Supplemental Material for

Single dose testosterone administration impairs cognitive reflection in men

Gideon Nave, Amos Nadler, David Zava, Colin Camerer

1. Participants

There were n = 243 male-only participants. Most (217, 89%) were students from a southern Californian college. Non-student participants were community members from surrounding cities. n = 125 of subjects were randomly assigned to receive a standard dose of T and n = 118 received placebos of matched viscosity in a double blind exogenous administration paradigm.

Pre-screening criteria excluded everyone with relevant medical and psychological conditions (5α-reductase deficiency, Klinefelter's syndrome, brain tumor, cancer, psychiatric diagnosis/diagnoses, high blood pressure, liver disease, kidney disease, angina, cancer, hepatitis, renal/kidney impairment, history of epileptic seizures and hypersensitivity to soy/ alcohol), subjects using prescription drugs that may interfere with the study (oxyphenbutazone, insulin, corticosteroids, opioids), subjects who self-reported consuming illegal drugs or excessive alcohol in the last 24 hours and non-native English speakers.

Personal, demographic, and treatment expectancy characteristics of the two treatment groups are summarized in table S1 (note that 5 subjects did not report their age and were therefore excluded from all analyses in which age is used as a control variable). The right column of Table S1 also reports the p-value of two sample t-tests for differences between T and placebo group characteristics (a check on whether random assignment resulted in balance on all such variables). Two subjects (one from each treatment group) self-reported taking T treatment on a regular basis; all analyses include these subjects and are robust to excluding them.

Table S1: Self-reported demographic data summary (standard errors in parentheses)

	All	T	Placebo	p-values for t-test of difference
N	243	125	118	
Age	23.63	24.42	22.78	0.08
	(0.46)	(0.77)	(0.49)	
Left-handed (proportion)	0.074	0.064	0.085	0.54
	(0.02)	(0.02)	(0.03)	
Heterosexual (proportion)	0.90	0.91	0.89	0.56
	(0.02)	(0.03)	(0.03)	
Treatment expectancy[1]	2.76	2.67	2.85	0.16
	(0.06)	(0.08)	(0.09)	
Married (proportion)	0.08	0.09	0.08	0.74
	(0.02)	(0.03)	(0.03)	
In a relationship (proportion)	0.38	0.34	0.42	0.20
	(0.03)	(0.05)	(0.04)	
Has children	0.06	0.08	0.04	0.23
	(0.02)	(0.02)	(0.02)	
Personal monthly income[2]	2.05	2.02	2.07	0.84
	(0.11)	(0.14)	(0.16)	

1 5 point scale, 1 = definitely did not get testosterone, 2 = probably not, 3 = unsure, 4 = probably got testosterone, 5 = definitely got testosterone

2 5 point scale, 1 = under $500/month, 2 = $501-$1,000, 3 = $1,001-$1,500/month, 4 = $1,501-$2,000/month 5 = over $2001/month

2. Hormonal assay procedure

Salivary steroids (estrone, estradiol, estriol, testosterone, androstenedione, DHEA, 5-alpha DHT, progesterone, 17OH-progesterone, 11-deoxycortisol, cortisol, cortisone, and corticosterone) were measured by LC-MS/MS using an AB Sciex Triple Quad 5500. Internal standards were added to 1 ml of saliva and the steroids then extracted by C18 column chromatography with 0.1 M NH4OH wash followed by 10% acetone. Steroids were eluted from the SPE with 10% methanol in acetone and dried under nitrogen. The dried samples were subjected to derivatization—the process of transforming a compound into a derivative product of similar chemical structure—with pyridine-3-sulfonyl chloride for the estrogens (estrone (E1), estradiol (E2), and estriol (E3)) as outlined by Xi and Spink (2008). 40 µL sodium bicarbonate (50mM, pH 10) and 40 µL pyridine-3-sulfonyl chloride (3 mg/mL in acetonitrile) were added to the dried samples, and incubated at 60oC for 10 minutes. After derivatization, the samples were diluted with 80 µL of water

and injected for LC-MS/MS analysis with analytical separation performed on an Agilent Poroshell 120 EC-C8 column and ionization by atmospheric pressure chemical ionization (APCI) in the positive ionization mode. Table S2 lists each analyte along with its validation results for the lower limit of quantitation (LLOQ is jargon for the lowest level of detection with coefficients of variation (CVs) < 20% over the linear range), linear range, and the inter-assay precision from the highest concentration to the LLOQ within the linear range. When salivary hormone levels of participants were below their LLOQ, we assigned values halfway between zero and their respective LLOQ (note that the true quantities of the hormone in the sample are never zero, even when they do not reach the detection threshold).

