Weak and variable effects of exogenous testosterone on cognitive reflection task performance in three experiments: Commentary on Nave et al. (2017)

Erik L. Knight [a,b,1]*, Blakeley B. McShane[c,1,]*, Hana H. Kutlikova [d]*, Pablo J. Morales [a], Colton B. Christian [a], William T. Harbaugh [e], Ulrich Mayr [a], Triana L. Ortiz [f], Kimberly Gilbert [f], Christine Ma-Kellams[g], Igor Riečanský[d,h], Neil V. Watson [i], Christoph Eisenegger [d,2], Claus Lamm[d], Pranjal H. Mehta [a,j], & Justin M. Carré[f*]

a. Department of Psychology, University of Oregon
b. Center for Healthy Aging, Pennsylvania State University
c. Kellogg School of Management, Northwestern University
d. Department of Basic Psychological Research and Research Methods, University of Vienna
e. Department of Economics, University of Oregon
f. Department of Psychology, Nipissing University
g. Department of Psychology, San Jose State University
h. Centre of Experimental Medicine, Slovak Academy of Sciences
i. Department of Psychology, Simon Fraser University
j. Department of Experimental Psychology, University College London

*Correspondence to:
Erik L. Knight
elk24@psu.edu

Blakeley B. McShane
b-mcshane@kellogg.northwestern.edu

Hana H. Kutlikova
hana.kutlikova@univie.ac.at

Justin M. Carré
justinca@nipissingu.ca

Notes:
1. Shared first authorship
2. Deceased February 27, 2017

Abstract

Nave and colleagues (2017) presented a single experiment (*n*=243) finding that exogenous testosterone caused a decrease in performance on the Cognitive Reflection Test (CRT) by increasing intuitive-but-incorrect responses. We report three new experiments (total *n*=628) that also examine the effect of exogenous testosterone on CRT performance. When pooling the data across experiments, we find (i) substantial variation in CRT performance across experiments, treatment groups, and participants and (ii) variable treatment effects of testosterone on CRT performance across experiments with any average effect being weak relative to this underlying variability – regardless of whether we considered the three new experiments or all four. Given our modeling assumptions, an average treatment effect of a 7% decrease in the odds of correctly responding to a CRT item is the value most compatible with the data from the three new experiments; however, anything from a 53% decrease to a 99% increase is also reasonably compatible. Similarly, a 27% decrease is the value most compatible with the data from all four experiments; however, anything from a 62% decrease to a 58% increase is also reasonably compatible. We explore potential explanations for the pattern of results observed across the four experiments.

Keywords: exogenous testosterone; cognitive reflection; replication; meta-analysis

Introduction

Testosterone is associated with behaviors such as aggression, sensation seeking, and impulse control disorders, including drug addiction and eating disorders, but if and how testosterone affects cognition and decision-making remains unclear. Given the role of testosterone in mating and reproduction, Nave, Nadler, Zava, and Camerer (2017; hereafter NNZC) suggested the "facilitation of rapid intuitive responses by testosterone could be biologically adaptive in contexts in which reproductive success depends on instincts (e.g., during copulation) and when responding slowly might be especially costly (e.g., during physical challenges)" (p. 1404). This led them to hypothesize that testosterone biases decision making away from reflective and deliberate responses and toward rapid and intuitive ones, thereby elucidating one potential mechanism by which testosterone might cause behaviors such as aggression, sensation seeking, and impulse control disorders.

To study their hypothesis, NNZC conducted a single experiment (n=243) in which they randomly administered either exogenous testosterone or placebo to participants and then measured their performance on the Cognitive Reflection Test (CRT), a simple three-item assessment of intuitive versus deliberate decision making (Frederick, 2005). Each CRT item has an intuitive but incorrect response with which most people respond; discerning the correct response requires one to inhibit this intuitive response and to perform deliberate but easy calculations. For example, one item reads

> *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?*

When asked this question, many people automatically respond with the perhaps intuitive but ultimately incorrect response of 24 days; discerning the correct response (47 days) requires inhibiting the automatic response and deliberating on the question. Consistent with their hypothesis, the NNZC experiment found that exogenous testosterone caused a decrease in performance on the CRT by increasing intuitive-but-incorrect responses.

