universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Investigating the Influence of Testosterone on Aggressive and Prosocial Behavior"

verfasst von / submitted by

Felix Dörflinger, BA, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Univ.-Prof. Dr. Claus Lamm
Mitbetreut von / Co-Supervisor:	

Table of Contents

Introduction

The history of research on the effects of testosterone in social interactions is long and diverse. In the beginning, scientists focused on the role of testosterone in the context of aggressive behavior. Experimental studies with animals provided evidence that testosterone is associated with aggressive behavior in a variety of species (Almeida, Cabral, & Narvaes, 2015; Archer, 1988). As early as the 1930s, Allee, Collias, and Lutherman (1939) observed increased aggressive behavior in hen and successively also a changed social order in the flock after testosterone administration. In female mice, testosterone administration similarly elevated aggressive behavior (Edwards, 1969). Early correlational studies in humans also attempted to establish this link between testosterone and aggression. Indeed, higher testosterone levels were found in male prison inmates who were convicted for more severe and violent crimes than in inmates who were convicted for non-violent crimes (Dabbs, Carr, Frady, & Riad, 1995; Dabbs, Frady, Carr, & Besch, 1987). However, one issue remained unclear due to the design of these studies: whether more pronounced aggressive behavior was the effect of higher testosterone levels or if higher testosterone levels were caused, for example, by environmental factors that might have also increased the likelihood of aggressive behavior. The highly competitive, stressing, and challenging social environment of a prison might have been such a factor. Overall, only a very small correlation (r = .08) between testosterone levels and aggressive behavior was found in a review of 30 correlational studies (Archer, Graham-Kevan, & Davies, 2005).

The formulation of the challenge hypothesis by Wingfield, Hegner, Dufty, and Ball (1990) consequently led to change in perspective. Wingfield et al. (1990) observed that testosterone levels in male birds increased during the breeding season, that challenging social situations (e.g., competitive situations) caused a further increase in testosterone levels, and that these changes in testosterone levels were associated with aggressive behavior. Based on these observations, the authors proposed the challenge hypothesis which states that: "circulating levels of testosterone correlate with aggression only during periods of social instability and, male-male interactions over status and access to sexually receptive females tend to increase testosterone secretion" (Wingfield, Lynn, & Soma, 2001, p. 242). The challenge hypothesis thereby established the idea that the effects of testosterone are context-dependent and was used as a framework for investigating the relationship of testosterone and aggression in humans (Archer, 2006). Instead of focusing on correlational associations between trait-level testosterone and aggressive behavior, studies have now acknowledged the

influence of the social context (Carré & Archer, 2018). To clarify the relationship between testosterone and aggressive behavior, more refined scientific approaches have then been employed. Studies began to use behavioral tasks to measure if aggressive behavior can be increased by administering testosterone instead of relying on correlations between self-report measurements of aggression and measurements of endogenous testosterone levels. Carré et al. (2017) revealed that testosterone increased aggressive reactions in a sample of 121 men in a Point Subtraction Aggression Paradigm (PSAP) task. The PSAP task is a decision-making game that requires participants to gain points to earn money. During the task, which is disguised as an online game with another player, the participant experiences provocations in the form of point subtractions, and the participant can react by retaliating, thereby reducing the other player's points. This act is considered an aggressive one that is intended to harm the other person. In a placebo-controlled double-blind study, Carré et al. (2017) identified more point reductions (aggressive reactions) in the testosterone compared with the placebo group. However, the effect was present only in highly dominant individuals or individuals with low self-control. Using the PSAP task, Geniole et al. (2019) again found evidence that testosterone increased aggressive behavior in men with certain personality traits. To simplify the analysis by reducing the number of moderators, the authors combined the measurements of personality traits associated with aggressive behavior (dominance, self-control, and self-construal) into a single score. Men who scored high on this score (described themselves to be high in dominance, low in self-control, and with an independent self-construal) demonstrated increased aggressive behavior when given testosterone (Geniole et al., 2019). In a large sample of 308 men, the authors also identified related physiological mechanisms. Genetic polymorphisms (fewer cytosine-adenine-guanine (CAG) repeats in exon 1), which lead to increased androgen receptor efficiency, strengthened the association between testosterone and aggressive behavior (Geniole et al., 2019). These studies, based on the PSAP task, support the idea of a causal effect of testosterone on reactive aggressive behavior that is used as a response to provocation (the point subtractions in the PSAP) and provide support for the challenge hypothesis.

