

# Decision

## **No Evidence for a Causal Effect of Exogenous Testosterone on Risky Decision-Making in Women: An Experiment and Meta-Analysis**

Lena Schaefer, Iris Iking, Isabel Woyke, Vivian Heuvelmans, Karin Roelofs, and Bernd Figner

Online First Publication, August 18, 2022. <http://dx.doi.org/10.1037/dec0000192>

### CITATION


Schaefer, L., Iking, I., Woyke, I., Heuvelmans, V., Roelofs, K., & Figner, B. (2022, August 18). No Evidence for a Causal Effect of Exogenous Testosterone on Risky Decision-Making in Women: An Experiment and Meta-Analysis. *Decision*. Advance online publication. <http://dx.doi.org/10.1037/dec0000192>

## BRIEF REPORT

## No Evidence for a Causal Effect of Exogenous Testosterone on Risky Decision-Making in Women: An Experiment and Meta-Analysis

Lena Schaefer<sup>1, 2</sup>, Iris Ikink<sup>1, 3, 4</sup>, Isabel Woyke<sup>1</sup>, Vivian Heuvelmans<sup>1, 4</sup>,  
Karin Roelofs<sup>1, 4</sup>, and Bernd Figner<sup>1, 4</sup><sup>1</sup> Behavioural Science Institute, Radboud University, Nijmegen<sup>2</sup> Department of Psychological and Brain Sciences, Boston University<sup>3</sup> Department of Experimental Psychology, Ghent University<sup>4</sup> Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen


While strong claims have been made that testosterone increases risk-taking, the existing literature is inconclusive. Thus, our experiment aimed at addressing some shortcomings of previous work. First, risk-taking was assessed using the Columbia Card Task, which allows to decompose overt risk-taking into three task factors—gain amount, loss amount, and the probability of losing—in both a dynamic, more affective (“hot”) and a static, more deliberative (“cold”) decision-making context. Second, we conducted a testosterone administration study in 80 females using a *triple-blind* (i.e., blinded participants, experimenters, and data-analysts), placebo-controlled, randomized, between-subjects design to investigate the causal effect of exogenous testosterone on risk-taking. Reviewers were also blind to the treatment conditions during the reviews. Third, we preregistered our analyses. We investigated (a) the main effect of testosterone, (b) the influence of gain amount, loss amount, and the probability of losing on risky decision-making, each in a more affective and more deliberative decision-making context, and (c) whether testosterone moderated any of those effects. Although we replicated previous


Lena Schaefer  <https://orcid.org/0000-0001-8607-5094>  
Isabel Woyke  <https://orcid.org/0000-0002-6263-942X>  
Lena Schaefer and Iris Ikink contributed equally to this article and share first authorship.

The authors would like to thank Aniek Wols for her help with the data collection, Mari-Liis Burket for her input on the preregistration, Gero Lange for his help on coordinating the blinding-procedure of this study, and Floor Burghoorn for her help with the Region of Practical Equivalence analysis. The authors have no conflict of interest to disclose.

Lena Schaefer played lead role in writing of original draft and equal role in formal analysis, visualization and writing of review and editing. Iris Ikink played lead role in project administration and equal role in data curation, formal analysis, visualization and writing of review and editing. Isabel Woyke played equal role in project administration. Vivian

Heuvelmans played equal role in project administration. Karin Roelofs played supporting role in writing of review and editing and equal role in conceptualization and supervision. Bernd Figner played lead role in supervision and equal role in conceptualization and writing of review and editing.

 The data are available at <https://osf.io/vejat/>

 The preregistered design and analysis plan is accessible at <https://osf.io/vejat/>

Correspondence concerning this article should be addressed to Lena Schaefer, Department of Psychological and Brain Sciences, Boston University, 677 Beacon Street, Boston, MA 02215, United States or Iris Ikink, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium or Bernd Figner, Behavioural Science Institute, Radboud University, Nijmegen, Houtlaan 4, 6525 XZ Nijmegen, Netherlands. Email: lenascha@bu.edu or iris.ikink@ugent.be or bernd.figner@ru.nl

studies showing that risk-taking was affected by gains, losses, probabilities, and decision-making context, we found no evidence for a main or interaction effect involving testosterone. This finding was further supported by our meta-analysis, which suggests that the effect of administered testosterone on risk-taking in women is smaller than a small effect size. In conclusion, these results provide no evidence for an effect of exogenous testosterone on decision-making under risk, raising some doubts about the commonly suggested direct link between the two.

*Keywords:* risk-taking, testosterone, Columbia Card Task, decision-making

*Supplemental materials:* <https://doi.org/10.1037/dec0000192.supp>

Events such as the financial crisis of 2008 raised the question whether heightened testosterone levels were related to exacerbated financial risk-taking, which in turn caused the collapse of the market (Coates & Herbert, 2008; Cueva et al., 2015; Nadler et al., 2018). Previous research, however, has yielded inconclusive results regarding the role of testosterone in risky decision-making: Although some *correlational* studies indeed showed that higher levels of endogenous testosterone (i.e., as naturally produced by the body) are related to increased risk-taking (Apicella et al., 2014, 2008; Coates & Herbert, 2008; Dariotis et al., 2016; Evans & Hampson, 2014; Stanton, Liening, & Schultheiss, 2011), others found this relationship to be nonlinear (Sapienza et al., 2009; Stanton, Mullette-Gillman et al., 2011), to be driven by males (Reavis & Overman, 2001; Schipper, 2012), to be moderated by cortisol (Mehta et al., 2015; Smith & Apicella, 2017), or to be absent (Derntl et al., 2014; Doi et al., 2015). Studies examining the *causal* effect of exogenous testosterone (i.e., by administering testosterone) on risk-taking report even more inconsistent results: While some reported increased risk-taking after testosterone administration (Cueva et al., 2015; van Honk et al., 2004), others found moderation by either a genetic variant (Wagels et al., 2017) or a prior loss outcome (Wu et al., 2016), or found no effects (Boksem et al., 2013; Nadler et al., 2021; Stanton et al., 2021; Woyke et al., under revision; Zethraeus et al., 2009).

Given these mixed findings, drawing strong conclusions on the influence of testosterone on decision-making under risk is difficult. Likely, some of these differences can be attributed to differences in study-designs (e.g., causal vs. correlational, within- vs. between-subjects), participant samples (e.g., gender and age distribution of the samples), risk-taking paradigms (e.g., loss/

gain/mixed lotteries, balloon analog risk task, Iowa Gambling Task), as well as social versus nonsocial task contexts (Heany et al., 2016). Moreover, Stanton (2017) suggested that the inconsistency in findings might be attributed to the potential existence of publication bias (but see Kurath & Mata, 2018, who found no support for this in correlational endogenous studies) as well as data-contingent analyses in the field of testosterone research, stressing the necessity of methodically strong and preregistered studies with unbiased designs and analysis approaches. In line with this idea, and in order to examine whether testosterone affects risk-taking behavior in a non-social context, we conducted a testosterone-administration study with a *triple-blind* (i.e., blinded participants, blinded experimenters, and blinded data-analysis using labels A/B for group instead of testosterone/placebo), placebo-controlled, randomized, between-subjects design, and we preregistered our analyses (<https://osf.io/vejat/>). To avoid any potential biases in the review process, the reviewers and authors remained blind to which group (labeled A and B) received placebo versus testosterone until the article was accepted for publication.

We assessed risky decision-making via the Columbia Card Task (CCT), which provides two advantages over other commonly used task paradigms: First, it allows to decompose overt risk-taking levels into three underlying psychological processes. In the decision sciences, agents faced with a risky choice are assumed to consider potential gains, losses, and the probabilities with which the gains and losses occur. Based on these three pieces of information (i.e., the three economic primitives), decision-makers are thought to make their choice on how much risk to take. In the CCT, the three economic primitives are varied orthogonally and systematically across game rounds to investigate how each of them

contributes to the decision on how many cards to turn over. The number of cards turned over in each game round serves as the measure of risk-taking as turning over more cards increases outcome variability (i.e., risk as defined in the decision sciences) and the probability of incurring a loss (i.e., risk as defined in everyday language; see Figner & Weber, 2011). Thus, this design feature of the CCT allows for the decomposition of overt risk-taking levels into three underlying psychological processes, namely, sensitivity to gains, losses, and probabilities.

Second, the CCT exists in a hot and a cold version: In the hot version, participants are assumed to make decisions based on gut feeling, affect, and arousal (Type 1 decision processes) compared to the cold CCT, where participants are assumed to make their decisions based on mathematical reasoning and deliberation (Type 2 decision processes; e.g., Buelow, 2015; Figner et al., 2009; Figner & Weber, 2011; Weller et al., 2019). The hypothesized differential involvement of affective versus deliberative processes in the two CCT versions has been confirmed using self-reports and the assessment of electrodermal activity (Figner et al., 2009): Both self-reported as well as physiological emotional arousal (as measured through skin conductance response) were higher in the hot compared to the cold CCT.

This latter distinction opens up the possibility to examine two contradicting claims about the working mechanisms via which testosterone is suggested to influence risk-taking: Some claim that testosterone influences intuitive decisions and unconsciously motivated behavioral responses via *affective* circuits in subcortical regions (Nave et al., 2017; van Honk et al., 2004), suggesting that testosterone moderates Type 1 decision processes. This would suggest that testosterone effects on risk-taking may be more pronounced in the hot compared to the cold CCT. A different claim in the literature states that testosterone influences risk-taking by increasing risk-neutrality (i.e., a change toward financially optimal risk-taking levels; Apicella et al., 2008, 2014; Heany et al., 2018; Sapienza et al., 2009), which would suggest that testosterone relates to Type 2 decision processes. In that case, we would expect decisions in the hot and the cold CCT to be more similar to each other and closer to risk-neutrality in the testosterone compared to the placebo group.

