding suggestive evidence of a positive interaction in line with our hypothesis ($\beta = 0.391$, 95% CI = [0.109, 0.673], 99.5% CI = [−0.013, 0.796], $t(962) = 2.721$, $p = .007$, $N = 966$).

In two additional exploratory analyses, we tested whether testosterone administration affects the response time for the seven cognitive reflection test questions. Specifically, we tested whether participants who received testosterone provided correct answers more slowly, and incorrect answers more quickly, than participants who received a placebo. We tested these hypotheses by estimating a linear mixed-model regression with response time as the dependent variable (logged for normalization), fixed effects for treatment, and an indicator for an error on the CRT question, and the interaction between treatment and error. The model also included random effects for participant and question number. The model included seven observations per participant, with one observation per CRT question. We found no evidence of a testosterone effect on response time in these two

exploratory analyses. Finally, we tested for an interaction effect between treatment and trait impulsivity on the CRT, as reported in the exploratory analysis by Knight et al. (2020). We found no evidence of such an interaction.

Nonpreregistered robustness tests

Based on reviewers' suggestions, we tested the robustness of our three primary hypotheses after controlling for behavior in the tasks that preceded the CRT and the confidence tasks; see the Supplemental Material for details. When we included 10 variables from the preceding tasks in our ordered logistic regressions, we found statistically significant (CRT score and relative confidence score) or suggestive (confidence score) evidence of associations between these 10 variables and the three outcome measures, on the basis of a joint Wald test (see Table S5 in the Supplemental Material). However, adding these controls did not affect the conclusions about the treatment effects on the three outcome measures. We also tested for interactions between these 10 variables and the testosterone treatment, but found no evidence of such interactions for any of the three outcome measures, after joint Wald tests (see Table S5 in the Supplemental Material). Finally, using nonparametric Mann-Whitney U tests, we tested whether CRT score, the confidence score, the confidence score minus the CRT Score, and the relative confidence score