The challenge hypothesis and research stemming from this background may seem to primarily support the idea of testosterone as a promoter for aggressive behavior. However, this hypothesis is not the only theoretical approach for explaining the effects of testosterone on human behavior. Mazur and Booth (1998) established the idea that testosterone primarily influences dominance (status-related) behavior. This idea is often referred to as social status hypothesis. An individual is said to "act *dominantly* if [the] apparent intent is to achieve or

maintain high status – that is, to obtain power, influence, or valued prerogatives – over a conspecific" (Mazur & Booth, 1998, p. 353). Although aggressive behavior can be used as a means of achieving higher status in certain situations, among humans, other forms of behavior (e.g., prosocial or generous actions aimed at increasing the value of an individual in a social group) may serve this purpose equally well (Anderson & Kilduff, 2009). If testosterone is indeed linked to status relevant behavior, experimental evidence that testosterone influences behavior in social interactions in forms other than aggressive reactions should therefore be possible to find as well.

Eisenegger, Naef, Snozzi, Heinrichs, and Fehr (2010) investigated this assumption using the ultimatum game. In the ultimatum game, one player proposes another player a share of a fixed amount of money. If the other player accepts, both players get their share; if the other player rejects the offer, neither player earns anything. Eisenegger et al. (2010) argued that "if testosterone induces a higher concern for social status in the sense that subjects want to prevent their proposal from being turned down, we should observe fairer offers among proposers who received testosterone" (p. 357). In a sample of 121 women, the authors found that testosterone increased fair bargaining behavior in the ultimatum game. Women who received a dose of testosterone offered more money to the other player than subjects who received a placebo; furthermore, no differences were found in the rejection rates between the groups, ruling out the possibility that testosterone increases altruistic behavior, in which case lower offers should also be accepted in the testosterone group (Eisenegger et al., 2010). This result indicates that in women, testosterone influences non-aggressive status-relevant behavior.

To examine the possible effects of testosterone on status-relevant behavior in men, Dreher et al. (2016) used a modified version of the ultimatum game in a double-blind placebo-controlled study. Their modification of the ultimatum game enabled the participants to punish or reward the other player after the offer was rejected or accepted, by giving or taking money at a cost to themselves. In this study, the participants could only respond to the offers of the other player, and they were never in the position to propose. To prevent the strategic use of the options to punish or reward the other player, the participants knew that the behavior of the other player was prerecorded. The results indicated that participants who received testosterone chose to punish the other player more often when confronted with an unfair offer and that more and greater rewards were given from participants who received testosterone if the other player offered a generous amount of money (Dreher et al., 2016). This outcome provided further evidence that testosterone is associated with behavior that

promotes social status. In an aggressive or provocative context, testosterone increases aggressive behavior; by contrast, friendly or prosocial actions are promoted in a friendly context because in such situations, aggressive behavior would not be constructive to gain or maintain social status (Dreher et al., 2016).

Boksem et al. (2013) also found evidence that testosterone increases prosocial behavior. In a one-shot trust game, 54 female participants decided if (and how much) they wanted to invest in a trustee from an endowment of €20. The investment would be tripled, then the trustee had the opportunity to return a share of the investment to the investor. Each participant played both roles; for the trustees, the amount of money they received was fixed and always high (€60), indicating the high trust of the investor. Interestingly, participants who received testosterone exhibited less trust and decided to invest less money but returned more money to the investor when confronted with the generous investment compared with the placebo group (Boksem et al., 2013).