Given the inconclusive literature on the relation between testosterone and risk-taking, we decided

to preregister three (partly competing) hypotheses, each consistent with one of the contradictory claims in the literature: (a) The *increased risk-taking* hypothesis, (b) the *risk-neutrality* hypothesis, and (c) the *null-effect* hypothesis. After analyzing our data, we then evaluated which hypothesis was most consistent with our results. Thus, the main goal of this preregistration was to commit to a set of specific a priori analyses (not hypotheses), to avoid any possible bias that could result from (implicit or explicit) wishes to find specific effects or results. First, the increased risk-taking hypothesis predicts that the testosterone group makes more risky decisions compared to the placebo group (Cueva et al., 2015; van Honk et al., 2004). This effect could be driven by increased reward-sensitivity (van Honk et al., 2004), decreased punishment-sensitivity (Stanton, Liening, & Schultheiss, 2011), and/or decreased probability sensitivity (Cueva et al., 2015). Based on previous work suggesting that testosterone acts via an affective-motivational pathway (van Honk et al., 2004), we would expect a stronger testosterone effect in the hot than cold version of the CCT. Second, the risk-neutrality hypothesis predicts that testosterone increases risk-neutrality (Apicella et al., 2008, 2014; Heany et al., 2018; Sapienza et al., 2009). Although participants typically display risk-aversion in risk-taking paradigms (i.e., lower risk-taking levels than financially optimal), participants in the CCT usually display risk-seeking behavior (i.e., higher risk-taking levels than financially optimal; Figner et al., 2009). In this case, the risk-neutrality hypothesis thus predicts a decrease in risk-taking. Such a decrease might be explained by either increased or decreased reward-sensitivity, increased punishment-sensitivity, and/or increased probability sensitivity. Furthermore, we expect increased risk-neutrality to impact both the hot and cold CCT, such that the hot/cold difference might be smaller in the testosterone than placebo group. Finally, given (a) the inconsistent findings in the literature, (b) the potential existence of publication bias and data-contingent analyses in existing work (see also Stanton, 2017), and (c) our attempt at an unbiased design and analysis approach, we also deemed it possible that our study would reveal null effects of the testosterone administration. In this case, we would expect no main effect nor interactions involving testosterone.

To supplement our work and as suggested by a reviewer, we additionally conducted a systematic

literature search and a meta-analysis on the effect of administered testosterone on risk-taking in women. Given the differential effects of testosterone in men and women and across development, we focused on adult female-only samples. To further limit the scope of this analysis, we restricted our search to studies that include behavioral measures of risk-taking. We ultimately combined the results of six eligible studies using meta-analytical procedures to summarize the evidence for an effect of exogenous testosterone on risk-taking in women.

## Methods of Experiment

### Participants and Exclusion Criteria

Eighty women (age range 18–27 years;  $M = 21.37$ ,  $SD = 2.10$ ) were recruited via the participant-recruitment system of the Radboud University in Nijmegen, the Netherlands. Only females participated in the study, as the time course and dosage for inducing neurophysiological effects of the specific sublingual testosterone administration method we employed in this study has only been validated in females (Tuiten et al., 2000; van Rooij et al., 2012). Exclusion criteria (following Tuiten et al., 2000; van Rooij et al., 2012) are reported in Supplemental Materials (SM)-Appendix A. Eligibility of participants was determined based on self-report screening questionnaires.

Participant groups (placebo/testosterone; 40 each) did not differ significantly in age, income, digit span, positive and negative affect, or self-reported risk-taking across five domains (i.e., Domain-Specific Risk-Taking [DOSPERT]: ethical, financial, health/safety, recreational, social; see Blais and Weber (2006), and SM-Appendix B, for more information) as measured after drug intake but before the drug-active window 3.5 hr later, indicating comparable groups. The sample size was based on an a priori simulation-based power analysis, which indicated that given a sample size of 72 participants and a medium effect size (a Cohen's  $d$  of approximately 0.38), this study would achieve approximately 80% power (see SM-Appendix C and <https://osf.io/vejat/> for more details). Ultimately, 80 participants were tested to avoid a reduction in power in case of technical issues or participant exclusions. In addition to compensation with money or course credits, every

participant entered a lottery with 20% chance of receiving an additional payment that could range between €0 and €100 based on their decisions in one of all the decision-making tasks that were completed in the drug active window (see SM-Figure A, for an overview).

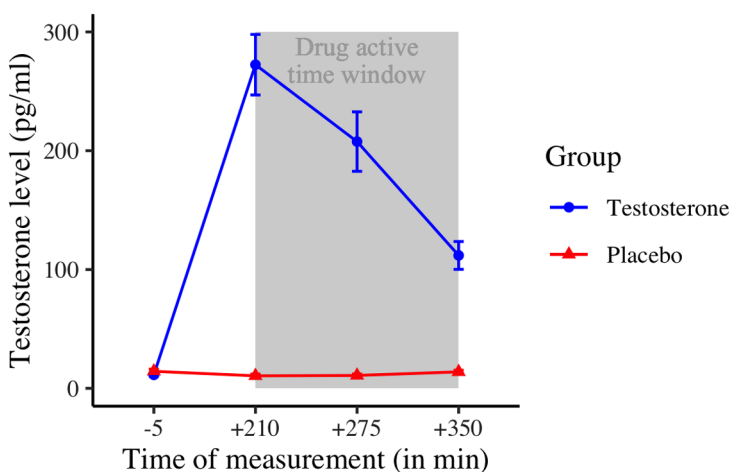
### Procedure

After signing the consent-form and being screened for exclusion criteria, participants received either a single dose of 0.5 mg testosterone suspended in a clear solution (0.5 ml with 0.5 mg hydroxypropyl- $\beta$ -cyclodextrin, 0.005 ml ethanol 96%, and distilled water) or a matched placebo using double-blind between-subjects randomization. Participants were instructed to hold the solution under their tongue for 1 min before swallowing it. This standardized testosterone administration procedure has shown to induce consistent behavioral and psychophysiological effects 3.5–6 hr after administration (Tuiten et al., 2000; van Rooij et al., 2012).

Accordingly, starting 3.5 hr after drug administration, participants completed two 1-hr task-blocks with several decision-making (see SM-Figure A, for an overview of the whole procedure). In addition to the CCT, participants conducted several other tasks, including a risk and ambiguity task (Tymula et al., 2012), probability, money, and time ratings, as well as an intertemporal choice task (Figner et al., 2010). These results will be reported elsewhere. These tasks were selected to study the effects of exogenous testosterone on economic, nonsocial decision-making; each of them targeting different aspects of decision-making.

Both versions of the CCT were completed in the first task-block (hot/cold order counterbalanced between-subjects), separated by two other tasks. At the end of the experiment, participants entered the lottery for additional performance-based payouts and received compensation for their participation. Over the course of the experiment, participants also provided a total of five saliva samples (passive drool method), which confirmed a successful testosterone administration procedure (see Figure 1). The whole experiment had a duration of 6.5 hr, started at 10 am for each participant, and was approved by an accredited local Research Ethics Committee (Commissie Mensgebonden Onderzoek Regio Nijmegen-Arnhem, protocol ID: NL49277.091.14).

**Figure 1**  
*Testosterone Levels (in pg/ml) per Group During the Time Course of the Experiment, as Measured Using Saliva Samples*



*Note.* CCT = Columbia Card Task. The first measurement was taken 5 min prior to testosterone administration (−5 min); the last almost 6 hr later (+350 min), at the end of the experiment. The CCT was administered during the drug active window (starting from +210 min). We additionally collected a saliva sample 1 hr and 30 min after testosterone administration, but since this returned only missing values in one group, we do not show this time point in the plot. Although we are not sure what may have caused this, the missing time point is not essential in evaluating whether the testosterone administration was effective. To avoid unblinding any of the authors until the article was accepted for publication, an otherwise noninvolved party sent us the data after changing all participant IDs, such that even if we tried we would not be able to link these samples to the CCT data. See the online article for the color version of this figure.

### **Blinding Procedure**

This study employed a *triple-blind* design, meaning that participants, experimenters, and data-analysts were blind with respect to which participants received placebo versus testosterone (using labels A/B for group instead of testosterone/placebo). The same labels were used during the review process. The study was unblinded only after acceptance of the article on June 27, 2022.

