Commentary

Psychological Science
2020, Vol. 31(7) 890–897
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797619885607
www.psychologicalscience.org/PS

Weak and Variable Effects of Exogenous Testosterone on Cognitive Reflection Test Performance in Three Experiments: Commentary on Nave, Nadler, Zava, and Camerer (2017)

Erik L. Knight[1,2], Blakeley B. McShane[3], Hana H. Kutlikova[4], Pablo J. Morales[1], Colton B. Christian[1], William T. Harbaugh[5], Ulrich Mayr[1], Triana L. Ortiz[6], Kimberly Gilbert[6], Christine Ma-Kellams[7], Igor Riečanský[4,8], Neil V. Watson[9], Christoph Eisenegger[4], Claus Lamm[4], Pranjal H. Mehta[1,10], and Justin M. Carré[6]

[1]Department of Psychology, University of Oregon; [2]Center for Healthy Aging, The Pennsylvania State University; [3]Kellogg School of Management, Northwestern University; [4]Department of Cognition, Emotion, and Methods in Psychology, University of Vienna; [5]Department of Economics, University of Oregon; [6]Department of Psychology, Nipissing University; [7]Department of Psychology, San Jose State University; [8]Centre of Experimental Medicine, Slovak Academy of Sciences; [9]Department of Psychology, Simon Fraser University; and [10]Department of Experimental Psychology, University College London

Received 10/27/17; Revision accepted 9/5/19

Testosterone is associated with behaviors such as aggression and sensation seeking as well as behavioral disorders such as impulse-control disorders including drug addiction and eating disorders, but to what degree 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) suggested that 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 and behavioral disorders.

To study their hypothesis, Nave et al. 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?

Corresponding Authors:
Erik L. Knight, Center for Healthy Aging, The Pennsylvania State University, 423 Biobehavioral Health, University Park, PA 16802
E-mail: elk24@psu.edu

Blakeley B. McShane, Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208
E-mail: b-mcshane@kellogg.northwestern.edu

Hana H. Kutlikova, Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna
E-mail: hana.kutlikova@univie.ac.at

Justin M. Carré, Department of Psychology, Nipissing University, 100 College Drive, North Bay, Ontario, Canada, P1B 8L7
E-mail: justinca@nipissingu.ca

When asked this question, many people automatically respond with the perhaps intuitive but ultimately incorrect response of 24 days; discerning the correct response of 47 days requires inhibiting the automatic response and deliberating on the question. Nave et al. reported that the results of their experiment were consistent with their hypothesis, namely 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 found (a) substantial variation in CRT performance across experiments, treatment groups, and participants and (b) 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 can be found at https://osf.io/b7ngf/. Data from the three new experiments and the original Nave et al. experiment, as well as scripts that implement all analyses presented in this article and its Supplemental Material available online, can be found at https://osf.io/kefqp/. Materials for and data from Nave et al.'s experiment were obtained from the corresponding article; its permanent repository, which can be found at https://osf.io/jbq9v/; and personal communication with Nave. Further detail on methodology, results, and other matters related to this article are provided in our Supplemental Material.

Method

The three new experiments were designed independently of and executed prior to the publication of Nave et al. (2017) and therefore differ from one another and from Nave et al.'s experiment with regard to the experimental design, as discussed below (see Table S1 in the Supplemental Material for a comparison of the experiments). Like Nave et al.'s experiment, the three new experiments contained tasks completed prior to the CRT as part of larger protocols, including competitive and prosocial decision-making tasks (Experiment 1); an aggression task and public-goods game (Experiment 2); and emotion-recognition tasks, empathy tests, and prosocial decision-making tasks (Experiment 3).

Testosterone versus placebo administration

Prior to the CRT, exogenous testosterone or placebo was administered to three samples of men between the ages of 18 and 41 years (Experiment 1: N = 116, Oregon, United States; Experiment 2, N = 396, Ontario, Canada; Experiment 3, N = 116, Bratislava, Slovakia) either topically (150-mg dose, Experiments 1 and 3) or intranasally (11-mg dose, Experiment 2).