Table S2: Detection levels, precision and normality tests of hormonal assay

Analyte	LLOQ	Range	Precision	Proportion undetected, pre-treatment sample A	Proportion undetected, first post-treatment sample B	K-S test p-value	K-S test (log) p-value
Estrone pg/mL	0.5	0.5 - 510	8.7 - 13.7%	0.132	0.257	<0.01	0.56
Estradiol pg/mL	0.3	0.3 - 510	4.3 - 18.7%	0.128	0.329	0.06	0.88
Testosterone pg/mL	3.0	3.0 - 5100	3.0 - 18.1%	0	0.008	$<10^{-20}$	<0.01
Androstenedione pg/mL	5.0	5.0 - 2300	5.2 - 6.6%	0	0.008	$<10^{-20}$	0.008
DHEA pg/mL	20.0	20.0 - 1800	4.1 - 15.2%	0.004	0.012	0.002	0.98
DHT pg/mL	10.0	10.0 - 920	3.6 - 17.7%	0.786	0.473	$<10^{-11}$	0.02
Progesterone pg/mL	10.0	10.0 - 10000	4.8 - 10.8%	0.794	0.753	<0.01	0.03
17OH-Progesterone pg/mL	5.0	5.0 - 630	3.9 - 13.8%	0.004	0.061	0.003	0.98
11-Deoxycortisol pg/mL	5.0	5.0 - 410	6.8 - 16.6%	0.132	0.473	<0.01	0.04
Cortisol ng/mL	0.1	0.1 - 52	5.1 - 17.9%	0	0.008	<0.01	0.92
Cortisone ng/mL	0.1	0.1 - 81	4.1 - 14.9%	0	0.008	0.07	0.59
Corticosterone pg/mL	5.0	5.0 - 1800	4.6 - 17.5%	0.313	0.312	<0.01	0.08
Aldosterone pg/mL	10.0	10.0 - 560	8.9 - 18.8%	0.272	0.272	<0.06	0.39
Melatonin pg/mL	2.5	2.5-10000	5.2 - 15.9%	0.502	0.500	0.07	0.14

Note: P-values are calculated using a Kolmogorov-Smirnov test for the distributions of the second saliva sample compared to Gaussian, and for the log-transform (the null hypothesis is normal Gaussian distribution).

3. Hormonal changes following treatment and manipulation check

As expected, there were significant post-treatment differences between groups with respect to all hormones influenced by T treatment, either as an upstream (androstenedione) or downstream (5-alpha DHT) metabolite of T (S*1*). There was also a decrease in progesterone 170H resulting from an increase in T (which is common, according to personal communication from ZRT Laboratories chief scientist Dr. David Zava). The changes in saliva T measures were similar in magnitude to those reported in previous studies following topical gel administration of T and progesterone, e.g. (S*2, S3*).

We observed no significant differences between treatment groups in hormones that were not expected to change following short-term T treatment (e.g., aldosterone, cortisol, cortisone, melatonin) in all four saliva measurements throughout the experiment (i.e., the pre-treatment and the three post-treatment measurements). The pre-treatment and first post-treatment mean hormonal saliva levels are summarized in table S3; note that differences between morning and afternoon hormonal levels were affected by diurnal cycles in both treatment groups.

From assays conducted during the first 13 (out of 17) sessions of the study, we noted that 72 out of 184 pre-treatment baseline saliva samples (in both treatment groups) presented measurements with higher T level that are expected in normal young men (greater than 400 pg/mL). All other measurements (including T metabolites) were hormonally typical. The effects of T on the CRT were robust to excluding the subjects with abnormal measurements (see below).