### **Materials**

#### **Columbia Card Task**

We used versions of the hot and the cold CCT with 24 game rounds each (e.g., Figner & Weber, 2011). At the beginning of each game round, 32 cards were presented face down on the computer screen in four rows of eight cards. Out of the 32 cards, one or three were loss cards, the rest were gain cards. The number of loss cards thus varied across game rounds (representing the task factor

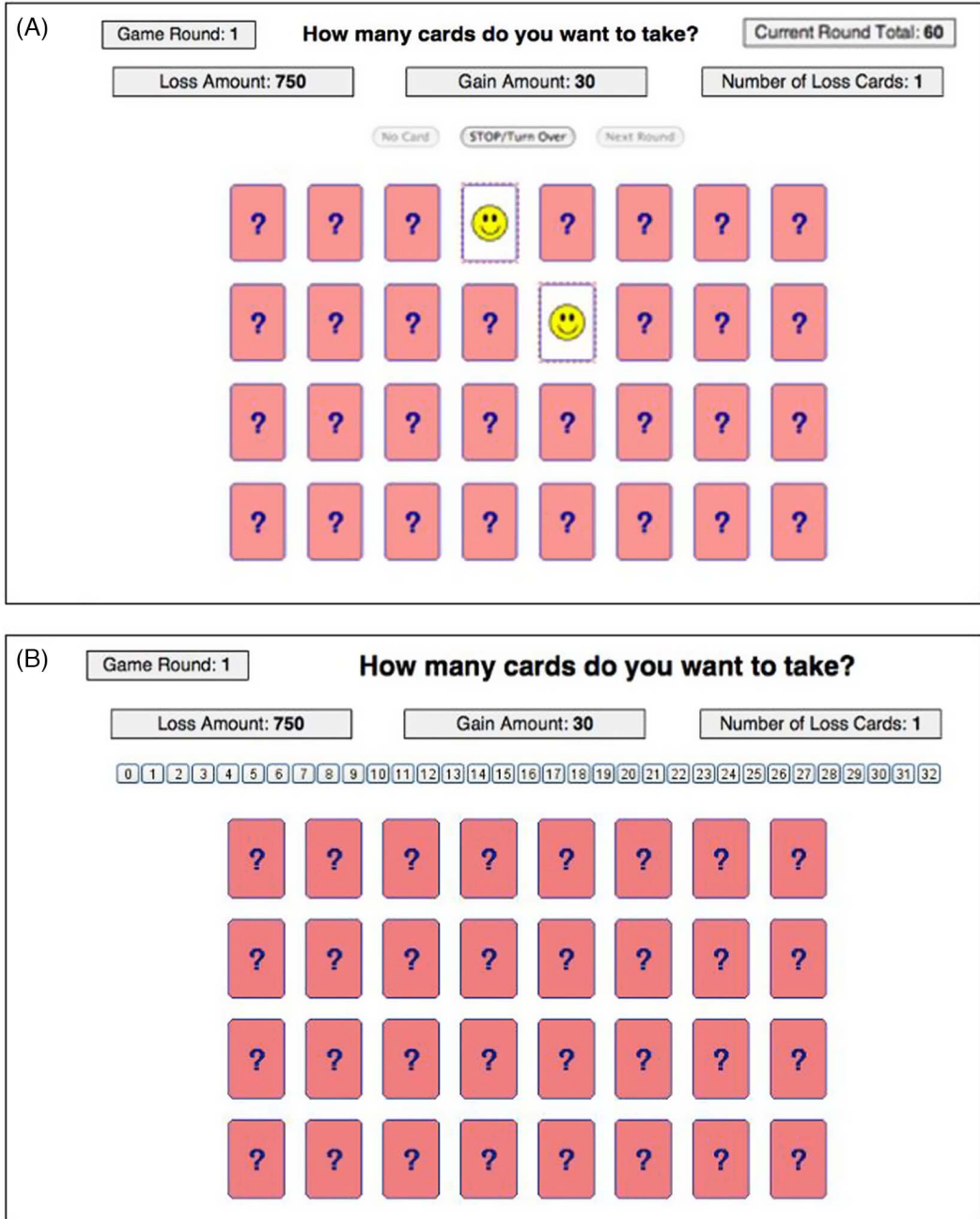
*loss probability*). By turning over a gain card, a specified number of points was added to the total score (i.e., *gain amount*: 10 or 30 points), and by turning over a loss card, a specified number of points was subtracted from the total score (i.e., *loss amount*: −250 or −750 points) and the game round ended. At the top of the screen, information about gain amount, loss amount, and the number of loss cards in the current game round was presented; each game round started with a score of 0 points. Number of loss cards (1/3), loss amounts (−250/−750), and gain amounts (10/30) were full factorially crossed to create eight different combinations that were repeated three times (in blocks unnoticeable to participants), resulting in 24 game rounds per CCT version. The primary dependent variable was the number of cards turned over per game round, with more cards indicating more risk-taking.

In the hot CCT, participants could select any card by clicking on it, which would then turn over, revealing whether it was a gain or loss card (Figure 2A). If it was a gain card, the gain amount



**Figure 2**

Examples of (A) a Hot and (B) Cold CCT Trial



*Note.* CCT = Columbia Card Task. Information about the current game round, loss amount (per loss card), gain amount (per gain card), and number of loss cards was always provided on top of the screen. (A) In the hot CCT, currently two gain cards (out of 32) are turned over. Participants could continue turning over cards by selecting additional cards, with direct feedback being provided in terms of added/lost points (visible at "Current Round Total"). A game round ended when the participant turned over a loss card or pressed the STOP button. (B) In the cold CCT, participants only had to indicate how many cards they want to turn over (ranging from 0 to 32) and no feedback was provided until all game rounds were played. See the online article for the color version of this figure.

was added to the total score, which was constantly visible and changed with every card turned over. Participants could then continue to turn over cards, or end the current game round by pressing the STOP button at the top of the screen. If a loss card was turned over, however, the loss amount was subtracted from the total score, and the game round ended.<sup>1</sup> When a game round ended, all cards were always turned over, revealing which of the remaining cards were gain and loss cards.

The cold CCT (Figure 2B) was very similar to the hot CCT, except here at the top of the screen, a sequence of buttons labeled from 0 to 32 was presented. In the beginning of each game round, participants needed to indicate the number of cards they wanted to turn over by clicking one of these buttons. Participants also received no outcome feedback until all game rounds were finished.

Importantly, in the hot CCT, some game rounds ended when the participant turned over a loss card. In these right-censored game rounds, we do not know whether participants would have turned over more cards if they had not encountered the loss card. In contrast, in the cold CCT participants were always able to indicate how many cards they wanted to turn over, without any censoring. This difference in censoring was accounted for in our analyses (see below).

### Self-Report

Using short self-report questionnaires after each CCT version, the constructs Type 1 decision processes (i.e., intuitive, gut-based decision-making; 5 items), and Type 2 decision processes (i.e., deliberative, mathematical decision-making; 5 items) were assessed, based on and adapted from Figner et al. (2009). Responses were given on continuous visual analog scales ranging from 0 (*does not apply at all*) to 100 (*strongly applies*), and internal consistency of the constructs was .77 and .82, respectively. For more details, see SM-Appendix D.

### Data Analysis

Data and scripts are available at (<https://osf.io/vejat/>). All analyses were conducted using mixed-effects models in a Bayesian framework, calculating credible intervals (CIs) using the `brm`-function of the R-package `brms` (Bürkner, 2017), which provides an interface to Stan (Carpenter et al., 2017). We used `brms`' weakly informative

default priors and fit the models using six chains with 8,000 iterations each (4,000 warmup). Model convergence was inspected by checking the Rhats (Rhat should be  $<1.01$  and  $>0.99$ ) and visually inspecting the trace- and densityplots of all parameters. If the 95% CI of an effect did not include 0, we concluded that there was a significant effect. To account for the repeated-measures nature of the data and to avoid inflated Type I errors, we used a maximal random-effects structure in all models as recommended by Barr et al. (2013). Thus, all models included a random intercept per participant, random slopes for all within-subjects effects, and all possible random correlations.

The CCT data were analyzed at the game-round level without aggregation. The number of cards turned over per game round (range: 0–32) was analyzed as a function of group (testosterone/placebo), CCT version (hot/cold), gain amount (10/30 points), loss amount (–250/–750 points), and number of loss cards (1/3 loss cards). We also included all possible two-way interactions between group and the other predictors, and three-way interactions between group, CCT version, and either number of loss cards, loss amount, or gain amount. All predictors were sum-to-zero contrast coded, and group was the only between-subjects predictor. To account for the right-censoring of our data, we used the addition term `resp_cens()` as implemented in `brms` (Bürkner, 2017). When using this function, censored observations are integrated out (see Section 4.3 in Stan User's Guide, for more details; Stan Development Team, 2021). Furthermore, since the data are censored, we report estimated marginal means (EMMs) and 95% CIs using the `ggeffects` package (Lüdtke, 2018) instead of the raw data.

To specifically test whether testosterone affects risk neutrality, we used the exact same predictors as described above, but now with deviation scores as dependent variable. To obtain the deviation scores, we first computed the number of cards that maximized the expected value (EV) for each game round. In short (for details, see Figner et al., 2009), the normative solution maximizing EV says that no further card should be turned over

<sup>1</sup> Note that with each card turned over, both the probability of turning over a loss card and the outcome variability increases (the latter one increases up to a point, after which it decreases again).



if the EV of turning over an additional card is smaller than 0, or formally, when fewer than  $n_{\text{cards}}$  remain to be turned over, calculated by:

$$n_{\text{cards}} = \frac{n_{\text{loss cards}}(g + l)}{g}, \quad (1)$$

where  $n_{\text{loss cards}}$  is the number of loss cards,  $g$  is the gain amount and  $l$  is the loss amount. For example, with one loss card, a loss amount of 250, and a gain amount of 30,  $n_{\text{cards}}$  is 9.33, indicating that the participant should turn over  $(32 - 9.33) = 22.67$  cards to maximize EV. We then computed the deviation score by subtracting the *actual* from this *optimal* number of cards turned over, and multiplied that by  $-1$ . Thus, a positive deviation score represented risk-seeking (i.e., more cards than optimal) and a negative deviation score risk-aversion (i.e., less cards than optimal).