Cognitive Reflection Test

In the three new experiments, the CRT items were presented in random order. In Experiment 3, they were presented in Slovak, the native language of the participants. Financial incentives for CRT performance were not included in Experiments 1 and 2, but they were included in Experiment 3 as in Nave et al.'s experiment in an effort to increase attention and engagement with the task; specifically, €0.30 was paid per correct response in Experiment 3, 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, results from a large-scale meta-analysis suggest that financial incentives do not impact CRT performance (Brañas-Garza, Kujal, & Lenkei, 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. In Experiment 1, each participant was assigned to one of two blinding conditions (single blind, n = 58; double blind, n = 58); experimenters remained blind in both conditions, whereas participants were informed whether they had been assigned to receive testosterone or placebo in the single-blind condition. In Experiment 3, each participant was assigned to one of two experimental stressors (cold pressor, n = 39; socially evaluated cold pressor, n = 37) or to a control condition (warm pressor, n = 40).

Experimenter gender. Male and female experimenters were used in Experiments 1 and 2, whereas only female experimenters were used in Experiment 3; only male experimenters were used in Nave et al.'s experiment. 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 that 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, Lukaszewski, & Simmons, 2007). To what degree experimenter gender impacts the effect of testosterone treatment on CRT performance is unknown.

Time of day. The CRT was administered at a range of times from approximately 11:00 a.m. to 7:00 p.m. in the three new experiments; it was administered at approximately 4:00 p.m. in Nave et al.'s experiment. 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, von Eckardstein, Fehr, & von Eckardstein, 2013; Geniole et al., 2019). Although testosterone levels fluctuate with a diurnal rhythm, to what degree the time of day impacts the effect of testosterone treatment on CRT performance is unknown.

In sum, testosterone or placebo was administered in the three new experiments prior to the CRT following Nave et al.'s procedure, but the new experiments differed from one another and from Nave et al.'s experiment with regard to some details. These differences provide a valuable opportunity to examine the generalizability of the Nave et al. report 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, Tackett, Böckenholt, & Gelman, 2019).

Individual-difference variables

Prior research has shown that several individual-difference variables, including basal cortisol, the ratio between the lengths of the second and fourth digits of the hand (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 treatment on sociocognitive behavior among men. Accordingly, prior work has shown a negative effect of testosterone treatment on empathic accuracy in individuals with lower 2D:4D ratios (Carré et al., 2015; van Honk et al., 2011; but see also Nadler et al., 2019). 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 of each hand between the ventral proximal creases of the digits to the fingertips.

Trait impulsivity. Recent work shows that the effect of testosterone treatment 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, Kuhlman, Joireman, Teta, & Kraft, 1993), Experiment 2 used a summed composite of the Barratt Impulsiveness Scale (Patton & Stanford, 1995) and Brief Self-Control Scale (Tangney, Baumeister, & Boone, 2004), and Experiment 3 used the fun-seeking subscale of the Behavioral Inhibition System and Behavioral Activation System 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 & 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 (a) each experiment across all items, (b) each experiment for each item, (c) each treatment group (i.e., primary treatment condition crossed with blinding or stressor condition as applicable) across all items, (d) each treatment group for each item, and (e) 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 reported by Nave et al.

Secondary. For comparability with Nave et al.'s analysis, we also meta-analyzed aggregated data. We did so by fitting 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 our primary model to include covariates included by Nave et al., namely, age, treatment expectancy, right-hand 2D:4D ratio, basal cortisol levels, positive and negative affect (Experiments 1, 3, and Nave et al. only), and mathematics aptitude (Experiment 1 and Nave et al. only). Like Nave et al., we also report the effect of testosterone treatment separately for each CRT item as well as the effect on intuitive but incorrect responses (intuitive but incorrect = 1, all other responses = 0) instead of correct responses.