We traced the cause of these abnormal measurements to T gel transfer to common surfaces (e.g., door knobs, mouse pads). Crucially, the high measurements were caused by local spread of T into saliva *tubes*, but *physiological* levels were unaffected by superficial contact with the dry nuisance T gel, as (a) we observed normal pre-treatment levels of T metabolites, namely DHT and androstenedione in all subjects; (b) none of the placebo group participants showed abnormally high values of T metabolites in any of the post-treatment measurements; (c) Only five out of 118 subjects from the placebo group showed consistently elevated T measurements in all of the 3 post-treatment saliva samples; (d) previous investigations found that interpersonal T transfer is highly unlikely even with skin-to-skin contact, (S*4*). Thus, we found convergent evidence that *biofluid* levels were unaffected by superficial contact. This conclusion was supported by ZRT Laboratories chief scientist Dr. David Zava.

In response to this finding during the course of the experimental period, we identified all surfaces and objects through which T could spread in the facility and improved sterile isolation protocol to eliminate the spread of the dried T gel. This protocol included thorough cleaning of keyboards, computer mice, chair backs, displays, and all doorknobs with a bleach-alcohol solution after each session as well as asking subjects to carefully wipe hands with a wet tissue before collecting each saliva sample. New pens were used for each session while all previously used pens were removed from testing area. Clipboards and other miscellaneous objects that participants did or could interact with

were cleaned, and an aerosol "air sanitizer" that bonds to VOCs (volatile organic compounds) was sprayed into the air. Following the adoption of this strict sterilization protocol, we found a drastic reduction in incidence of high T samples in the pre-treatment measurements, to a total of 5 participants out of 58 in the following four sessions (sessions 14-17).

Finally, we conducted additional robustness checks by examining the effects of T on the CRT when (a) excluding subjects with pre-treatment saliva T of greater than 400 pg/ml from both treatment groups; (b) excluding placebo subjects with post-treatment saliva T (sample B) greater than 400 pg/ml; (c) excluding all subjects in either condition (a) or (b); and (d) repeating the analysis with a more conservative cutoff of 250 pg/ml. We found that the effect of T administration on the CRT were highly significant (all p's<0.02) regardless of the exclusion criteria used.

Table S3: hormone panel data measurements log(pg/mL) summary statistics (standard errors in parentheses)

	Placebo		**Testosterone**		**Two-tailed p-value from t-test of T-Placebo equality**	
Sampling time[1]	**9am**	**2pm**	**9am**	**2pm**	**9am**	**2pm**
Testosterone	**5.743**	**5.111**	**5.580**	**8.373**	**0.267**	**1.06E-13**
	(0.094)	**(0.085)**	**(0.084)**	**(0.151)**		
Androstenedione	**4.510**	**4.205**	**4.525**	**5.462**	**0.634**	**3.11E-09**
	(0.039)	**(0.044)**	**(0.034)**	**(0.084)**		
DHT	**1.984**	**1.867**	**1.905**	**3.482**	**0.745**	**2.38E-06**
	(0.069)	**(0.051)**	**(0.060)**	**(0.114)**		
Progesterone	1.937	2.002	1.829	1.883	0.36	0.41
	(0.058)	(0.064)	(0.052)	(0.055)		
Progesterone170H	**3.245**	**2.675**	**3.217**	**2.463**	**0.792**	**0.008**
	(0.050)	**(0.058)**	**(0.049)**	**(0.058)**		
Estrone	-0.088	-0.557	-0.007	-0.389	0.29	0.42
	(0.063)	(0.066)	(0.064)	(0.056)		
Estradiol	-0.743	-1.158	-0.766	-1.066	0.86	0.44
	(0.052)	(0.059)	(0.054)	(0.054)		
DHEA	5.198	4.570	5.116	4.557	0.30	0.76
	(0.053)	(0.058)	(0.051)	(0.054)		
Deoxycortisol11	2.579	1.650	2.568	1.584	0.66	0.35
	(0.079)	(0.072)	(0.083)	(0.064)		
Cortisol	1.047	0.062	1.045	0.077	0.68	0.81

	(0.058)	(0.065)	(0.057)	(0.058)		
Cortisone	2.539	1.952	2.539	2.003	0.70	0.76
	(0.030)	(0.060)	(0.034)	(0.050)		
Corticosterone	2.442	1.274	2.646	1.290	0.37	0.76
	(0.126)	(0.065)	(0.123)	(0.060)		
Aldosterone	2.640	2.516	2.634	2.395	0.82	0.14
	(0.067)	(0.071)	(0.068)	(0.066)		
Melatonin	1.045	0.276	1.221	0.353	0.27	0.23
	(0.093)	(0.029)	(0.101)	(0.051)		