Finally, to test for the effects of CCT version and group on Type 1 and Type 2 decision processes, these dependent variables were separately analyzed as a function of group, CCT version, and their two-way interaction. These models only included a random intercept per participant.

## Results of Experiment

### Testosterone Effects

Group (placebo/testosterone; originally analyzed using the labels A/B) did not significantly influence the number of cards turned over, placebo: EMM = 8.75, 95% CI [7.67, 9.84]; testosterone: EMM = 9.80, 95% CI [8.74, 10.87], estimated regression coefficient:  $B = -0.53$ ; 95% CI [-1.29, 0.25]. Figure 3 illustrates that the two groups are highly similar with regard to their central tendencies and distributions. Furthermore, all interaction effects involving group were nonsignificant (Table 1). Interestingly, visual inspection of the CIs suggests that specifically the main effect of group was estimated with less precision than other significant and nonsignificant effects (i.e., the 95% CIs are wider for the group main effect compared to the other effects; Figure 4).<sup>2,3</sup>

To specifically test whether testosterone affected risk neutrality, we used the same fixed- and random-effects structure as in the model above but with a different dependent variable, namely, deviation scores. As expected, participants were generally risk-seeking in the CCT, EMM = 4.77, 95% CI [3.97, 5.56]. Group did not

influence the deviation scores, placebo: EMM = 4.28, 95% CI [3.11, 5.37], testosterone: EMM = 5.26, 95% CI [4.16, 6.39], suggesting that neither group was closer to risk-neutrality than the other. All interactions effects involving group were nonsignificant (Table 2).

### Task Effects

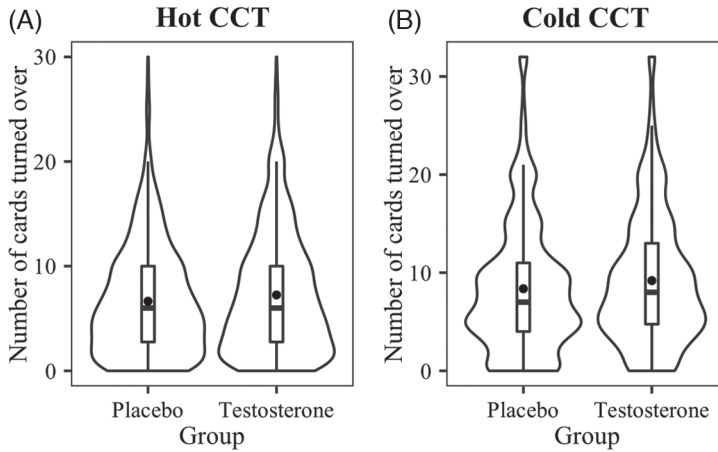
As expected, all three task factors were significant (Table 1). Participants selected more cards (a) when the probability of losing was low compared to high, 1 loss card: EMM = 11.85, 95% CI [10.95, 12.74]; 3 loss cards: EMM = 6.71, 95% CI [6.02, 7.38], (b) when the loss amount was low compared to high, 250-point loss: EMM = 11.39, 95% CI [10.53, 12.25]; 750-point loss: EMM = 7.15, 95% CI [6.38, 7.96], and (c) when the gain amount was high compared to low, 30-point gain: EMM = 10.24, 95% CI [9.47, 11.03]; 10-point gain: EMM = 8.31, 95% CI [7.52, 9.12]. These findings indicate that participants were sensitive to changes in the gain amount, the loss amount, and the probability of losing, and adjusted their level of risk-taking accordingly.

Furthermore, participants selected significantly more cards in the hot than the cold CCT, hot: EMM = 9.77, 95% CI [8.92, 10.64]; cold: EMM = 8.78, 95% CI [7.95, 9.65]. In addition, there was a significant interaction between CCT version and loss amount, indicating that the change from a small to a large loss amount reduced the number of turned over cards more in the cold than the hot CCT (i.e., a 41.80% vs.

<sup>2</sup> As suggested by a reviewer, we reran this model using the mean DOSPERT score across the five scales as a main effect to control for possible baseline differences in risk-taking. We found the same pattern of significance, confirming that group did not significantly influence the number of cards turned over, even after controlling for baseline differences in risk-taking, placebo: EMM = 8.71, 95% CI [7.60, 9.83]; testosterone: EMM = 9.80, 95% CI [8.66, 10.90], estimated regression coefficient:  $B = -0.52$ ; 95% CI [-1.29, 0.25]; see SM-Table C.

<sup>3</sup> As suggested by a reviewer, we also conducted a Bayesian ROPE (Region of Practical Equivalence) analysis to quantify the support for a null effect. In summary (see details in SM-Appendix E): We can conclude with high credibility that the true effect is smaller than a medium standardized effect size ( $|Cohen's d| = |Hedge's g| < 0.46$ ). Given our sample size, the correction factor for computing Hedge's  $g$  has only a minor influence, such that Hedge's  $g$  rounds to the same value as Cohen's  $d$ . The conclusion of a null effect is only supported with low credibility, suggesting that our study was underpowered to find small effects.

**Figure 3**  
 Combined Violin- and Boxplots Based on the Raw Data, Showing the Number of Cards Turned Over in the (A) Hot and (B) Cold CCT Version per Game Round, as a Function of Group (Placebo/Testosterone)



*Note.* CCT = Columbia Card Task. The black dots indicate the mean number of cards turned over in the placebo and the testosterone group. As can be seen, central tendencies and distributions are very similar in both groups.

32.76% reduction, respectively). Thus, participants seemed less sensitive to losses in the hot than the cold CCT, in line with the notion that affect can decrease attention to choice-relevant information (e.g., Pachur et al., 2014 found this for probabilities).

### Self-Report

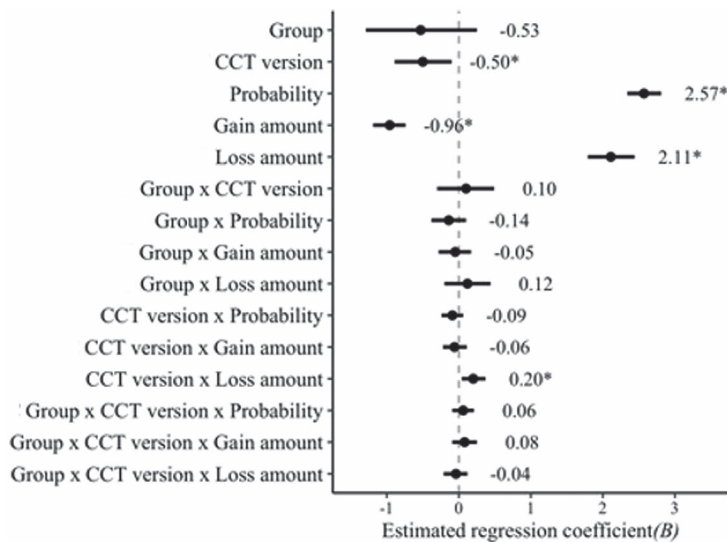
As expected (Figner et al., 2009), participants reported greater reliance on Type 1 decision processes (e.g., decisions based on excitement and gut feelings) in the hot compared to the cold version of

**Table 1**  
 Results of Choice Model With Number of Cards Turned Over per Game Round as Dependent Variable

Predictor	<i>B</i>	Est. error	Lower 95% CI	Upper 95% CI	Sign.
Intercept	9.28	0.39	8.51	10.05	s
Group (placebo/testosterone)	-0.53	0.39	-1.29	0.25	ns
CCT version (hot/cold)	-0.50	0.20	-0.89	-0.10	s
Probability of losing (3 or 1 loss cards)	2.57	0.12	2.34	2.81	s
Gain amount (30 or 10 points)	-0.96	0.12	-1.19	-0.74	s
Loss amount (-750 or -250 points)	2.11	0.16	1.79	2.44	s
Group × CCT version	0.10	0.20	-0.30	0.49	ns
Group × Probability	-0.14	0.12	-0.38	0.10	ns
Group × Gain amount	-0.05	0.12	-0.28	0.17	ns
Group × Loss amount	0.12	0.16	-0.20	0.44	ns
CCT version × Probability of losing	-0.09	0.08	-0.24	0.06	ns
CCT version × Gain amount	-0.06	0.09	-0.22	0.11	ns
CCT version × Loss amount	0.20	0.08	0.04	0.37	s
CCT version × Group × Probability of losing	0.06	0.08	-0.09	0.21	ns
CCT version × Group × Gain amount	0.08	0.09	-0.09	0.25	ns
CCT version × Group × Loss amount	-0.04	0.08	-0.21	0.12	ns

*Note.* *B* = estimated regression coefficient; Est. Error = estimated standard error; lower 95% CI = lower boundary of the 95% posterior credible interval; upper 95% CI = upper boundary of the 95% posterior credible interval; Sign = significance of effect. If the 95% CI does not include 0, we interpret the effect as significant, with s = significant; ns = nonsignificant; CCT = Columbia Card Task.

**Figure 4**  
*Unstandardized Estimated Regression Coefficients (B) With 95% Posterior Credible Intervals (CIs) for the Fixed Effects of the Main Model*



*Note.* CCT = Columbia Card Task; CI = credible interval. Since all categorical predictors are sum-to-zero coded, the magnitude of  $B$  can be compared. The number of cards turned over was the dependent variable in this model, with \* indicating that the 95% CI did not include 0, in which case we rejected the null hypothesis for that effect and deemed the effect significant. As can be seen, specifically the group effect has wider CIs compared to the other effects, indicating that this effect was estimated with less precision.

the CCT, hot: EMM = 64.80; 95% CI [61.64, 67.77]; cold: EMM = 55.87, 95% CI [52.87, 58.99], and, vice versa, they reported greater reliance on Type 2 decision processes (e.g., decisions based on deliberative and mathematical strategies) in the cold compared to the hot CCT, cold: EMM = 51.86, 95% CI [48.16, 55.36]; hot: EMM = 43.71, 95% CI [40.14, 47.31]. Group did not influence Type 1 or Type 2 decision processes nor were any interactions involving group significant. For full results of these models, see Table 3.