We report three new experiments (total n=628) that also examine the effect of exogenous testosterone on CRT performance. When pooling the data across experiments, we find (i) substantial variation in CRT performance across experiments, treatment groups, and participants and (ii) variable treatment effects of testosterone on CRT performance across experiments with any average effect being weak relative to this underlying variability – regardless of whether we considered the three new experiments or all four. We explore potential explanations for the pattern of results observed across the four experiments.

Materials for the three new experiments; data from the three new experiments and the original NNZC experiment; and code that implements all analyses presented in this manuscript and in our Supplementary Online Materials, which provides further detail on Methods, Results, and other matters related to this manuscript, are available at https://osf.io/6ppdv/. Materials for and data from the original experiment were obtained from the corresponding publication, supplemental materials (https://osf.io/79r2v), and personal communication with Nave.

Methods

The three new experiments were designed independently of and executed prior to the publication of NNZC and therefore differ with one another and with NNZC with regards to the experimental design as discussed below (**Table S1**). Like NNZC, these experiments included tasks completed prior to the CRT as part of larger protocols, including competitive and pro-

social decision-making tasks (Experiment 1); an aggression task and public goods game (Experiment 2); and emotion recognition tasks, empathy tests, and pro-social decision-making tasks (Experiment 3).

Testosterone versus Placebo Administration

Exogenous testosterone or placebo was administered to three samples of men aged 18-41 (Experiment 1: *n*=116, Oregon, USA; Experiment 2, *n*=396, Ontario, Canada; Experiment 3, *n*=116, Bratislava, Slovakia) prior to the CRT topically (150-mg dose, Experiments 1 and 3) or intranasally (11-mg dose, Experiment 2).

Cognitive Reflection Task

Each of the three new experiments presented the CRT items in random order. Experiment 3 presented them in Slovak, the native language of the participants. Experiments 1 and 2 did not include financial incentives for CRT performance, but, in an effort to increase attention and engagement with the task Experiment 3 did as did NNZC; specifically, Experiment 3 paid €0.30 per correct response, a value chosen to reflect the local, part-time job salary for students (€4 per hour at the time of experiment). We note that financial incentives may improve effort but not performance in laboratory experiments, or they may improve performance only for individuals with higher cognitive skills (Camerer & Hogarth, 1999). Consistent with this reasoning, a large-scale meta-analysis suggests that financial incentives do not impact CRT performance (Brañas-Garza et al., 2015).

Methodological Difference Variables

Experimental Manipulations. All manipulations in the three new experiments were randomized and administered prior to the CRT. Two of the experiments included manipulations

in addition to testosterone or placebo. Experiment 1 manipulated blinding by informing half of all participants (n=58) whether they had been assigned to testosterone or placebo; experimenters remained fully blind. Experiment 3 assigned participants to one of two experimental stressors (cold pressor, n=39; a socially-evaluated cold pressor, n=37) or control (warm pressor, n=40).

Experimenter Gender. Experiments 1 and 2 used male and female experimenters while Experiment 3 used female only; NNZC used male only. Limited prior work suggests that experimenter gender may alter testosterone levels and behavior in young men in an ecological setting (Ronay & von Hippel, 2010). Other work has shown such experimenter gender effects may generalize to a laboratory setting, but the effects may be weaker and may depend on the time of day (Roney et al., 2007). To our knowledge, no experiment has found that experimenter gender impacts the effect of testosterone treatment on behavior.

Time of Day. The new experiments administered the CRT at a range of times from approximately 11:00 AM to 7:00 PM; NNZC administered it at approximately 4:00 PM. In all experiments, the CRT was administered in the time period that pharmacokinetic analyses suggest should coincide with peak testosterone levels for each method (Eisenegger et al., 2013; Geniole et al., 2019). Although testosterone levels fluctuate with a diurnal rhythm, whether the time of day impacts the effect of testosterone treatment on CRT performance is unknown.