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

Models fitted 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% credible interval (CI) estimates for each parameter 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 (see Table S2 in the Supplemental Material) and the distribution of the scores of the participants (see Fig. S1 and Table S3 in the Supplemental Material); performance was lower in the three new experiments, compared with performance in Nave et al.'s experiment. 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 Nave et al.'s experiment was relatively more similar to the distributions in the highest-performing samples in prior research (e.g., Massachusetts Institute of Technology and Princeton University students; Frederick, 2005; Iyer, Koleva, Graham, Ditto, & Haidt, 2012).

Primary

We begin by discussing the estimates of the variance components from our primary model because they inform our discussion of the estimates of the treatment effect (see Table S4 in the Supplemental Material; estimates reported in this paragraph and the subsequent one are derived from the estimates of the variance components reported in Table S4 using the script named variability.analysis.R). These estimates indicated substantial variation in CRT performance from experiment to experiment, thus reflecting the differences in CRT performance across experiments discussed above, regardless of whether we considered the three new experiments (point estimate = 1.15, 95% CI = [0.54, 2.99]) or all four (point estimate = 1.30, 95% CI = [0.72, 2.92]). They also indicated substantial variation in CRT performance from treatment group to treatment group, thus reflecting differences in the treatment effect from experiment to experiment, regardless of whether we considered the three new experiments (point estimate = 0.43, 95% CI = [0.08, 1.34]) or all four (point estimate = 0.54, 95% CI = [0.12, 1.45]). They finally indicated substantial variation in CRT performance from participant to participant, thus reflecting individual differences in CRT performance, regardless of whether we considered the three new experiments (point estimate = 3.20, 95% CI = [2.77, 3.70]) or all four (point estimate = 3.24, 95% CI = [2.85, 3.67]).

To illustrate the extent of this variation in a more interpretable manner, we scale the point estimates of the variance components presented above by the point estimate of the meta-analytic average treatment effect. First, the difference in CRT performance from experiment to experiment was estimated to be 16.68 times larger than the meta-analytic average treatment effect when we considered the three new experiments or 4.10 times larger when we considered all four. Second, the difference in the treatment effect from experiment to experiment was estimated to be 6.16 times larger than the meta-analytic average treatment effect when we considered the three new experiments or 1.71 times larger when we considered all four. Third, the difference in CRT performance from participant to participant was estimated to be 46.21 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 that the larger relative estimates when we considered the three new experiments compared with all four do not so much reflect differences in the estimates of the variance components but instead 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 compared with 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]; Fig. 1 and Table S4 in the