### Meta-Analysis

#### Literature Search

To identify potential work for inclusion in our meta-analysis, we employed multiple search strategies. First, we searched Google Scholar using the keyword testosterone in combination with keywords related to sex (women, female) and risk (risk, risky, risky choice, risk-taking,

reward). See SM-Appendix F, for all used keywords. We also screened the reference lists of review articles on the relation between testosterone and risk-taking (Apicella et al., 2015; Kurath & Mata, 2018; Stanton, 2017) as well as of two recent empirical studies (Nadler et al., 2021; Stanton et al., 2021).

#### Inclusion Criteria

To be included in the meta-analysis, records had to fulfill the following criteria:

1. Studies include an adult female participant sample: Participants were all female, and the mean age of the sample was above 18 years. If no information on the mean age was provided, we used the midpoint of the age range.
2. Studies contained a risky decision-making task: Tasks were required to measure risky decision-making, operationalized as

**Table 2***Results of Choice Model With Deviation Scores per Game Round as Dependent Variable*

Predictor	<i>B</i>	Est. error	Lower 95% CI	Upper 95% CI	Sign.
Intercept	4.77	0.41	3.97	5.56	s
Group (placebo/testosterone)	-0.50	0.40	-1.29	0.30	ns
CCT version (hot/cold)	-0.83	0.20	-1.24	-0.44	s
Probability of losing (3 or 1 loss cards)	-1.12	0.13	-1.36	-0.88	s
Gain amount (30 or 10 points)	2.25	0.12	2.02	2.50	s
Loss amount (-750 or -250 points)	-1.07	0.17	-1.41	-0.73	s
Group × CCT version	0.08	0.20	-0.31	0.48	ns
Group × Probability	-0.12	0.13	-0.37	0.13	ns
Group × Gain amount	-0.07	0.12	-0.31	0.17	ns
Group × Loss amount	0.14	0.17	-0.19	0.48	ns
CCT version × Probability of losing	-0.24	0.10	-0.43	-0.04	s
CCT version × Gain amount	0.06	0.10	-0.13	0.25	ns
CCT version × Loss amount	0.05	0.10	-0.13	0.24	ns
CCT version × Group × Probability of losing	0.04	0.10	-0.15	0.23	ns
CCT version × Group × Gain amount	0.09	0.10	-0.11	0.28	ns
CCT version × Group × Loss amount	-0.06	0.10	-0.25	0.13	ns

*Note.* *B* = estimated regression coefficient; Est. Error = estimated standard error; lower 95% CI = lower boundary of the 95% posterior credible interval; upper 95% CI = upper boundary of the 95% posterior credible interval; Sign = significance of effect. If the 95% CI does not include 0, we interpret the effect as significant, with s = significant; ns = nonsignificant; CCT = Columbia Card Task.

choosing the option with higher outcome variability compared to other options (Figner & Weber, 2011).

3. Studies implemented a testosterone administration manipulation.
4. Studies contained (or authors provided) sufficient statistics to calculate the effect size for the difference in risk taking between the testosterone and the placebo group.
5. Studies were written in English.

### Screening and Selection Procedures

Figure 5 presents the Preferred Reporting Items for Systematic Reviews Meta-Analyses (Moher et al., 2009) diagram depicting the selection and exclusion process. In short, we identified 455 studies through the Google Scholar search and 670 records by manually inspecting reference lists. Studies that clearly did not meet the inclusion criteria (based on reading the title and abstract) were excluded at this stage ( $N = 893$ ). We screened the remaining studies ( $N = 232$ ) in more detail for the inclusion criteria. At this stage, 227 records were excluded: 103 studies were excluded as they were duplicates, 123 did not pass the inclusion criteria, and one record was excluded as the results were still blinded.<sup>4</sup> In addition to the current project, we thus identified

a total of five studies, resulting in a sample of six studies to be included in the analysis.

### Coding of Study Characteristics

Two of the authors independently evaluated the suitability of studies for inclusion and coded the study characteristics. Interrater agreement between coders was high (interrater reliability = .94) and all discrepancies were resolved.

We coded the study's sample size (ranging from 12 to 134) and mean age of the sample (ranging from 21.37 to 62.5). With respect to the risky decision-making task, we coded the name of the task, whether task performance was incentive-compatible, whether choice probabilities were known or unknown, whether immediate performance feedback was provided, and whether a safe option (i.e., no outcome variability associated with one of the available options) was available. In addition to this, we also determined the type of testosterone administration procedure as well as the study design.

<sup>4</sup> This study comes from our group and uses the same participant sample but a different task, which assesses decision-making under ambiguity and under risk. To date, we have not yet unblinded the conditions, that is, we do not yet know which group received testosterone and which placebo (Woyke et al., under revision).

**Table 3**

Results of the Self-Report Questionnaire Data, With Type 1 or Type 2 Decision Processes as a Dependent Variable

Dependent variable	Predictor	<i>B</i>	Est. error	Lower 95% CI	Upper 95% CI	Sign.
Type 1 decision processes	Intercept	60.32	1.35	57.65	62.95	s
	CCT version (hot/cold)	-4.47	0.81	-6.04	-2.88	s
	Group (placebo/testosterone)	-1.36	1.37	-4.05	1.31	ns
	Group × CCT version	1.17	0.81	-0.40	2.74	ns
Type 2 decision processes	Intercept	47.77	1.57	44.65	50.82	s
	CCT version (hot/cold)	4.07	0.98	2.11	6.01	s
	Group (placebo/testosterone)	-0.13	1.58	-3.22	2.95	ns
	Group × CCT version	-0.38	0.99	-2.32	1.56	ns

*Note.* *B* = estimated regression coefficient; Est. Error = estimated standard error; lower 95% CI = lower boundary of the 95% posterior credible interval; upper 95% CI = upper boundary of the 95% posterior credible interval; Sign = significance of effect. If the 95% CI does not include 0, we interpret the effect as significant, with s = significant; ns = nonsignificant; CCT = Columbia Card Task.

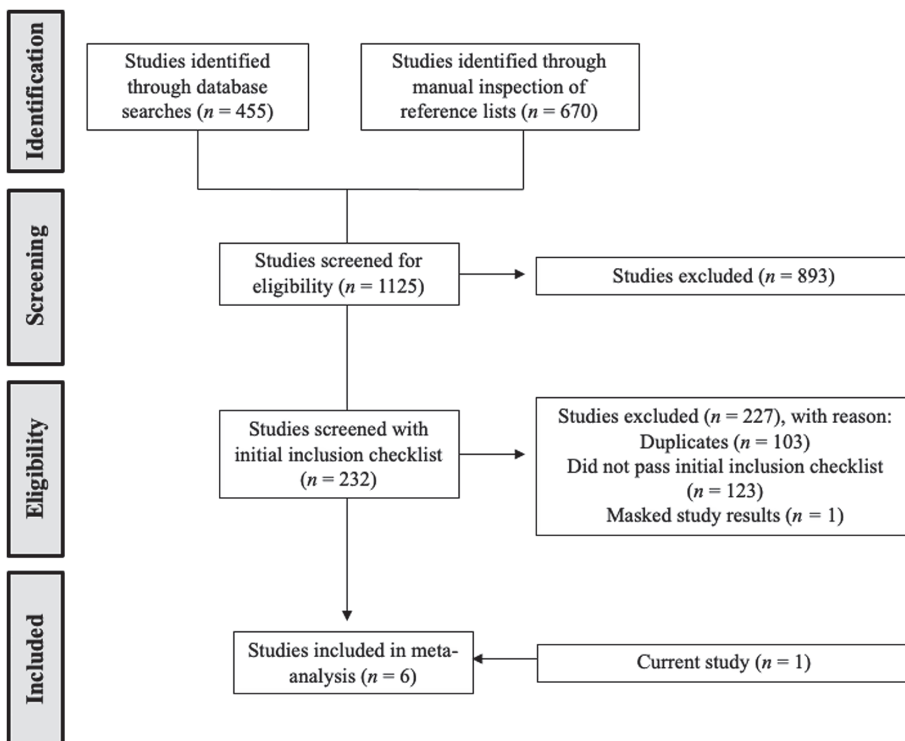
### Calculation of Effect Sizes

Effect size and variances were calculated to reflect the difference between risky decision-

making in the testosterone and the placebo group. They were coded so that a positive effect reflects more risk-taking in the testosterone group compared to the placebo group. We first computed

**Figure 5**

Preferred Reporting Items for Systematic Reviews Meta-Analyses (PRISMA) Diagram Depicting the Record Selection and Exclusion Process





Cohen's  $d$  and transformed it into Hedge's  $g$  which corrects for inflated effect sizes that may occur when the sample size is small (Formulas 4.22–4.24 in Borenstein et al., 2009). Hedge's  $g$  is interpreted in the same way as Cohen's  $d$ . Please note that the correction is very small for even moderate sample sizes: for example, for our own study with  $N = 80$ , the correction factor is about 0.99, that is, Hedge's  $g$  is only about 1% smaller than Cohen's  $d$ . See SM-Appendix F, for more details on the effect size calculation and comparability of effect sizes.