In sum, the three new experiments administered testosterone or placebo prior to the CRT, but they differed with one another and with NNZC with regard to some details. These differences provide a valuable opportunity to examine the generalizability of the NNZC finding regarding the effect of testosterone treatment on CRT performance across diverse experimental populations and designs as well as heterogeneity in the treatment effect that may result from these or other

unknown factors – an important consideration when conducting replications of psychological research studies (McShane et al., 2019).

Individual Difference Variables

Prior research reports that several individual difference variables including basal cortisol, the ratio between the lengths of the second and fourth digits (2D:4D ratio), and trait impulsivity may affect the relationship between testosterone and social cognition and behavior. Although these variables have not been examined in studies of the effect of testosterone on CRT performance, exploring their effects may provide insight into potential moderators that could be investigated in future studies.

Basal Cortisol. A recent meta-analysis suggests that testosterone is more strongly associated with status-relevant behavior when cortisol levels are low, though heterogeneity is evident in the direction and magnitude of this interaction effect across studies (Dekkers et al., 2019). In the three new experiments, basal cortisol was measured prior to testosterone or placebo administration.

2D:4D Ratio. The 2D:4D ratio is believed to be associated with prenatal testosterone exposure, which in turn may moderate the effects of testosterone administration on socio-cognitive behavior among men. Accordingly, prior work has reported reduced empathic accuracy in individuals with lower 2D:4D ratios (van Honk et al., 2011; Carré et al., 2015). In the three new experiments, participants' left and right hands were scanned on a flatbed scanner; trained research assistants digitally measured the lengths of the second and fourth digits between the ventral proximal creases of the digits to the fingertips.

Trait Impulsivity. Recent work reports that the effect of testosterone on reactive aggression is associated with trait impulsivity (Carré et al., 2017; Geniole et al., 2019). In the three new experiments, trait impulsivity was measured via three questionnaires: Experiment 1 used the impulsivity subscale of the Zuckerman-Kuhlman Impulsive Sensation-Seeking Scale (Zuckerman et al., 1993); Experiment 2 used a summed composite of the Barrett Impulsivity (Patton et al., 1995) and Brief Self-Control Scale (Tangey et al., 2004); and Experiment 3 used the fun-seeking subscale of the Behavioral Inhibition/Behavior Activation Scales (Carver & White, 1994).

Models

Primary. To estimate the effect of testosterone treatment on CRT performance, we meta-analyzed the data from the three new experiments as well as all four experiments by fitting a multilevel logistic regression to the response of each participant to each CRT item (Correct=1, Incorrect=0) jointly (McShane and Böckenholt, 2017; 2018). The model treated effects for the interaction of each item and primary treatment condition (i.e., testosterone or placebo) as "fixed" and effects for (i) each experiment across all items, (ii) each experiment for each item, (iii) each treatment group (i.e., primary treatment condition crossed with blinding or stressor condition as applicable) across all items, (iv) each treatment group for each item, and (v) each participant across all items as "random." We also expanded the model to directly compare the degree to which the treatment effect pooled across the three new experiments differed from the treatment effect in NNZC.

Secondary. For comparability with NNZC, we also meta-analyzed aggregated data. Specifically, we fit a multilevel linear regression specified *mutatis mutandis* analogously to our primary model to the score of each participant (i.e., number of CRT items correct out of three).

We also expanded the primary model to include covariates included by NNZC, namely age, treatment expectancy, right hand 2D:4D ratio, basal cortisol levels, positive/negative affect (Experiments 1, 3, and NNZC only), and mathematics aptitude (Experiment 1 and NNZC only). As in NNZC, we also report the effect of testosterone treatment separately for each CRT item as well as the effect on intuitive-but-incorrect responses (1=intuitive-but-incorrect, 0=all other responses) instead of correct responses.