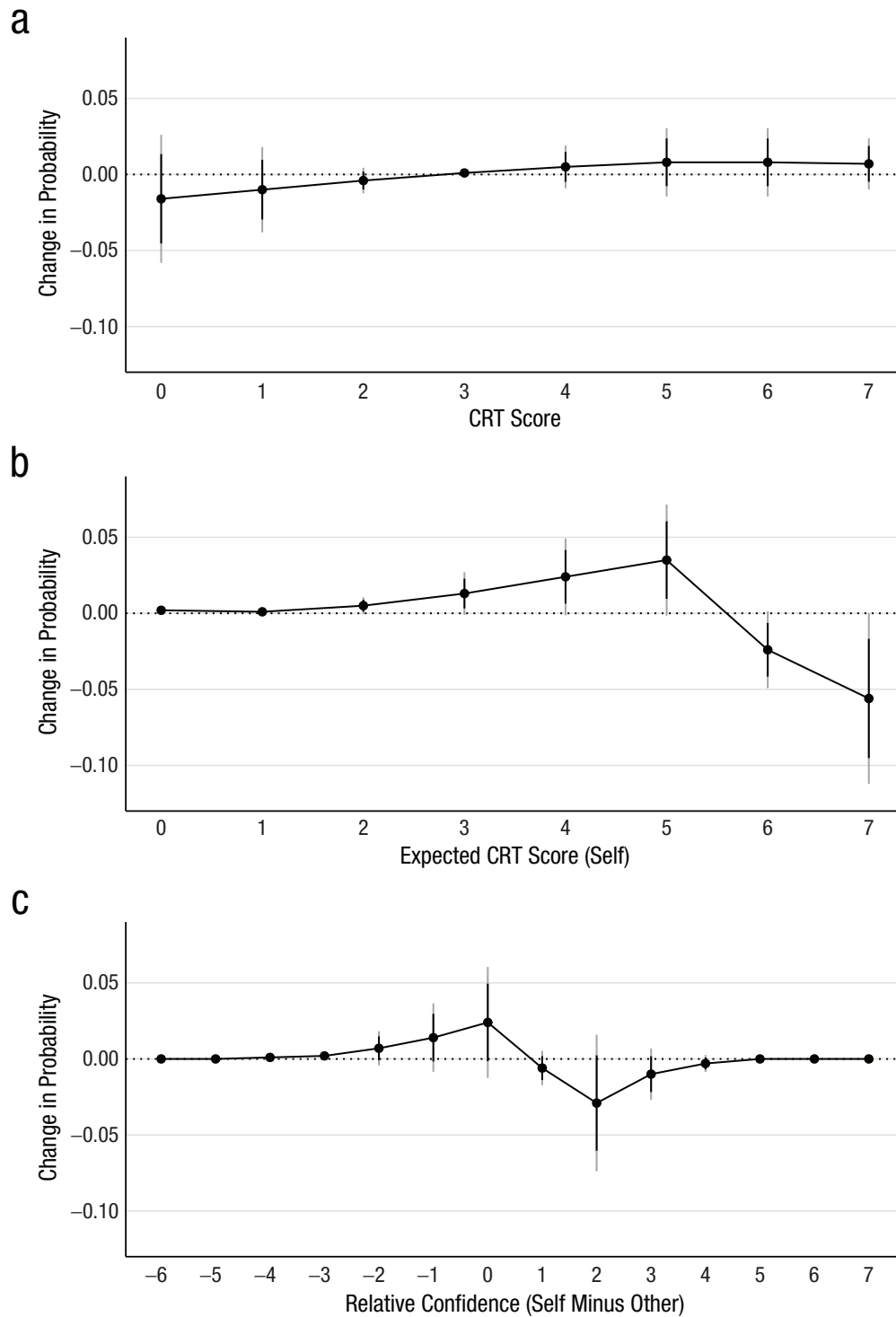


Fig. 2. Marginal effects of testosterone treatment compared with placebo from ordinal logistic regression models for (a) CRT performance, (b) confidence (guessed number correct), and (c) relative confidence (guessed number correct minus guessed number correct for average participant). Black error bars denote 95% confidence intervals; gray error bars denote 99.5% confidence intervals. CRT = cognitive reflection test.

differed between the testosterone and placebo groups. These analyses did not affect our conclusions, except that the evidence for a treatment effect on the confidence

score (in the opposite direction of our hypothesis) is now suggestive rather than statistically significant (see Table S6 in the Supplemental Material).

Table 3. Preregistered Exploratory Analyses

Dependent variable: CRT score (ordered logistic regression)					
Independent variable	Hypothesized sign	N	Regression coefficient (SE)	95% CI (99.5% CI)	p value
Basal testosterone	Negative	988	-0.158 (0.056)	-0.266, -0.049 (-0.314, -0.002)	0.005
Basal cortisol	Negative	966	0.039 (0.060)	-0.078, 0.156 (-0.128, 0.207)	0.511
Dependent variable: CRT score (ordinary least squares regression with robust standard errors)					
Independent variable	Hypothesized sign	N	Regression coefficient (SE)	95% CI (99.5% CI)	t value (df) p value
Basal Testosterone × Basal Cortisol	Positive	965	-0.108 (0.070)	-0.246, 0.030 (-0.306, 0.090)	-1.531 0.126
Basal Cortisol × Treatment	Positive	966	0.391 (0.144)	0.109, 0.673 (-0.013, 0.796)	2.721 0.007
Trait impulsivity × Treatment	Positive	996	0.106 (0.075)	-0.041, 0.253 (-0.104, 0.316)	1.419 (992) 0.156
Dependent variable: Log of response time (linear mixed-model regression; seven observations per participant)					
Independent variable	Hypothesized sign	N (obs)	Regression coefficient (SE)	95% CI (99.5% CI)	t value (df) p value
Testosterone treatment	Positive (on CRT questions answered correctly)	999 (6,993)	-0.023 (0.034)	-0.089, 0.042 (-0.118, 0.071)	-0.695 (1909.5) 0.487
Testosterone treatment ¹	Negative (on CRT questions answered incorrectly)		-0.010 (0.031)	-0.071, 0.050 (-0.097, 0.076)	-0.338 (1461.2) 0.736

Note: The first five exploratory analyses are carried out in separate regressions, and the last two exploratory analyses are carried out in the same regression. Degrees of freedom were calculated using the Satterthwaite method. CRT = cognitive reflection test; obs = observations; CI = confidence interval.