### Data-Analysis Plan

We estimated the average effect size for the effect of administered testosterone on risky decision-making in women using a random-effects model in metafor (Viechtbauer, 2010). We additionally conducted equivalence tests in the context of quantifying support for the null hypothesis. We compared the meta-analytic effect size against (a) a small effect size ( $|g| > 0.2$ ; as defined by Cohen, 1988) and (b) half of a small effect size ( $|g| > 0.1$ ) as suggested by Kruschke (2013, 2018) as default boundaries for defining a region of practical equivalence (ROPE) around a null effect.

### Results of Meta-Analysis

For a complete overview of the characteristics of the included studies, see SM-Table D. In short, five out of six studies included a sample with the participants' mean age between 21 and 23 years and administered a single-dose of 0.5 mg sublingual testosterone. The remaining study recruited participants between 60 and 65 years of age and used testosterone undecanoate 40 mg daily for 4 weeks. Different behavioral measures were employed, ranging from the Iowa Gambling Task and the CCT (the present study) to different variants of gambling tasks. While five out of six studies provided performance-based incentives, only four out of six studies included known task probabilities. Three tasks as well as the hot version of the CCT provided immediate feedback but the two remaining tasks as well as the cold version of the CCT did not provide immediate feedback. Half of the studies included a safe option and four out of the six included studies employed a between-subjects design. Sample sizes ranged from 12 to 134 participants.

The random-effects model indicated that the average effect size is small and not significant, Hedge's  $g = 0.01$ , 95% CI  $[-0.16, 0.17]$ ,  $z = 0.08$ ,  $p > .05$ . Equivalence tests suggested that we can reject effects larger than small effects ( $|g| = 0.2$ ,  $z = -2.28$ ,  $p = .011$ ) but we cannot reject effects that are larger than half of a small effect ( $|g| = 0.1$ ,  $z = -1.10$ ,  $p = .139$ ). Thus, our meta-analysis suggests that the effect of administered testosterone on risk-taking in women is smaller than a small effect size.

### Discussion

The present study investigated the causal effect of exogenous testosterone on decision-making under risk in women. We preregistered our analyses, employed a triple-blind design with initial reviews using the labels A/B for group instead of testosterone/placebo, and assessed both overt risk-taking levels as well as the relative contribution of the underlying psychological processes (sensitivity to gains, losses, and probabilities) in a more affective and a more deliberative decision-making context. While we replicated the results of previous work with the CCT, showing that the three task factors (i.e., gain amount, loss amount, and probability of losing) as well as decision-making context (affective vs. deliberative) affected risk-taking (Buelow, 2015; Figner et al., 2009; Weller et al., 2019), we did not find a significant effect of testosterone on risk-taking, neither as a main nor as an interaction effect.

Our results thus provide no support for the increased risk-taking or for the risk-neutrality hypothesis. Based on our ROPE analysis, we can conclude with high credibility that the true effect of administered testosterone on risk-taking in women is smaller than a medium, standardized effect size ( $|Cohen's\ d| = |Hedge's\ g| < 0.46$ ). The results of the meta-analysis are consistent with the notion of a very small, if any, effect (smaller than a small, standardized effect size;  $|Hedge's\ g| < 0.2$ ). When interpreting the empirical results of our experiment, it is important to consider the question whether the results reflect the true absence of an effect or whether they might be due to other causes. We believe that the results of our experiment align most with the idea that testosterone either has no or only a small effect on risky choice in women: First, we replicated all typical task effects, suggesting that the absence of a group effect cannot be attributed to unreliable

task measurements or to participants misunderstanding the task or not paying attention. Second, groups (i.e., testosterone vs. placebo) were comparable in terms of age, income, digit span, positive and negative affect as well as self-reported risk-taking before the drug-active window, thus excluding the possibility that such preexisting differences influenced our results. Third, the central tendencies and distributions per CCT version were highly similar in both groups, further strengthening the argument that a potential causal effect of testosterone on risk-taking would be likely rather small (see Figure 3). Fourth and perhaps most importantly, these empirical results are in line with the results of our meta-analysis, which suggests that the true effect of administered testosterone on risk-taking in women is most likely very small (i.e., smaller than a small, standardized effect size), if it exists at all.

These results may raise doubts about the efficacy of commonly used testosterone administration procedures, a topic critically discussed in the literature (see, e.g., Nadler et al., 2019). Note that the procedure used in the present study has been validated (Tuiten et al., 2000; van Rooij et al., 2012) and has shown behavioral effects in several studies using a similar timeline and comparable or smaller samples (see, e.g., Boksem et al., 2013; Enter et al., 2014; Hermans et al., 2010; Hutschemaekers et al., 2021; Mehta et al., 2015; Terburg et al., 2012; van Honk et al., 2004, 2012, 2016). Furthermore, our manipulation check confirmed increased testosterone levels in the saliva of the testosterone group compared to the placebo group when the CCT was administered (see Figure 1<sup>5</sup>). For these reasons, we deem it unlikely that the observed absence of a testosterone effect in our empirical study can be explained by an unsuccessful testosterone administration.

Note, however, that it is unclear to what extent testosterone levels in the saliva (as measured in our manipulation check) reflect testosterone levels in the brain. Generally, this field of pharmacological research would benefit from more basic science and larger validation studies: To date, there is no consensus as to when or how much testosterone reaches the brain for any of the typically used administration methods (i.e., sublingual, intranasal, or gel). Consequently, there is little guidance on the optimal dosage and time-course for finding reliable behavioral effects, nor

is it clear to what extent different administration methods may yield comparable results.

Although several testosterone administration studies already reported a nonsignificant main effect on risk-taking in women (Boksem et al., 2013; Buskens et al., 2016; Wu et al., 2016; Zethraeus et al., 2009), we believe that our independent replication of this null effect makes a valuable contribution to the literature. Specifically, we conceptually replicate the null effect using a different risk-taking task that is more comprehensive and arguably more naturalistic (Figner et al., 2009) than the simpler risky gambles used in other studies (Boksem et al., 2013; Bürkner, 2016; Wu et al., 2016; Zethraeus et al., 2009). In addition, although most (if not all) behavioral measures of risk-taking may have limited predictive validity (see, e.g., Frey et al., 2017, for a large study showing poor predictive validity across a large set of behavioral risk tasks), the CCT has demonstrated at least a somewhat higher temporal stability relative to other behavioral measures of risk-taking (i.e., a 6-month test-retest reliability of about .6 compared to the average of about .4). Last, given the inconclusive literature and potential existence of publication bias and data-contingent analyses in the field of testosterone research (see Stanton, 2017), conceptually replicating work using unbiased designs and analysis approaches (i.e., a triple-blind, placebo-controlled randomized design and preregistered analyses) is valuable.

We think it is worth to further compare and embed our empirical study in the context of existing studies: Three out of the four existing studies reporting a nonsignificant effect of testosterone on risk-taking in women had a sample that was similar to or smaller than our sample. One study—Zethraeus et al. (2009)—used a larger sample size with a between-subjects design with  $n = 67$  participants in the placebo group and  $n = 67$  in the testosterone group. While compared to this study, our own study used a more comprehensive and arguably more reliable measure of risk-taking, there are two additional key differences between both studies so that replicating the null effect from Zethraeus et al.

<sup>5</sup> While one could argue that the increased testosterone levels might stem from contamination—a possibility we ultimately cannot rule out—it at least confirms that the testosterone group indeed received testosterone and the placebo group did not.

(2009) further strengthens our confidence in the conclusion that testosterone has only a very small effect on risk-taking in women, if at all. First, Zethraeus et al. (2009) recruited an older sample of postmenopausal women aged 50–65 years, in contrast to the young adult female participants aged 18–27 years that were recruited in the present study. Second, Zethraeus et al. (2009) used a prolonged 4-week treatment with testosterone undecanoate, while we used a one-time sublingual testosterone administration method. It is likely that the physiological testosterone curves resulting from each of the administration procedures are very different, making the replication of a nonsignificant effect valuable.

It is important to point out that a nonsignificant effect does not necessarily indicate that testosterone has no effect on risk-taking. However, such a pattern of null results across several studies such as in our meta-analysis, suggests that one should realistically expect the effect of administered testosterone on risk-taking in women to be smaller than a small, standardized effect size. As a matter of fact, none of the published studies had enough power to detect even a small effect size (e.g., to detect an effect size of Cohen's  $d = 0.2$  in a one-tailed  $t$  test with 80% power and an  $\alpha$  level of .05, one would require at least 310 participants per group in a between-subjects design).<sup>6</sup> This inconsistency between reported effects and used sample sizes seems to point to the potential existence of  $p$ -hacking or data-contingent analyses. Indeed, Schäfer and Schwarz (2019) compared the reported effect sizes of publications in psychology journals with and without preregistration and found that effect sizes reported in articles without preregistration were twice as big as effect sizes reported in articles with preregistration. To draw clear conclusions about the true magnitude of administered testosterone's effects on risk-taking, studies with strong methodologies are thus needed. This would involve (a) the use of preregistration and registered reports as well as (b) blinding procedures that extend to data-analysts and reviewers to reduce data-contingent analysis, unintentional biases, and file-drawer effects, (c) blinding procedures that extend to data-analysts and reviewers to further reduce file-drawer effects and unintentional biases, and (d) ideally sufficiently powered studies to produce reliable estimates.