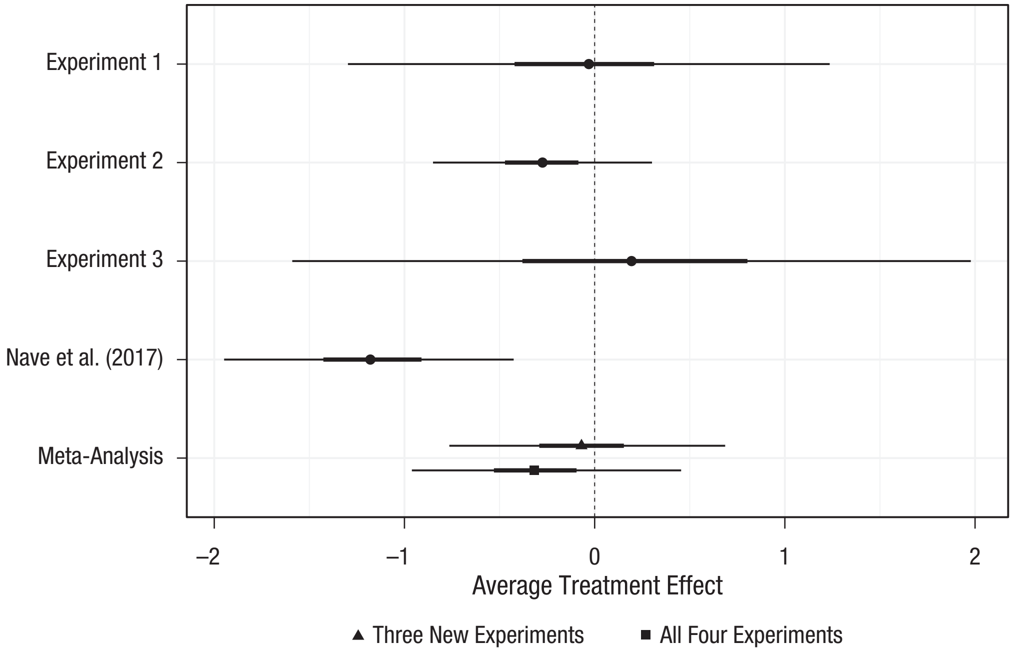

Fig. 1. Primary results: estimate of the average treatment effect in each of the three new experiments; the experiment of Nave, Nadler, Zava, and Camerer (2017); and the meta-analyses of the three new experiments as well as all four experiments. Point estimates are given by the points; 50% and 95% interval estimates are given by the thick and thin lines, respectively. Estimates are from models fitted to data from each experiment separately and from our primary model fitted to the data from all experiments jointly (see Tables S4 and S5 in the Supplemental Material available online).

Supplemental Material) or all four (point estimate = −0.32, 95% CI = [−0.96, 0.46]; Fig. 1 and Table S4 in the Supplemental Material). 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 with the treatment effect reported by Nave et al. within our modeling framework also resulted, again unsurprisingly, in an estimate with considerable uncertainty (point estimate = −1.01, 95% CI = [−2.44, 0.53]; see Table S6 in the Supplemental Material); although the point estimate suggests a stronger (i.e., more negative) treatment effect in Nave et al.'s experiment compared with 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. Results for the score of each participant (i.e., number of CRT items correct out of three) were in line with those presented above (see Table S7 in the Supplemental Material). Estimates of the variance components 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 the three new experiments (point estimate = −0.03, 95% CI = [−0.31, 0.28]) or all four (point estimate = −0.13, 95% CI = [−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 analyses included covariates included by Nave et al. (see Table S8 in the Supplemental Material). Results also remained substantively similar when the meta-analytic treatment effect was examined at the CRT item level (see Table S4 in the Supplemental Material). Finally, results remained substantively similar when we examined the treatment effect on intuitive but incorrect (as opposed to correct) responses both on average and at the item level (see Fig. S2 and Table S9 in the Supplemental Material).

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 (see Table S10 in the Supplemental Material); the result concerning the stability of the estimates of the variance components is particularly notable because it suggests that our conclusions regarding differences in the treatment effect from experiment to experiment were not driven by the single-blind or stressor conditions of the respective experiments. In addition, methodological-difference variables showed no substantial moderating effects (see Table S11 in the Supplemental Material).

Individual-difference variables. Individual-difference variables showed no substantial moderating effects (see Table S12 in the Supplemental Material), 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 potentially negative treatment effect at lower levels of trait impulsivity and a potentially positive treatment effect at higher levels of trait impulsivity (point estimate = 0.52, 95% CI = [0.06, 0.99]; Fig. S3 and Table S12 in the Supplemental Material).