We also examined potential moderators of the effect of testosterone on CRT performance. We did so for methodological differences across the experiments in two ways: (i) by re-fitting our primary model with the single-blind and stressor groups removed from Experiments 1 and 3 respectively and (ii) by expanding our primary model to include the interaction of each item, primary treatment condition, and various methodological difference variables [experimenter gender, time of day, experimental blinding conditions (Experiment 1), and experimental stressor conditions (Experiment 3)]. We did so for individual difference variables (basal cortisol, right- and left-hand 2D:4D ratio, and trait impulsivity) by expanding the model to include interactions in the same manner.

Models fit to subsets of experiments (e.g., because one or more did not measure a given variable) were specified analogously to our primary model with effects treated as random removed when they were not identified.

Estimation. We estimate all models in a fully Bayesian manner (Gelman, et al., 2013) and present point and 95% interval estimates for each coefficient or effect of interest. All estimates are presented on the scale of a logistic regression coefficient unless otherwise noted, with point estimates given by the median of the estimated posterior distribution and interval

estimates given by the 2.5 and 97.5 percentiles. Positive estimates imply better CRT performance.

Results

Distributions of CRT Performance

Experiments differed in terms of CRT performance as reflected in the mean (**Table S2**) and the distribution of the scores of the participants (**Figure S1** and **Table S3**) with lower performance in the three new experiments as compared to NNZC. The distributions in the three new experiments were relatively more similar to the distribution in a large meta-analysis (Brañas-Garza et al., 2015) whereas the distribution in NNZC was relatively more similar to the distributions in the highest performing samples in prior research (e.g., MIT and Princeton students; Frederick, 2005; Iyer et al., 2012).

Primary

We begin by discussing the estimates of the variance components from our primary model as they inform our discussion of the estimates of the treatment effect. These estimates indicated substantial variation in CRT performance from (i) experiment to experiment, thus reflecting the differences in CRT performance across experiments discussed above; (ii) treatment group to treatment group, thus reflecting differences in the treatment effect from experiment to experiment; and (iii) participant to participant, thus reflecting individual differences in CRT performance – regardless of whether we considered the three new experiments or all four (**Table S4**).

To illustrate the extent of this variation, we use our point estimates of the variance components to create three comparisons that correspond to each of the above three points respectively and that are scaled relative to the point estimate of the meta-analytic average

treatment effect. First, the difference in CRT performance from experiment to experiment was estimated to be 15.50 times larger than the meta-analytic average treatment effect when we considered the three new experiments or 4.01 times larger when we considered all four. Second, the difference in the treatment effect from experiment to experiment was estimated to be 6.71 times larger than the meta-analytic average treatment effect when we considered the three new experiments or 2.05 times larger when we considered all four. Third, the difference in CRT performance from participant to participant was estimated to be 46.25 times larger than the meta-analytic average treatment effect when we considered the three new experiments or 10.20 times larger when we considered all four. We note the larger relative estimates when we considered the three new experiments as compared to all four do not so much reflect differences in the estimates of the variance components; instead, they primarily reflect the scaling by the estimate of the meta-analytic average treatment effect, which, as we discuss immediately below, was considerably smaller when we considered the three new experiments as compared to all four.

Given this degree of variation, the meta-analytic average treatment effect was unsurprisingly estimated with considerable uncertainty regardless of whether we considered the three new experiments (*Point Estimate*: -0.07; *95%CI*: [-0.76, 0.69]; **Figure 1** and **Table S4**) or all four (-0.32; [-0.96, 0.46]; **Figure 1** and **Table S4**). Put differently, given our modeling assumptions, an average treatment effect of a 7% decrease in the odds of correctly responding to a CRT item is the value most compatible with the data from the three new experiments; however, anything from a 53% decrease to a 99% increase is also reasonably compatible. Similarly, a 27% decrease is the value most compatible with the data from all four experiments; however, anything from a 62% decrease to a 58% increase is also reasonably compatible. A comparison of the treatment effect in the three new experiments to the treatment effect in NNZC within our

modeling framework also resulted, again unsurprisingly, in an estimate with considerable uncertainty (-1.01; [-2.44, 0.53]; **Table S6**); while the point estimate suggests a stronger (i.e., more negative) treatment effect in NNZC as compared to the three new experiments, a similar or even weaker treatment effect is also reasonably compatible with the data from all four experiments given our modeling assumptions.