Another factor that may have contributed to the observed null results is that testosterone is a

hormone that may be specifically relevant in social (rather than nonsocial) situations. Across species, in birds, rodents, nonhuman primates, and humans with and without social anxiety symptoms, upcoming social challenges result in the release of testosterone (Bateup et al., 2002; Hutschemaekers et al., 2020; Muller & Wrangham, 2004; Neave & Wolfson, 2003; Wingfield et al., 1990, 2001). Likewise, it has been suggested that testosterone may affect risky behavior specifically in socially challenging situations, for example, when social hierarchies have to be defended (Wingfield et al., 2001). Even though the CCT has a clear affective component, it does not have a social component. Therefore, future work on the role of testosterone in risk-taking might profit from systematically comparing risky behavior in social versus nonsocial settings.

Last, it is also important to consider that our participant sample consisted of women only. While there are articles that have shown effects of exogenous testosterone on decision-making under risk both in female and male participants (Cueva et al., 2015; van Honk et al., 2004), a direct comparison is difficult since there is no validated testosterone administration procedure that can be used in both women and men. Given that basal testosterone levels are generally 5- to 25-fold higher in men than in women (Salameh et al., 2010), one might expect to find stronger testosterone effects in male than in female samples. However, Sapienza et al. (2009) have shown that while increased basal testosterone relates to lower risk aversion in females, this relation does not exist in males. If anything, this would suggest stronger effects of testosterone on risk-taking in women than in men. Furthermore, two recent studies with relatively large sample sizes found no effect of testosterone administration on risk-taking in men (Nadler et al., 2021; Stanton et al., 2021). Future research should thus evaluate to

<sup>6</sup> This inconsistency is generally in line with the results of the meta-analysis of Kurath and Mata (2018), who assessed the effects of *endogenous* testosterone on risk-taking (including studies with behavioral tasks and questionnaire-based measures). They found a small but significant correlation between testosterone and risk-taking ( $r = .12$ , corresponding to a Cohen's  $d$  of 0.24; Borenstein et al., 2009). Importantly, 96% of the studies included in their meta-analysis were underpowered (i.e., well powered studies would require approximately 430 participants to find an effect with 80% power).

what extent our findings generalize to males, which, however, requires ideally the development of a reliable testosterone administration technique that is similarly applicable and comparable in all genders.

To conclude, our testosterone administration study with a triple-blind, placebo-controlled, randomized, between-subjects design provides no evidence for an effect of testosterone on risky decision-making. Using the best-known practices to guarantee unbiased data-collection and analyses, we replicated all expected task effects, yet found no significant main effect or interactions involving testosterone. While we are not the first to report a null effect (e.g., Zethraeus et al., 2009), we believe that our study makes a valuable contribution to the mixed literature on the effects of testosterone on risk-taking. Specifically, our conceptual and independent replication adopted high methodological standards to safeguard against bias and shows that the true effect of administered testosterone on risk-taking in women is smaller than a medium effect size. Furthermore, our meta-analysis across six independent studies is consistent with our observed null effect, as the meta-analysis also found no evidence for a significant effect of testosterone and suggests that the true effect is smaller than a small, standardized effect size. These results contradict strong claims regarding the role of testosterone in risk-taking, raising doubts whether heightened testosterone levels were involved in the financial crisis of 2008, while also being highly relevant to applied scholars targeting excessive risk-taking in the real world.

## References

- Apicella, C. L., Carré, J. M., & Dreber, A. (2015). Testosterone and economic risk taking: A review. *Adaptive Human Behavior and Physiology, 1*(3), 358–385. <https://doi.org/10.1007/s40750-014-0020-2>
- Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., & Little, A. C. (2008). Testosterone and financial risk preferences. *Evolution and Human Behavior, 29*(6), 384–390. <https://doi.org/10.1016/j.evolhumbehav.2008.07.001>
- Apicella, C. L., Dreber, A., & Mollerstrom, J. (2014). Salivary testosterone change following monetary wins and losses predicts future financial risk-taking. *Psychoneuroendocrinology, 39*, 58–64. <https://doi.org/10.1016/j.psychneuen.2013.09.025>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bateup, H. S., Booth, A., Shirtcliff, E. A., & Granger, D. A. (2002). Testosterone, cortisol, and women's competition. *Evolution and Human Behavior, 23*(3), 181–192. [https://doi.org/10.1016/S1090-5138\(01\)00100-3](https://doi.org/10.1016/S1090-5138(01)00100-3)
- Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1*, 33–47. <https://doi.org/10.1037/t13084-000>
- Boksem, M. A. S., Mehta, P. H., Van den Bergh, B., van Son, V., Trautmann, S. T., Roelofs, K., Smids, A., & Sanfey, A. G. (2013). Testosterone inhibits trust but promotes reciprocity. *Psychological Science, 24*(11), 2306–2314. <https://doi.org/10.1177/0956797613495063>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. *Introduction to meta-analysis* (pp. 45–49). Wiley. <https://doi.org/10.1002/97804707443386.ch7>
- Buelow, M. T. (2015). Predicting performance on the Columbia Card Task: Effects of personality characteristics, mood, and executive functions. *Assessment, 22*(2), 178–187. <https://doi.org/10.1177/1073191114539383>
- Bürkner, P. C. (2016). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P. C. (2017). *Brms: Bayesian regression models using stan (Version 1.5.0)*. <https://cran.r-project.org/web/packages/brms/index.html>
- Buskens, V., Raub, W., van Miltenburg, N., Montoya, E. R., & van Honk, J. (2016). Testosterone administration moderates effect of social environment on trust in women depending on second-to-fourth digit ratio. *Scientific Reports, 6*, 1–8. <https://doi.org/10.1038/srep27655>
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Coates, J. M., & Herbert, J. (2008). Endogenous steroids and financial risk taking on a London trading floor. *Proceedings of the National Academy of Sciences of the United States of America, 105*(16), 6167–6172. <https://doi.org/10.1073/pnas.0704025105>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.
- Cueva, C., Roberts, R. E., Spencer, T., Rani, N., Tempest, M., Tobler, P. N., Herbert, J., & Rustichini, A. (2015). Cortisol and testosterone increase



- financial risk taking and may destabilize markets. *Scientific Reports*, 5(1), Article 11206. <https://doi.org/10.1038/srep11206>
- Dariotis, J. K., Chen, F. R., & Granger, D. A. (2016). Latent trait testosterone among 18–24 year olds: Methodological considerations and risk associations. *Psychoneuroendocrinology*, 67, 1–9. <https://doi.org/10.1016/j.psyneuen.2016.01.019>
- Derntl, B., Pintzinger, N., Kryspin-Exner, I., & Schöpf, V. (2014). The impact of sex hormone concentrations on decision-making in females and males. *Frontiers in Neuroscience*, 8, Article 352. <https://doi.org/10.3389/fnins.2014.00352>
- Doi, H., Nishitani, S., & Shinohara, K. (2015). Sex difference in the relationship between salivary testosterone and inter-temporal choice. *Hormones and Behavior*, 69, 50–58. <https://doi.org/10.1016/j.yhbeh.2014.12.005>
- Enter, D., Spinhoven, P., & Roelofs, K. (2014). Alleviating social avoidance: Effects of single dose testosterone administration on approach-avoidance action. *Hormones and Behavior*, 65(4), 351–354. <https://doi.org/10.1016/j.yhbeh.2014.02.001>
- Evans, K. L., & Hampson, E. (2014). Does risk-taking mediate the relationship between testosterone and decision-making on the Iowa Gambling Task? *Personality and Individual Differences*, 61, 57–62. <https://doi.org/10.1016/j.paid.2014.01.011>
- Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in inter-temporal choice. *Nature Neuroscience*, 13(5), 538–539. <https://doi.org/10.1038/nn.2516>
- Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 709–730. <https://doi.org/10.1037/a0014983>
- Figner, B., & Weber, E. U. (2011). Who takes risks when and why? Determinants of risk taking. *Current Directions in Psychological Science*, 20(4), 211–216. <https://doi.org/10.1177/0963721411415790>
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10), Article e1701381. <https://doi.org/10.1126/sciadv.1701381>
- Heany, S. J., Bethlehem, R. A. I., van Honk, J., Bos, P. A., Stein, D. J., & Terburg, D. (2018). Effects of testosterone administration on threat and escape anticipation in the orbitofrontal cortex. *Psychoneuroendocrinology*, 96, 42–51. <https://doi.org/10.1016/j.psyneuen.2018.05.038>
- Heany, S. J., van Honk, J., Stein, D. J., & Brooks, S. J. (2016). A quantitative and qualitative review of the effects of testosterone on the function and structure of the human social-emotional brain. *Metabolic Brain Disease*, 31, 157–167. <https://doi.org/10.1007/s11011-015-9692-y>
- Hermans, E. J., Bos, P. A., Ossewaarde, L., Ramsey, N. F., Fernández, G., & van Honk, J. (2010). Effects of exogenous testosterone on the ventral striatal BOLD response during reward anticipation in healthy women. *NeuroImage*, 52(1), 277–283. <https://doi.org/10.1016/j.neuroimage.2010.04.019>
- Hutschmaekers, M. H., de Kleine, R. A., Hendriks, G. J., Kampman, M., & Roelofs, K. (2021). The enhancing effects of testosterone in exposure treatment for social anxiety disorder: A randomized proof-of-concept trial. *Translational Psychiatry*, 11, 1–7. <https://doi.org/10.1038/s41398-021-01556-8>
- Hutschmaekers, M. H. M., de Kleine, R. A., Davis, M. L., Kampman, M., Smits, J. A. J., & Roelofs, K. (2020). Endogenous testosterone levels are predictive of symptom reduction with exposure therapy in social anxiety disorder. *Psychoneuroendocrinology*, 115, Article 104612. <https://doi.org/10.1016/j.psyneuen.2020.104612>
- Kruschke, J. K. (2013). *How much of a Bayesian posterior distribution falls inside a region of practical equivalence (ROPE)*. Doing Bayesian Data Analysis. <http://doingbayesiandataanalysis.blogspot.com/2013/08/how-much-of-bayesian-posterior.html>
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280. <https://doi.org/10.1177/2515245918771304>
- Kurath, J., & Mata, R. (2018). Individual differences in risk taking and endogenous levels of testosterone, estradiol, and cortisol: A systematic literature search and three independent meta-analyses. *Neuroscience and Biobehavioral Reviews*, 90, 428–446. <https://doi.org/10.1016/j.neubiorev.2018.05.003>
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26), Article 772. <https://doi.org/10.21105/joss.00772>
- Mehta, P. H., Welker, K. M., Zilioli, S., & Carré, J. M. (2015). Testosterone and cortisol jointly modulate risk-taking. *Psychoneuroendocrinology*, 56, 88–99. <https://doi.org/10.1016/j.psyneuen.2015.02.023>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & for the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, 6, Article e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Muller, M. N., & Wrangham, R. W. (2004). Dominance, aggression and testosterone in wild chimpanzees: A test of the ‘challenge’ hypothesis. *Animal Behaviour*, 67(1), 113–123. <https://doi.org/10.1016/j.anbehav.2003.03.013>