1 Main effects of time (afternoon vs. treatment) are due to the diurnal cycles of the hormones (*S5-S7*)

4. Results

4.1 Mood questionnaire

Table S4 shows a modest decrease in both affect measures over time (morning vs. afternoon), and no treatment or time x treatment interaction, indicated by the output of 2-way analysis of variance (ANOVA) with an interaction term, ruling out this indirect way in which T might affect cognition and behavior. Three subjects did not answer all of the negative affect items in their questionnaires, and five subjects did not complete all of the positive affect items; these subjects were excluded from analyses that include these scales as control variables.

Table S4: Positive / negative affect (PANAS-X) summary statistics

	All		**Testosterone**		**Placebo**		**ANOVA: p-values**		
Time	**Morning**	**Afternoon**	**Morning**	**Afternoon**	**Morning**	**Afternoon**	**T**	**Time**	**T x time**
Positive affect	2.72	2.61	2.72	2.63	2.72	2.60	0.85	0.16	0.85
	(0.05)	(0.06)	(0.06)	(0.08)	(0.07)	(0.09)			
Negative affect	1.53	1.45	1.53	1.46	1.53	1.43	0.77	0.13	0.84
	(0.04)	(0.04)	(0.06)	(0.05)	(0.05)	(0.05)			

4.2 Cognitive reflection test

CRT scores were comparable to those previously found in equivalent samples (Brañas-Garza et al. 2015), although at the high end of the range. This is likely due to high analytical skill in the sampled college population (conducted in one of the top ranked schools in the US). We tested our main hypothesis by estimating linear regression models with the three-item total CRT score as the dependent variable (DV). All of the analyses were conducted using the function 'lm' implemented in 'R' and the results are

summarized in table S5. Model A1 included only treatment (testosterone=1, placebo=0) as an independent variable (IV); Model A2 also included the math task performance. Model A3 also included age, positive and negative affect, treatment expectancy and the right hand digit ratios (the results hold when the left hand or the average between the two hands are used). Model A4 included all of the IVs of model A3 with the addition of all of the hormonal levels that were not affected by the treatment, as measured from the first post-treatment saliva sample (i.e., the second overall sample); all of the results hold when the measurements are replaced with the second post-treatment saliva sample (i.e., the third overall measurement, see table S7).
In models (B1-B4), summarized in table S6, we repeated the analyses of models (A1)-(A4), where the binary treatment variable was replaced by the measurements of the hormones that are affected by the treatment (T, DHT, androstenedione and progesterone 170H).

Finally, models (C1-C2) in Table S7 replicate the results of models (A4) and (B4) using the hormonal measurements of from the second post-treatment (and third overall) saliva sample.

4.3 CRT, question level

We further examined the effect of T on each of the three CRT questions separately. For each question, we classified the responses as either (a) an intuitive incorrect answer, i.e., 10 cents in the "bat and the ball" question, 100 minutes in the "widgets" question, 24 in the "lily pads" question; (b) the reflective, correct answer, i.e., 5 cents in the "bat and the ball" question, 5 minutes in the "widgets" question, 47 in the "lily pads" question; or (c) another incorrect answer, i.e., different than in (a) or (b).

We estimated two logistic regressions for each question, one that included a binary DV that was equal "1" for incorrect intuitive answers and the other included a binary DV that was equal "1" for correct answers. The analysis revealed that the likelihood of the incorrect intuitive response was significantly greater in the T group for each one of the three questions and that the proportion of correct answers was greater in the placebo group for each of the CRT questions in isolation (see Fig. 1 and table S8). Intriguingly, both of the subjects who self-reported taking T supplements regularly (one from each group) scored 0 out of 3 in the CRT, and all of their answers were the incorrect intuitive ones. Although the latter finding suggests that long term T treatment might have larger effects on CRT performance compared to a single dose, the small number of such subjects does not allow for making inferences that can be considered more than anecdotal. Moreover, as the long-term treatment was not assigned at random, causality cannot be inferred from these two data points (e.g., it is possible that subjects with low CRT scores are more likely to use T supplements, rather than vice versa).