In sum, our results suggest variable treatment effects of testosterone on CRT performance across experiments with any average effect being weak relative to this underlying variability.

Secondary

CRT Score. The multilevel linear regression fit to the score of each participant (i.e., number of CRT items correct out of three) yielded results in line with those presented above (**Table S7**). Variance component estimates again indicated substantial variation across experiments, treatment groups, and participants. The meta-analytic average treatment effect was again estimated with considerable uncertainty regardless of whether we considered only the three new experiments (-0.03; [-0.31, 0.28]) or all four (-0.13; [-0.39, 0.21]). Put differently, given our modeling assumptions, an average treatment effect of a 0.03 point decrease in the score is the value most compatible with the data from the three new experiments; however, anything from a 0.31 point decrease to a 0.28 point increase is also reasonably compatible. Similarly, a 0.13 point decrease is the value most compatible with the data from all four experiments; however, anything from a 0.39 point decrease to a 0.21 point increase is also reasonably compatible.

Covariates, Individual Item Responses, and Intuitive-but-Incorrect Responses. Results remained substantively similar when controlling for covariates included by NNZC (**Table S8**). Results also remained substantively similar when the meta-analytic average treatment effect was broken down at the CRT item-level (**Table S4**). Finally, results remained

substantively similar when examining the effect of testosterone on intuitive-but-incorrect (as opposed to correct) responses both on average and at the item-level (**Figure S2** and **Table S9**).

Methodological Difference Variables. Estimates of variance components and the meta-analytic average treatment effect remained substantively similar to those from the primary model when we excluded the single-blind and stressor groups from Experiments 1 and 3 respectively (**Table S10**); the result concerning the stability of the variance component estimates is particularly notable because it suggests our conclusions regarding differences in the treatment effect from experiment to experiment are not driven by the single-blind or stressor conditions of the respective experiments. In addition, methodological difference variables showed no substantial moderating effects (**Table S11**).

Individual Difference Variables. Individual difference variables showed no substantial moderating effects (**Table S12**) but the results may suggest a moderating effect of trait impulsivity. Specifically, the effect of testosterone on CRT performance may be associated with trait impulsivity, with a potential negative treatment effect at lower levels of trait impulsivity and a potential positive treatment effect at higher levels of trait impulsivity (0.52; [0.06, 0.99]; **Figure S3** and **Table S12**).

Discussion

NNZC presented a single experiment suggesting that exogenous testosterone causes a decrease in CRT performance. We report three new experiments that also examine the effect of exogenous testosterone on CRT performance. When pooling the data across experiments, we find (i) substantial variation in CRT performance across experiments, treatment groups, and participants and (ii) variable treatment effects of testosterone on CRT performance across experiments with any average effect being weak relative to this underlying variability –

regardless of whether we considered the three new experiments or all four. The extent of this relative variability suggests the notion of *the* effect of testosterone on CRT performance is not particularly meaningful. Instead, to the degree that testosterone does affect CRT performance, focusing on potential moderators that drive this variability would seem to be of greater interest.

Our results suggest two possible moderators. First, CRT performance in the three new experiments was relatively more similar to that in a large meta-analysis whereas CRT performance in NNZC was relatively more similar to that in the highest performing samples in prior research. It is therefore possible that testosterone causes impaired CRT performance only in high performing populations. Second, our results suggest that trait impulsivity may moderate the effect of testosterone on CRT performance with a potential negative treatment effect at lower levels of trait impulsivity and a potential positive treatment effect at higher levels of trait impulsivity. This finding may be related to recent work suggesting that trait impulsivity moderates the effect of testosterone on reactive aggression but with a positive treatment effect at lower levels of trait impulsivity and a negative treatment effect at higher levels of trait impulsivity (Carré et al., 2017; Geniole et al., 2019).