- 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*, 286(1910), Article 20191062. <https://doi.org/10.1098/rspb.2019.1062>
- Nadler, A., Jiao, P., Johnson, C. J., Alexander, V., & Zak, P. J. (2018). The bull of wall street: Experimental analysis of testosterone and asset trading. *Management Science*, 64(9), 4032–4051. <https://doi.org/10.1287/mnsc.2017.2836>
- Nadler, A., Wibrál, M., Dohmen, T., Falk, A., Previtro, A., Weber, B., Camerer, C., Almenberg, A. D., & Nave, G. (2021). Does testosterone increase willingness to compete, confidence, and risk-taking in men? Evidence from two randomized placebo-controlled experiments. *PsyArXiv*. <https://doi.org/10.31234/osf.io/62af7>
- Nave, G., Nadler, A., Zava, D., & Camerer, C. (2017). Single-dose testosterone administration impairs cognitive reflection in men. *Psychological Science*, 28(10), 1398–1407. <https://doi.org/10.1177/0956797617709592>
- Neave, N., & Wolfson, S. (2003). Testosterone, territoriality, and the 'home advantage'. *Physiology & Behavior*, 78(2), 269–275. [https://doi.org/10.1016/S0031-9384\(02\)00969-1](https://doi.org/10.1016/S0031-9384(02)00969-1)
- Pachur, T., Hertwig, R., & Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, 1(1), 64–78. <https://doi.org/10.1037/dec0000006>
- Reavis, R., & Overman, W. H. (2001). Adult sex differences on a decision-making task previously shown to depend on the orbital prefrontal cortex. *Behavioral Neuroscience*, 115(1), 196–206. <https://doi.org/10.1037/0735-7044.115.1.196>
- Salameh, W. A., Redor-Goldman, M. M., Clarke, N. J., Reitz, R. E., & Caulfield, M. P. (2010). Validation of a total testosterone assay using high-turbulence liquid chromatography tandem mass spectrometry: Total and free testosterone reference ranges. *Steroids*, 75(2), 169–175. <https://doi.org/10.1016/j.steroids.2009.11.004>
- Sapienza, P., Zingales, L., & Maestriperieri, D. (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36), 15268–15273. <https://doi.org/10.1073/pnas.0907352106>
- Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, Article 813. <https://doi.org/10.3389/fpsyg.2019.00813>
- Schipper, B. C. (2012, January). *Sex hormones and choice under risk* (Working Paper, No. 12–7). <https://doi.org/10.2139/ssrn.2046324>
- Smith, K. M., & Apicella, C. L. (2017). Winners, losers, and posers: The effect of power poses on testosterone and risk-taking following competition. *Hormones and Behavior*, 92, 172–181. <https://doi.org/10.1016/j.yhbeh.2016.11.003>
- Stan Development Team. (2021). *Stan modeling language users guide and reference manual, Version 2.27*. <https://mc-stan.org>
- Stanton, S. J. (2017). The role of testosterone and estrogen in consumer behavior and social & economic decision making: A review. *Hormones and Behavior*, 92, 155–163. <https://doi.org/10.1016/j.yhbeh.2016.11.006>
- Stanton, S. J., Lienes, S. H., & Schultheiss, O. C. (2011). Testosterone is positively associated with risk taking in the Iowa Gambling Task. *Hormones and Behavior*, 59(2), 252–256. <https://doi.org/10.1016/j.yhbeh.2010.12.003>
- Stanton, S. J., Mullette-Gillman, O. A., McLaurin, R. E., Kuhn, C. M., LaBar, K. S., Platt, M. L., & Huettel, S. A. (2011). Low- and high-testosterone individuals exhibit decreased aversion to economic risk. *Psychological Science*, 22(4), 447–453. <https://doi.org/10.1177/0956797611401752>
- Stanton, S. J., Welker, K. M., Bonin, P. L., Goldfarb, B., & Carré, J. M. (2021). The effect of testosterone on economic risk-taking: A multi-study, multi-method investigation. *Hormones and Behavior*, 134, Article 105014. <https://doi.org/10.1016/j.yhbeh.2021.105014>
- Terburg, D., Aarts, H., & van Honk, J. (2012). Testosterone affects gaze aversion from angry faces outside of conscious awareness. *Psychological Science*, 23(5), 459–463. <https://doi.org/10.1177/0956797611433336>
- Tuiten, A., Van Honk, J., Koppeschaar, H., Bernaards, C., Thijssen, J., & Verbaten, R. (2000). Time course of effects of testosterone administration on sexual arousal in women. *Archives of General Psychiatry*, 57(2), 149–153. <https://doi.org/10.1001/archpsyc.57.2.149>
- Tymula, A., Rosenberg Belmaker, L. A., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), 17135–17140. <https://doi.org/10.1073/pnas.1207144109>
- van Honk, J., Montoya, E. R., Bos, P. A., van Vugt, M., & Terburg, D. (2012). New evidence on testosterone and cooperation. *Nature*, 485(7399), E4–E5. <https://doi.org/10.1038/nature11136>
- van Honk, J., Schutter, D. J. L. G., Hermans, E. J., Putman, P., Tuiten, A., & Koppeschaar, H. (2004). Testosterone shifts the balance between sensitivity for punishment and reward in healthy young women. *Psychoneuroendocrinology*, 29(7), 937–943. <https://doi.org/10.1016/j.psyneuen.2003.08.007>

- van Honk, J., Will, G. J., Terburg, D., Raub, W., Eisenegger, C., & Buskens, V. (2016). Effects of testosterone administration on strategic gambling in poker play. *Scientific Reports*, *6*(1), Article 18096. <https://doi.org/10.1038/srep18096>
- van Rooij, K., Bloemers, J., de Leede, L., Goldstein, I., Lentjes, E., Koppeschaar, H., Olivier, B., & Tuiten, A. (2012). Pharmacokinetics of three doses of sublingual testosterone in healthy premenopausal women. *Psychoneuroendocrinology*, *37*(6), 773–781. <https://doi.org/10.1016/j.psyneuen.2011.09.008>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Wagels, L., Votinov, M., Radke, S., Clemens, B., Montag, C., Jung, S., & Habel, U. (2017). Blunted insula activation reflects increased risk and reward seeking as an interaction of testosterone administration and the MAOA polymorphism. *Human Brain Mapping*, *38*(9), 4574–4593. <https://doi.org/10.1002/hbm.23685>
- Weller, J. A., King, M. L., Figner, B., & Denburg, N. L. (2019). Information use in risky decision making: Do age differences depend on affective context? *Psychology and Aging*, *34*(7), 1005–1020. <https://doi.org/10.1037/pag0000397>
- Wingfield, J. C., Hegner, R. E., Duffy, A. M., Jr., & Ball, G. F. (1990). The “challenge hypothesis”: Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *American Naturalist*, *136*(6), 829–846. <https://doi.org/10.1086/285134>
- Wingfield, J. C., Lynn, S., & Soma, K. K. (2001). Avoiding the ‘costs’ of testosterone: Ecological bases of hormone-behavior interactions. *Brain, Behavior and Evolution*, *57*(5), 239–251. <https://doi.org/10.1159/000047243>
- Woyke, I., Iking, I., Heuvelmans, V., Roelofs, K., & Figner, B. (under revision). The effect of exogenous testosterone on decision-making under risk and ambiguity. *Hormones and Behavior*.
- Wu, Y., Liu, J., Qu, L., Eisenegger, C., Clark, L., & Zhou, X. (2016). Single dose testosterone administration reduces loss chasing in healthy females. *Psychoneuroendocrinology*, *71*, 54–57. <https://doi.org/10.1016/j.psyneuen.2016.05.005>
- Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., von Schoultz, B., Hirschberg, A. L., & Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6535–6538. <https://doi.org/10.1073/pnas.081275710>

Received November 19, 2020

Revision received June 24, 2022

Accepted June 27, 2022 ■