Table S5: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the first post-treatment saliva sample.

	(A1)	(A2)	(A3)	(A4)
Treatment	-0.438***	-0.441***	-0.356**	-0.296**
	(0.142)	(0.136)	(0.142)	(0.145)
Math		0.079***	0.078***	0.073***
		(0.017)	(0.017)	(0.018)
Negative affect			0.058	-0.049
			(0.132)	(0.136)
Positive affect			-0.142*	-0.139*
			(0.075)	(0.076)
Age			-0.030***	-0.035***
			(0.010)	(0.010)
Treatment expectancy			0.062	0.080
			(0.076)	(0.076)
Digit ratio (right)			-1.456	-0.901
			(2.029)	(2.038)
Estrone				0.017
				(0.114)
Estradiol				0.147
				(0.119)
DHEA				-0.139
				(0.149)
Progesterone				-0.012
				(0.108)
Deoxycortisol11				0.215

				(0.133)
Cortisol				-0.278*
				(0.228)
Cortisone				-0.300*
				(0.298)
Corticosterone				0.052
				(0.120)
Aldosterone				0.209**
				(0.098)
Melatonin				0.0002
				(0.154)
Constant	2.102***	1.248***	3.425*	3.692
	(0.102)	(0.203)	(1.965)	(2.152)
Observations	243	243	229	227
R^2	0.038	0.122	0.150	0.205
Adjusted R^2	0.034	0.115	0.123	0.140
Residual Std. Error	1.109 (df = 241)	1.062 (df = 240)	1.050 (df = 221)	1.031 (df = 209)
F Statistic	9.447*** (df = 1; 241)	16.665*** (df = 2; 240)	5.582*** (df = 7; 221)	3.173*** (df = 17; 209)

Note: $^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Table S6: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the first post-treatment saliva sample.

	(B1)	(B2)	(B3)	(B4)
Testosterone	-0.205***	-0.202***	-0.184***	-0.198***
	(0.067)	(0.064)	(0.067)	(0.067)
Androstenedione	0.226	0.180**	0.150	0.204
	(0.165)	(0.159)	(0.164)	(0.169)
DHT	-0.006	-0.046	0.073	0.089
	(0.087)	(0.084)	(0.084)	(0.085)
Progesterone170H	-0.034	-0.043	-0.073	-0.055
	(0.117)	(0.112)	(0.115)	(0.127)
Math		0.076***	0.075***	0.073***
		(0.017)	(0.018)	(0.018)
Negative affect			0.007	-0.095
			(0.131)	(0.136)
Positive affect			-0.143**	-0.148*
			(0.075)	(0.076)
Age			-0.029***	-0.033***
			(0.010)	(0.010)
Treatment expectancy			0.067	0.078
			(0.076)	(0.076)
Digit ratio (right)			-0.547	0.058
			(2.018)	(2.032)
Estrone				0.034
				(0.114)
Estradiol				0.157
				(0.112)

DHEA				-0.141
				(0.149)
Progesterone				-0.010
				(0.107)
Deoxycortisol11				0.211
				(0.133)
Cortisol				-0.314
				(0.229)
Cortisone				-0.274
				(0.306)
Corticosterone				0.078
				(0.119)
Aldosterone				0.222**
				(0.096)
Melatonin				0.038
				(0.155)
Constant	2.265***	1.564**	2.997	2.840
	(0.503)	(0.506)	(2.031)	(2.181)
Observations	241	241	227	227
R^2	0.056	0.133	0.164	0.227
Adjusted R^2	0.040	0.115	0.125	0.152
Residual Std. Error	1.098 (df = 236)	1.054 (df = 235)	1.04 (df = 216)	1.024 (df = 206)
F Statistic	3.485*** (df = 4; 236)	7.233*** (df = 5; 235)	4.24*** (df = 10; 216)	3.031*** (df = 20; 206)

Note: $^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Table S7: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the second post-treatment saliva sample.