¹This is the sum of the testosterone treatment coefficient and the coefficient of the interaction between the treatment variable and the error variable (the sum of these coefficients measure the effect of testosterone treatment on the response time on CRT questions with incorrect answers).

Discussion

In a randomized, preregistered, placebo-controlled study with an unprecedented sample size of 1,000 men, conducted as an adversarial collaboration between the original research teams, we found a weak, insignificant effect of a single dose of testosterone on cognitive reflection, in the opposite direction of our hypothesis. This finding suggests that momentary variations in testosterone do not affect this behavioral outcome. Our study can be considered a highly powered conceptual replication of Nave et al. (2017). Still, it should not be interpreted as a direct replication, because the designs are not identical (Dreber & Johannesson, 2025).

Using the 99.5% confidence interval, we found strong evidence against effects larger than about $d = 0.11$ in the hypothesized direction. We also failed to find evidence for our two additional primary hypotheses on the confidence score and the relative confidence score. However, we surprisingly found a statistically significant effect ($p < .005$) for the confidence score in the opposite direction of our hypothesis, indicating that testosterone decreased confidence relative to placebo. This result should be interpreted cautiously unless confirmed in additional studies, but it provides strong evidence against the hypothesized effect. Relatedly, two randomized, double-blind studies with 91 and 242 participants reported weak, statistically insignificant effects of testosterone on confidence (Nadler et al., 2024). However, these studies were underpowered to detect the effect size observed for the confidence score in this study (even when we pooled the studies, there was less than 50% power to detect such an effect size at the 5% level). We also could not conceptually replicate the result reported by Knight et al. (2020), which found an interaction between testosterone treatment and trait impulsivity on cognitive reflection.

In an exploratory analysis, we found evidence for a significant negative correlation between basal testosterone and CRT performance. However, as this was only an exploratory analysis, this correlational evidence requires confirmation from new studies before we draw firm conclusions. Finally, another exploratory analysis found suggestive evidence of a positive interaction between basal cortisol and testosterone treatment on CRT performance. This interaction also requires confirmation in future studies.

Our findings, along with the current literature, have limitations that constrain their generalizability and leave important questions unanswered for future research. First, similar to previous research on this question, we included only men (Knight et al., 2020; Nave et al., 2017). We know from animal research and behavioral theories, such as the challenge hypothesis, that more potent effects of testosterone on behavior are expected in men. Nevertheless,

other work has suggested that there are also testosterone-dependent effects on the brain and behavior in women (Carré et al., 2023). Therefore, it will be valuable to evaluate the generalizability of our findings to women in future research.

Second, the dosage, the administration method, and the timing of hormone administration may be important for observing a behavioral effect. Previous work has shown that the intranasal administration paradigm used here yields an increase in serum testosterone concentrations of 60% to 80% within 30 min (Luberti et al., 2021) and that this affects outcome measures in men, such as aggression and dehumanization (Carré et al., 2017; Geniole et al., 2019; Luberti et al., 2024). Furthermore, previous work indicates that a single dose of testosterone affects threat-related amygdala function in both men (Goetz et al., 2014) and women (van Wingen et al., 2009) within 45 to 60 min of administration, suggesting that testosterone may exert relatively rapid, possibly nongenomic, effects on human brain function. These findings are also consistent with a growing body of work in animal models demonstrating rapid effects of testosterone on several behavioral outcomes (see Nyby, 2008, for a review). Nonetheless, it is still possible that a longer time lag is needed to observe testosterone's behavioral effects. Some prior work, such as Tuiten et al.'s (2002) study, suggested that behavioral changes may occur 3.5 to 4 hours after testosterone administration. However, that study had substantial limitations: It included only 8 participants, reported a weak treatment effect ($p = .04$), and measured outcomes after testosterone levels had already returned to baseline. Thus, although we cannot entirely rule out delayed effects, the weight of existing evidence favors a relatively rapid timescale for testosterone's influence on behavior.