It is perhaps of interest to consider these two possible moderators jointly and alongside prior work linking high CRT performance with low trait impulsivity (Frederick, 2005). Specifically, although trait impulsivity was not measured in NNZC, the participants in that higher performing sample may have been less impulsive and therefore more vulnerable to any negative effect of testosterone on CRT performance as compared to participants in the three new experiments. Nonetheless, this would suggest that testosterone causes impaired CRT performance only in populations low in trait impulsivity.

However, we urge due caution in interpreting our moderation results particularly given the number of experiments, the sample sizes of the experiments, and the number of moderator variables examined. We also note we examined the moderating effects only of variables studied either in NNZC or in testosterone research more broadly; other variables may moderate the effect of testosterone on CRT performance. Nonetheless, insofar as future research efforts continue to examine the effects of testosterone treatment on cognitive reflection – perhaps in search of such moderators – our results suggest something akin to a "one phenomenon, many labs" approach that features systematic variation of methodological difference variables and examines potential moderating effects of relevant variables in larger and more diverse samples seems necessary (McShane et al., 2019).

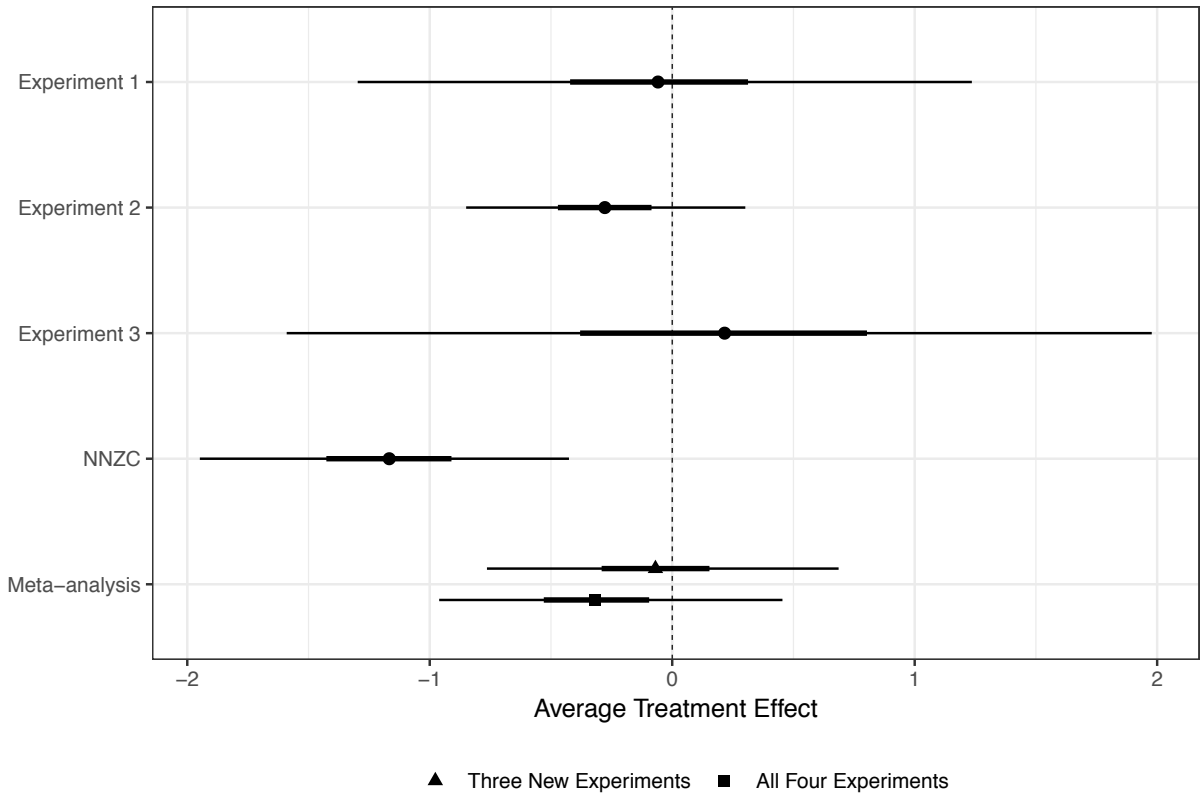

Figure 1. Primary Results. Point estimates are given by the circle; 50% and 95% interval estimates are given by the thick and thin lines, respectively. Estimates are based on models fit to data from each experiment separately and based on our primary meta-analytic model fit to the data from the experiments jointly (**Tables S4-S5**).