Meanwhile, Wagels et al. (2018) investigated the effects of exogenous testosterone on aggressive behavior using the Taylor Aggression Paradigm (TAP) in a sample of 103 men. The TAP is a game that allows the participant to decide if he wants to subtract money from an opponent if he reacts faster in a reaction time game. The participants won a fixed amount of money if they reacted faster but lost a variable amount of money determined by the opponent if they reacted slower. The results indicated that testosterone influenced how participants reacted during the TAP. After losing, participants who received testosterone reacted more aggressively (chose to reduce more money in the following trial) than participants who received a placebo, but they also acted less aggressively when the provocation was low than participants in the placebo group (Wagels et al., 2018).

Thus far, recent studies using single-dose testosterone administration procedures provided support for testosterone effects on both aggressive reactions to provocation and prosocial (fair or rewarding) actions, depending on the context and individual personality traits. Geniole and Carré (2018) proposed a new theoretical framework that incorporates these findings and other models on the relationship between testosterone and human social behavior, the so-called *Fitness Model of Testosterone Dynamics*. The idea is that the behavioral effects of testosterone are mediated by contextual cues, which can trigger either a threat or reward system; in turn, rapid changes in testosterone levels can occur, inducing behavior that responds to the situation in a manner that promotes and protects fitness (Geniole & Carré, 2018). The concept of “fitness” remains undefined but seems to have a broad meaning in the sense of reproductive success, status, access to resources, and the means of

achieving and maintaining all of these factors. Importantly, individual factors such as personality traits (as reported in the aforementioned PSAP studies) and physiological characteristics influence the mode of operation between the components in the model (e.g., how a situation is perceived, how strongly certain situations influence individual hormonal responses, and how these responses translate in behavioral changes) (Geniole & Carré, 2018). To illustrate this postulation, consider the following study for an example of a possible moderating effect on the relationship of testosterone and aggression. Prasad et al. (2017) presented preliminary (correlational) evidence that stress moderates the relationship between testosterone and aggressive behavior in a small sample of 39 participants (19 males, 20 females). Using the Trier Social Stress Task, the authors induced high stress in half of the participants and reduced stress using a relaxation task in the other half of the participants. In the ultimatum game, retaliation of unfair (very low) offers was used to measure aggressive behavior. The results indicated that higher testosterone predicted more rejections of unfair offers only in the low stress but not in the high stress condition. However, in men with high testosterone levels, the stress manipulation increased retaliatory behavior, whereas in women, high stress reduced retaliatory behavior (Prasad et al., 2017).

To summarize, previous research suggests that testosterone is associated with both aggressive and prosocial behavior. Individual characteristics such as certain personality traits and stress seem to moderate the effects of testosterone on aggressive behavior. Thus far, no study has systematically examined the possible moderating effects of personality traits and stress on the association between testosterone and prosocial behavior. With the study presented here, we intended to further clarify the relationship between testosterone, aggressive and prosocial behavior, different contextual cues, and individual personality traits. In a placebo-controlled double-blind testosterone administration study, male participants completed a modified PSAP task (described in detail below). The task offered the option to respond in an aggressive or prosocial manner. It was implemented in a way that allowed us to control for baseline differences in the participants' prosocial and aggressive behavior and to experimentally manipulate the context to systematically measure behavioral responses in both a friendly and a provocative situation. In accordance with previous studies, we hypothesized that testosterone increases the aggressive reactions of men to provocations and that testosterone also increases prosocial behavior in a situation that requires friendly responses. In addition, we controlled for relevant personality traits (dominance, self-control) and perceived stress of the participants because we assumed that similar to previous studies, these individual traits influence the effects of testosterone on both aggressive and prosocial behavior.