We also note that, with our single-dose paradigm, we cannot rule out the possibility that administering testosterone over a longer period—such as in clinical contexts, like gender transitions or hormone replacement therapy—might influence cognitive reflection. Chronic exposure may lead to neural or behavioral changes that differ from those observed here. However, experimentally investigating such long-term effects in healthy individuals poses significant ethical and practical challenges, particularly given the difficulty of maintaining placebo control, isolating causal mechanisms, and managing potential health risks associated with sustained hormonal manipulation.

Our findings also do not speak to the potential impact of testosterone exposure during critical developmental periods, such as in utero, postnatally, or during puberty, when hormonal influences may shape brain organization and long-term cognitive tendencies. However, experimentally studying these developmental effects in humans is not feasible because of ethical constraints, and existing

evidence relies heavily on indirect proxies, such as clinical populations or the 2D:4D proxy, the validity of which is questionable (Nadler et al. 2019; Nave et al., 2021; van Leeuwen et al., 2020). Thus, although our results challenge claims about short-term hormonal effects, the role of testosterone in shaping cognitive styles throughout development remains an open and important question for future research.

Transparency

Action Editor: Russell Poldrack

Editor: Simine Vazire

Author Contributions

Erik L. Knight: Conceptualization; Formal analysis; Visualization; Writing – original draft; Writing – review & editing.

Gideon Nave: Conceptualization; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Steven D. Shaw: Formal analysis; Validation.

Coren Apicella: Conceptualization; Project administration; Supervision; Writing – review & editing.

Pierre L. Bonin: Resources.

Anna Dreber: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing – review & editing.

Shawn N. Geniole: Conceptualization; Project administration; Writing – review & editing.

Magnus Johannesson: Conceptualization; Methodology; Funding acquisition; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Dylan Manfredi: Conceptualization; Methodology; Software.

Pranjal Mehta: Conceptualization; Methodology; Project administration; Supervision; Writing – review & editing.

Valentina Proietti: Investigation; Project administration; Writing – review & editing.

Steven J. Stanton: Conceptualization; Project administration; Writing – review & editing.

Francesca R. Luberti: Investigation; Project administration; Writing – review & editing.

Triana Ortiz: Investigation; Project administration; Supervision.

Justin M. Carré: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Writing – review & editing.

Erik L. Knight and Gideon Nave contributed equally to this article.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the Jan Wallander and Tom Hedelius Foundation (Grant Nos. P21-0091 and P23-0098 to A. Dreber), Knut and Alice Wallenberg Foundation and Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to A. Dreber), and Riksbankens Jubileumsfond (Grants NHS14-1719:1 and P21-0168 to M. Johannesson). This work was also supported by a National Science Foundation Faculty Early Career Development Program grant (Grant No.

1942917) to G. Nave and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to J. M. Carré. (RGPIN-2020-05120).

Artificial Intelligence

No artificial-intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

The study was approved by the Nipissing University Research Ethics Board (Protocol No. 101652).

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

ORCID iDs

Erik L. Knight  <https://orcid.org/0000-0002-0349-542X>
 Gideon Nave  <https://orcid.org/0000-0001-6251-5630>
 Steven D. Shaw  <https://orcid.org/0000-0002-5556-5629>
 Anna Dreber  <https://orcid.org/0000-0003-3989-9941>
 Shawn N. Geniole  <https://orcid.org/0000-0001-9866-1920>
 Magnus Johannesson  <https://orcid.org/0000-0001-8759-6393>
 Dylan Manfredi  <https://orcid.org/0000-0003-3621-9402>
 Pranjal Mehta  <https://orcid.org/0000-0001-7570-7404>
 Steven J. Stanton  <https://orcid.org/0009-0003-5514-0844>
 Francesca R. Luberti  <https://orcid.org/0000-0002-4673-8508>