	(C1)	(C2)
Treatment	-0.314**	
	(0.148)	
Testosterone		-0.169**
		(0.063)
Androstenedione		0.002
		(0.156)
DHT		0.167
		(0.104)
Progesterone170H		-0.225
		(0.139)
Math	0.075***	0.075***
	(0.018)	(0.018)
Negative affect	0.014	0.014
	(0.136)	(0.135)
Positive affect	-0.099	-0.103
	(0.077)	(0.076)
Age	-0.028***	-0.032***
	(0.010)	(0.011)
Treatment expectancy	0.054	0.043
	(0.077)	(0.076)
Digit ratio (right)	-1.648	-0.711
	(2.078)	(2.064)

Estrone	-0.076	-0.040
	(0.110)	(0.109)
Estradiol	0.125	0.163
	(0.124)	(0.124)
DHEA	-0.137	-0.099
	(0.131)	(0.130)
Progesterone	0.124	0.124
	(0.127)	(0.125)
Deoxycortisol11	0.285	0.393**
	(0.131)	(0.137)
Cortisol	-0.218	-0.329*
	(0.167)	(0.172)
Cortisone	-0.218	-0.209
	(0.167)	(0.225)
Corticosterone	0.293*	0.322**
	(0.146)	(0.145)
Aldosterone	0.069	0.042
	(0.149)	(0.147)
Melatonin	-0.015	0.048
	(0.165)	(0.165)
Constant	3.548	3.588
	(2.189)	(2.217)
Observations	228	228
R^2	0.200	0.234
Adjusted R^2	0.135	0.160
Residual Std. Error	1.039 (df = 210)	1.023 (df = 207)
F Statistic	3.086*** (df = 17; 211)	3.164*** (df = 20; 207)

Note: $^{*}p<0.1$ $^{**}p<0.05$ $^{***}p<0.01$

Table S8: CRT score response frequencies and statistics by question

	Testosterone		Placebo		Logistic regression stats: (intuitive = 1)		Logistic regression stats (deliberate = 1)	
Question	**% intuitive**	**% deliberate**	**% intuitive**	**% deliberate**	**z**	**p-value**	**z**	**p-value**
Bat and ball	0.42	0.53	0.31	0.66	2.05	0.04	-2.69	0.009
Widgets	0.42	0.45	0.30	0.59	2.057	0.04	-2.64	0.008
Lili pads	0.20	0.67	0.10	0.82	2.24	0.02	-2.77	0.006

4.4 Response times

The T group responded 6.74 seconds slower on average when making correct answers (T: 50.97s, placebo: 44.23s) and 7.11 second faster on average when providing incorrect answers (T: 51.35s, placebo: 58.46s). A Kolmogorov-Smirnov test revealed that the response times (RT) were highly non-Gaussian ($p < 10^{-18}$). Therefore the values were log-transformed for normalization purpose before statistical tests (post-transformation Kolmogorov-Smirnov test, $p = 0.25$). To formally examine the treatment's effect on RT, we estimated linear mixed model regressions with log(RT) as the dependent variable (DV), and treatment (binary variable), error indicator (incorrect = 1, correct = 0) and the interaction between those binary treatment and error dummies as independent variables (fixed effects). Random effects of subject and question number were also included (see table S9). The main treatment coefficient (models D1 and D2) was insignificant, implying that T subjects did not differ in their general response times relative to placebo. However, the interaction between treatment and incorrect answers was negative and marginally significant at the $p < 0.10$ level. That is, T group subjects adopted their incorrect intuitions more rapidly when providing incorrect answers in the CRT.