Discussion

Nave et al. reported a single experiment and claimed that exogenous testosterone causes a decrease in CRT performance. We report three new experiments that also examined the effect of exogenous testosterone on CRT performance. When pooling the data across experiments, we found (a) substantial variation in CRT performance across experiments, treatment groups, and participants and (b) 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 that the notion of *the* effect of testosterone on CRT performance is not particularly meaningful. Instead, to the degree that testosterone does affect CRT performance, 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 Nave et al.'s experiment 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 potentially negative treatment effect at lower levels of trait impulsivity and a potentially positive treatment effect at higher levels of trait impulsivity. This 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 by Nave et al., 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 compared with 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 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 that we examined the moderating effects only of variables studied either by Nave et al. or in testosterone research more broadly; other variables may moderate the effect of testosterone on CRT performance. Nonetheless, insofar as future research continues to examine the effects of testosterone treatment on cognitive reflection—perhaps in search of such moderators—our results suggest the need for 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 (McShane et al., 2019).

Transparency

Action Editor: D. Stephen Lindsay
Editor: D. Stephen Lindsay
Author Contributions
E. L. Knight and B. B. McShane share first authorship of this article. For Experiment 1, E. L. Knight, U. Mayr, W. T. Harbaugh, and P. H. Mehta designed the experimental

protocol, with input specific to the Cognitive Reflection Test from C. Ma-Kellams; E. L. Knight, C. B. Christian, and P. J. Morales collected the data. For Experiment 2, J. M. Carré, T. L. Ortiz, and N. V. Watson designed the experimental protocol; T.L. Ortiz performed hormone assays; and T. L. Ortiz and K. Gilbert collected the data. For Experiment 3, H. H. Kutlikova, I. Riečanský, C. Eisenegger, and C. Lamm designed the experimental protocol; H. H. Kutlikova collected the data. E. L. Knight and B. B. McShane analyzed the data for the manuscript. E. L. Knight wrote the initial manuscript; E. L. Knight and B. B. McShane wrote the revised manuscripts. J. M. Carré, P. H. Mehta, and H. H. Kutlikova wrote segments specific to their respective experiments in the Method section of the Supplemental Material available online. All surviving authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

Experiment 1 was supported by National Science Foundation grants to P. H. Mehta (No. 1451848) and to W. T. Harbaugh and U. Mayr (No. 1063561). Experiment 2 was supported by funds from the Natural Sciences and Engineering Research Council of Canada to J. M. Carré (RGPIN-2014-06676), a Northern Ontario Heritage Fund Corporation grant to J. M. Carré, and a Discovery Grant (RGPIN-2016-05706) from the Natural Sciences and Engineering Research Council of Canada to N. V. Watson. Experiment 3 was supported by the Vienna Science and Technology Fund (WWTF VRG 13-007). E. L. Knight is partially supported by a National Institute on Aging Grant (No. AG049676) to The Pennsylvania State University.

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/kefqp/ and https://osf.io/b7ngf/, respectively. The design and analysis plans for the experiments were not preregistered. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619885607. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Erik L. Knight https://orcid.org/0000-0002-0349-542X
Blakeley B. McShane https://orcid.org/0000-0002-4839-266X
Ulrich Mayr https://orcid.org/0000-0002-7512-4556
Christine Ma-Kellams https://orcid.org/0000-0002-5468-5514

Acknowledgments

C. Eisenegger passed away on February 27, 2017. See the Supplemental Material available online for a full list of acknowledgements.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619885607

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*, 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*, 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*, 319–333.

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.

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

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

Gelman, A., Carlin, J. B., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Boston, MA: Chapman and Hall/CRC.

Geniole, S. N., Procyshyn, T. L., Marley, N., Ortiz, T. L., Bird, B. M., Marcellus, A. L., . . .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.

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), Article e42366. doi:10.1371/journal.pone.0042366

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

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*, 255–271.

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*(Suppl. 1), 99–105.

Nadler, A., Camerer, C. F., Zava, D. T., Ortiz, T. L., Watson, N. V., 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. doi:10.1098/rspb.2019.1062

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

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*, 768–774.

Ronay, R., & von Hippel, W. (2010). The presence of an attractive woman elevates testosterone and physical risk taking in young men. *Social Psychological and Personality Science*, *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*, 271–324.

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, USA*, *108*, 3448–3452.

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*, 757–768.