Method

Participants

Participants were recruited using online advertisement on popular social media platforms and leaflets placed in the university buildings. Potential participants had to complete an online survey to assess their eligibility for the study. This study excluded individuals who reported to currently use drugs or medication, suffer from heart disease or psychiatric disorders, and presently study, or have studied, psychology; smokers were also excluded. In total, 219 eligible persons with a mean age of 24.9 years (*SD* = 3.8, age range: 19–38 years) participated in the study. Four participants did not complete the task because of technical problems, one dataset was corrupted and had to be dismissed, one participant quit the study early, and one participant was excluded as an outlier in the behavioral measurement (described in detail below). The final dataset consisted of 212 participants. Of these 212 participants, 166 (78.3%) reported German as their mother tongue, and 46 (21.7%) reported a different cultural background. Asked for their perceived ethnicity, the majority self-identified as Europeans (150 participants), three as Asians, two as South Americans, and the remaining participants declined to provide any information. Most participants were students (81%), and only a minority (9%) did report to have achieved merely lower levels of education. The sample size was based on the requirements of the other tasks in the study. Therefore, no power analysis was conducted with respect to the requirements of the study reported here. All the participants provided written informed consent to the procedures of the study and received monetary compensation. The ethics committee of the Medical University of Vienna approved the study.

Procedure

On arrival, the participants were instructed on the procedures of the study by the male research assistant conducting the study. They then provided informed consent and filled out questionnaires related to other tasks in the study. Afterwards, the gel containing either 150 mg of testosterone (Testogel®, Laboratoires Besins International, France) or a placebo was applied on the upper torso of the participant using a double-blind procedure. Ninety-five participants (44.8%) received the placebo gel and 117 participants (55.2%) the testosterone gel. To reduce the risk of contamination, only testosterone gel or only placebo gel was used in a single testing session. Participants were randomly assigned to different testing sessions (placebo or testosterone condition). Following the procedures described in other studies (Carré et al., 2017; Wagels et al., 2018), a waiting period (two hours) followed the application to wait for the gel to take effect. At the beginning of the waiting period, the participants filled

out personality questionnaires and questionnaires regarding their socioeconomic status and cultural background. After the waiting period, the participants completed four unrelated tasks, interrupted by short breaks and saliva samples. The results of these tasks are reported elsewhere. The modified PSAP task used in this study was completed last by the participants. During the study, six saliva samples were obtained at different time points to measure the participants' testosterone and cortisol levels (see Fig. 1 for a procedural timeline). At the end of the study, the participants filled out questionnaires concerning the tasks and their belief about receiving testosterone or a placebo. All the testing sessions were conducted in the afternoon, and each testing session lasted approximately five hours. All the participants (between one and six) were simultaneously tested in the same room.

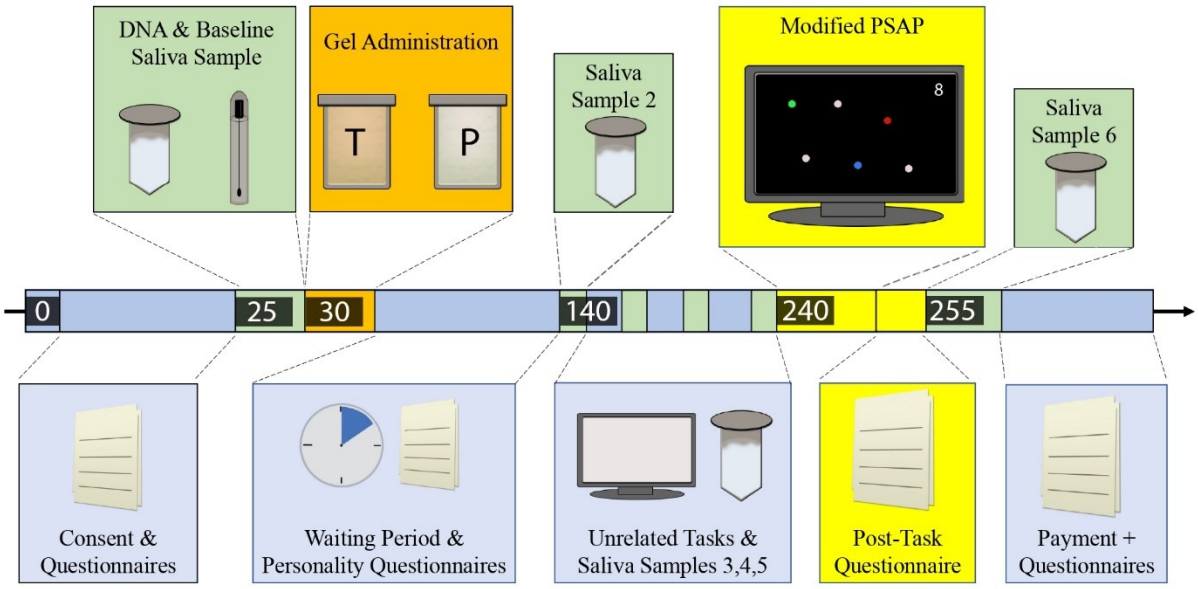


Figure 1. Timeline of the experiment. Numbers indicate minutes passed (approx.) since the start of the study.