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976261458325>

References

- Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80, 123–129.
- Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, 103, 718–735. <https://doi.org/10.1037/a0029395>
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.
- Carré, J. M., Geniole, S. N., Ortiz, T. L., Bird, B. M., Videto, A., & Bonin, P. L. (2017). Exogenous testosterone rapidly increases aggressive behavior in dominant and impulsive men. *Biological Psychiatry*, 82, 249–256.
- Carré, J. M., Hemsworth, C. T., & Alabi, I. A. (2023). Does testosterone modulate aggression and mating behavior in humans? A narrative review of two decades of single-dose testosterone administration research. *Adaptive Human Behavior and Physiology*, 9, 400–425.
- Dekkers, T. J., van Rentergem, J. A. A., Meijer, B., Popma, A., Wagemaker, E., & Huizenga, H. M. (2019). A meta-analytical evaluation of the dual-hormone hypothesis: Does cortisol moderate the relationship between testosterone and

- status, dominance, risk taking, aggression, and psychopathy? *Neuroscience & Biobehavioral Reviews*, 96, 250–271.
- Dreber, A., & Johannesson, M. (2025). A framework for evaluating reproducibility and replicability in economics. *Economic Inquiry*, 63, 338–356.
- Dreber, A., Johannesson, M., Nave, G., Apicella, C. L., Geniole, S. N., Imai, T., Knight, E. L., Manfredi, D., Mehta, P. H., Proietti, V., Stanton, S. J., Zeltikova, A., Luberti, F. R., Ortiz, T., & Carré, J. M. (2025). Investigating the effects of single-dose intranasal testosterone on economic preferences in a large randomized trial of men. *Proceedings of the National Academy of Sciences*, 122, Article e2508519122. <https://doi.org/10.1073/pnas.2508519122>
- Dreher, J.-C., Dunne, S., Pazderska, A., Frodl, T., Nolan, J. J., & O'Doherty, J. P. (2016). Testosterone causes both prosocial and antisocial status-enhancing behaviors in human males. *Proceedings of the National Academy of Sciences*, 113, 11633–11638.
- Eisenegger, C., Haushofer, J., & Fehr, E. (2011). The role of testosterone in social interaction. *Trends in Cognitive Sciences*, 15, 263–271.
- 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.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.
- Geniole, S. N., Bird, B. M., Ruddick, E. L., & Carré, J. M. (2017). Effects of competition outcome on testosterone concentrations in humans: An updated meta-analysis. *Hormones and Behavior*, 92, 37–50.
- Geniole, S. N., Procyshyn, T. L., Marley, N., Ortiz, T. L., Bird, B. M., Marcellus, A. L., Welker, K. M., Bonin, P. L., Goldfarb, B., Watson, N. V., & Carré, J. M. (2019). Using a psychopharmacogenetic approach to identify the pathways through which—and the people for whom—testosterone promotes aggression. *Psychological Science*, 30, 481–494.
- Goetz, S. M. M., Tang, L., Thomasson, M. E., Diamond, M. P., Hariri, A. R., & Carré, J. M. (2014). Testosterone rapidly increases neural reactivity to threat in healthy men: A novel two-step pharmacological challenge paradigm. *Biological Psychiatry*, 76, 324–331.
- Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters*, 110, 97–100.
- Knight, E. L., McShane, B. B., Kutlikova, H. H., Morales, P. J., Christian, C. B., Harbaugh, W. T., Mayr, U., Ortiz, T. L., Gilbert, K., Ma-Kellams, C., Riečanský, I., Watson, N. V., Eisenegger, C., Lamm, C., Mehta, P. H., & Carré, J. M. (2020). Weak and variable effects of exogenous testosterone on cognitive reflection test performance in three experiments: Commentary on Nave, Nadler, Zava, and Camerer (2017). *Psychological Science*, 31, 890–897.
- Luberti, F. R., Proietti, V., Geniole, S. N., Bird, B. M., Ortiz, T. L., Watson, N. V., Cearns, J., Goldfarb, B., & Carré, J. M. (2024). Testosterone rapidly increases men's emotion-based dehumanization of a conservatively dressed woman. *Psychoneuroendocrinology*, 170, Article 107173. <https://doi.org/10.1016/j.psyneuen.2024.107173>