Author Contributions

Experiment 1: E.L. Knight, U. Mayr, W.T. Harbaugh, P.H. Mehta designed the experimental protocol, with input specific to the CRT from C. Ma-Kellams; E.L. Knight, C.B. Christian, P.J. Morales executed data collection. Experiment 2: J.M. Carré, T.L. Ortiz, N.V. Watson designed the experimental protocol; T.L. Ortiz performed hormone assays; T.L. Ortiz & K. Gilbert executed data collection. Experiment 3: H.H. Kutlikova, I. Riečanský, C. Eisenegger, C. Lamm designed the experimental protocol; H.H. Kutlikova executed data collection. E.L. Knight and B.B. McShane analyzed data for the manuscript. E.L. Knight wrote the initial manuscript; E.L. Knight and B.B. McShane wrote the revised manuscript(s). J.M. Carré, P.H. Mehta, J.M. Carré and H.H. Kutlikova wrote segments specific to their respective experiments in the Supplementary Online Materials methods section. All surviving authors approved the final version of the manuscript for submission. Please see Supplementary Online Materials for acknowledgements.

Funding

Experiment 1 was supported by National Science Foundation grants to PHM (#1451848) and to WTH and UM (#1063561). Experiment 2 was supported by funds from the Natural Sciences and Engineering Research Council of Canada to JMC (RGPIN-2014-06676), a Northern Ontario Heritage Fund Corporation grant to JMC, and a Discovery Grant (RGPIN-2016-05706) from the Natural Sciences and Engineering Research Council of Canada to NVW. Experiment 3 was supported by the Vienna Science and Technology Fund (WWTF VRG 13-007). ELK is partially supported by a National Institute on Aging Grant AG049676 to The Pennsylvania State University.

References

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2015). Cognitive Reflection Test: Whom, how, when (ESI Working Paper 15-25). Retrieved from https://mpra.ub.uni-muenchen.de/68049/

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, *19*(1-3), 7-42.

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*(4), 249-256.

Carré, J. M., Ortiz, T. L., Labine, B., Moreau, B. J., 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.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of personality and social psychology*, *67*(2), 319.

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 and Biobehavioral Reviews*, *96*, 250-271.

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.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25-42.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, and Rubin, Donald B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, Third edition.

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 psychopharmacogenomic approach to identify the pathways through which and people for whom testosterone promotes aggression. *Psychological Science, 30*(4), 481-494.

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*(8), 3448-3452.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS One*, *7*(8), e42366.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research*, *43*(6), 1048-1063.

McShane, B.B. and Böckenholt, U. (2018) Multilevel Multivariate Meta-analysis with Application to Choice Overload. *Psychometrika*, *83*(1), 255-271.

McShane, B. B., Tackett, J. L., Gelman, A., & Bockenholt, U. (2019). Large scale replication projects in contemporary psychological research. *The American Statistician*, *73:sup1*, 99-105.

Nave, G., Nadler, A., Zava, D., & Camerer, C. (2017). Single dose testosterone administration impairs cognitive reflection in men. *Psychological Science, 28*(10), 1398-1407.

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768-774.

Ronay, R., & Hippel, W. V. (2010). The presence of an attractive woman elevates testosterone and physical risk taking in young men. *Social Psychological and Personality Science*, *1*(1), 57-64.

Roney, J. R., Lukaszewski, A. W., Simmons, Z. L. (2007). Rapid endocrine responses of young men to social interactions with young women. *Hormones and Behavior, 52*, 326–333.

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*(2), 271-324.

Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., & Kraft, M. (1993). A comparison of three structural models for personality: The Big Three, the Big Five, and the Alternative Five. *Journal of personality and social psychology*, *65*(4), 757.