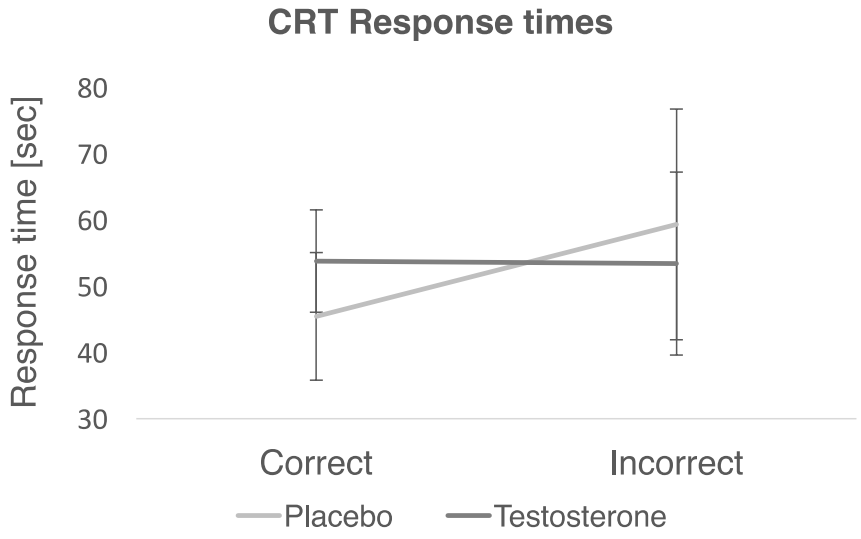
Additional simple slope analyses conducted separately for correct trials (456 observations) and incorrect trials (273 observations) revealed a marginally significant (at the $p < 0.10$ level) positive treatment effect in correct trials, implying that T treated participants were slower to respond when correct (model D3), and a negative, yet insignificant treatment effect in incorrect trials (model D4). The absence of statistically significant treatment effects in the latter might be due to the small number of observations and the noisiness of the RT measure in a task such as the CRT.

We did not directly record response time per-item in the math task. However, as the task was timed (5 minutes per participant) we could approximate the average response time per subject by dividing 300 (5 minutes) by the number of questions. We did not find differences between the two groups in this measure (P: 22.3 Sec; T = 21.4 Sec), $t(241) = 1.034$, $p > 0.30$).

Table S9: Mixed model linear regression. Dependent variable: logged response times, with subject and question random intercept

	Dependent variable:			
	log(Response time)			
	(D1)	(D2)	(D3)	(D4)
	All trials	All trials	Correct only	Incorrect only
Treatment	0.054	0.125	0.135*	-0.052
	(0.065)	(0.078)	(0.078)	(0.105)
Incorrect		0.089		
		(0.086)		
Treatment:Incorrect		-0.189*		
		(0.115)		
Constant	3.605***	3.578***	3.596***	3.732***
	(0.071)	(0.076)	(0.119)	(0.157)
Observations	729	729	456	273
Log Likelihood	-776.893	-778.680	-461.583	-303.984
Akaike Inf. Crit.	1,563.785	1,571.360	933.165	617.968
Bayesian Inf. Crit.	1,586.743	1,603.502	953.778	636.016
Note:				*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Figure S1: Response times (in seconds) as a function of treatment and response correctness. Error bars denote 95% confidence intervals.



5. References

S1. R. Horton, J. Tait, Androstenedione production and interconversion rates measured in peripheral blood and studies on the possible site of its conversion to testosterone. *Journal of Clinical Investigation* 45, 301 (1966).

S2. A. Mayo, H. Macintyre, A. Wallace, S. Ahmed, Transdermal testosterone application: pharmacokinetics and effects on pubertal status, short-term growth, and bone turnover. *The Journal of Clinical Endocrinology & Metabolism* 89, 681 (2004).

S3. J. Y. Du *et al.*, Percutaneous progesterone delivery via cream or gel application in postmenopausal women: a randomized cross-over study of progesterone levels in serum, whole blood, saliva, and capillary blood. *Menopause* 20, 1169 (2013).

S4. C. Rolf, U. Knie, G. Lemmnitz, E. Nieschlag, Interpersonal testosterone transfer after topical application of a newly developed testosterone gel preparation. *Clinical endocrinology* 56, 637 (2002).

S5. S. Hurwitz, R. J. Cohen, G. H. Williams, Diurnal variation of aldosterone and plasma renin activity: timing relation to melatonin and cortisol and consistency after prolonged bed rest. *Journal of Applied Physiology* 96, 1406 (2004).

S6. F. Hucklebridge, T. Hussain, P. Evans, A. Clow, The diurnal patterns of the adrenal steroids cortisol and dehydroepiandrosterone (DHEA) in relation to awakening. Psychoneuroendocrinology 30, 51 (2005).

S7. S. Nomura, M. Fujitaka, N. Sakura, K. Ueda, Circadian rhythms in plasma cortisone and cortisol and the cortisone/cortisol ratio. Clinica chimica acta 266, 83 (1997).