- Luberti, F. R., Reside, T.-L., Bonin, P. L., & Carré, J. M. (2021). Development of a single-dose intranasal testosterone administration paradigm for use in men and women. *Hormones and Behavior*, 136, Article 105046. <https://doi.org/10.1016/j.yhbeh.2021.105046>
- Manfredi, D., & Nave, G. (2019). *Beyond the bat and the ball: Overcoming familiarity effects in the cognitive reflection test by rewording its questions*. SSRN Working Paper. <https://dx.doi.org/10.2139/ssrn.3445138>
- Margittai, Z., Nave, G., Strombach, T., van Wingerden, M., Schwabe, L., & Kalenscher, T. (2016). Exogenous cortisol causes a shift from deliberative to intuitive thinking. *Psychoneuroendocrinology*, 64, 131–135.
- Martire, K. A., Robson, S. G., Drew, M., Nicholls, K., & Faasse, K. (2023). Thinking false and slow: Implausible beliefs and the Cognitive Reflection Test. *Psychonomic Bulletin & Review*, 30, 2387–2396. <https://doi.org/10.3758/s13423-023-02321-2>
- Mayer, J. D., & Gaschke, Y. N. (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, 55, 102–111.
- Nadler, A., Camerer, C. F., Zava, D. T., Ortiz, T. L., Watson, N. W., Carré, J. M., & Nave, G. (2019). Does testosterone impair men's cognitive empathy? Evidence from two large-scale randomized controlled trials. *Proceedings of the Royal Society B: Biological Sciences*, 286(1910), Article 20191062. <https://doi.org/10.1016/j.yhbeh.2024.105659>
- Nadler, A., Wibrall, M., Dohmen, T., Falk, A., Previtro, A., Weber, B., Camerer, C., Dreber, A., & Nave, G. (2024). Does a single dose of testosterone increase willingness to compete, confidence, and risk taking in men? Evidence from two randomised placebo-controlled experiments. *Hormones and Behavior*, 166, Article 105659. <https://doi.org/10.1016/j.yhbeh.2024.105659>
- Nave, G., Daviet, R., Nadler, A., Zava, D., & Camerer, C. (2020). Reflecting on the evidence: A reply to Knight, McShane, et al. (2020). *Psychological Science*, 31, 898–900.
- Nave, G., Koppin, C. M., Manfredi, D., Richards, G., Watson, S. J., Geffner, M. E., Yong, J. E., Kim, R., Ross, H. M., Serrano-Gonzalez, M., & Kim, M. S. (2021). No evidence for a difference in 2D:4D ratio between youth with elevated prenatal androgen exposure due to congenital adrenal hyperplasia and controls. *Hormones and Behavior*, 128, Article 104908. <https://doi.org/10.1016/j.yhbeh.2020.104908>
- Nave, G., Nadler, A., Zava, D., & Camerer, C. (2017). Single-dose testosterone administration impairs cognitive reflection in men. *Psychological Science*, 28, 1398–1407.
- Nyby, J. G. (2008). Reflexive testosterone release: A model system for studying the nongenomic effects of testosterone upon male behavior. *Frontiers in Neuroendocrinology*, 29, 199–210.
- Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, 90, Article 101614. <https://doi.org/10.1016/j.intell.2021.101614>
- Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, 51, 768–774.

- Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in Psychology*, 7, Article 1680. <https://doi.org/10.3389/fpsyg.2016.01680>
- Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychology Bulletin*, 20, 580–591.
- Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72, 271–324.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20, 147–168.
- Tuiten, A., van Honk, J., Verbaten, R., Laan, E., Everaerd, W., & Stam, H. (2002). Can sublingual testosterone increase subjective and physiological measures of laboratory-induced sexual arousal? *Archives of General Psychiatry*, 59, 465–466.
- 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*, 108, 3448–3452.
- van Leeuwen, B., Smeets, P., Bovet, J., Nave, G., Stieglitz, J., & Whitehouse, A. (2020). Do sex hormones at birth predict later-life economic preferences? Evidence from a pregnancy birth cohort study. *Proceedings. Biological Sciences*. 287, Article 20201756. <https://doi.org/10.1098/rspb.2020.1756>
- van Wingen, G. A., Zylicz, S. A., Pieters, S., Mattern, C., Verkes, R. J., Buitelaar, J. K., & Fernández, G. (2009). Testosterone increases amygdala reactivity in middle-aged women to a young adulthood level. *Neuropsychopharmacology*, 34, 539–547.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.