Questionnaires

The German version of the Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983) in its 10-item version was used for measuring the participants' self-reported stress. The PSS-10 asks for an assessment of how often during the last month a person had the sense of being unable to control its life or felt irritated or stressed. To measure trait dominance and prestige, the participants completed the German versions of the Dominance-Prestige Scales (Cheng, Tracy, & Henrich, 2010). Eight items measure dominance (attaining social status using intimidating or aggressive strategies) and nine items measure prestige (attaining social status by possessing expertise or skills). The German version of the Barratt Impulsiveness Scale (BIS-15; Meule, Vögele, & Kübler, 2011) was used for measuring the participants' self-

control (impulsivity). The BIS-15 measures impulsivity as a tendency to act inconsiderately without thinking about possible negative outcomes. At the end of the two-hour waiting period, the participants completed the Multidimensional Mood State Questionnaire in its German version (Steyer, 1997) to assess any mood state differences between the control and experimental group prior to the tasks. Additionally, after finishing the task, the participants completed a post-task questionnaire that asked for their opinion about the task itself, the other players' behavior, and their own behavior. The aim of this questionnaire was to assess how the participants perceived the task (difficult, funny, exhausting or demanding), how fair the participants rated their own behavior, as well as the behavior of the other players and to evaluate if the participants believed that they were playing with "real" people or if they realized that the setup was preprogrammed. At the end of the study, the participants filled out a questionnaire that asked for their belief about receiving testosterone or a placebo and how certain they were about this belief. Following the procedure described by Geniole et al. (2019), we decided to combine the personality measurements into a composite measurement (risk score). This measurement included averaging the standardized scores of the DPS (dominance and prestige) and BIS-15 (self-control) scales. A high value on the risk score represents persons who describe themselves as dominant and impulsive (low in self-control).

Modified Point Subtraction Aggression Paradigm

To answer the research questions, a modified version of the well-established PSAP task was used. The task was programmed and presented using PsychoPy (Peirce, 2008) and was designed as a tool for measuring not only aggressive behavior (like the standard PSAP) but also prosocial behavior. Similar to the original task, the modified version has the objective to gain points, and it is also disguised as an online game with other players. The more points the participants gained, the more money they earned. However, the participants did not know in advance the amount of money they could earn during this task and the worth of each point. They simply knew that more gained points result in a higher payout. In the original PSAP task, the participants must press a button for 100 consecutive times to gain one point. From time to time, the points of the participants are randomly reduced, and the participant has the option to retaliate by reducing the other player's points (the count of this action is the aggression measurement) or to choose a non-aggressive option to protect his own points from being reduced. The PSAP task is often used and considered to be a reliable measurement of reactive aggression to provocation (Geniole, MacDonell, & McCormick, 2017).

The modified PSAP task used in the current study substantially differs from the original task in several aspects. (a) The participants do not only have the option to reduce the other

player's points, but they can also give points to the other player. Similar to the original task, reducing points is considered an aggressive action in the modified PSAP task. The additional option to give points to the other player implemented in the modified version of the PSAP is considered a prosocial action. (b) During the task, the participants were confronted with three different types of players (experimental conditions), and the actions of the participants were recorded in each condition (how many points they gave to the other player and how many points they reduced). (c) Instead of repeatedly pressing a button, the modified PSAP requires the participants to respond correctly in a signal detection task to gain points. This step was undertaken to keep the task brief, gain a better control over the difficulty of gaining points (by adjusting the timing and visibility of the signal detection task), and ensure that gaining points is not too easy but is in need of effort in the form of attention.

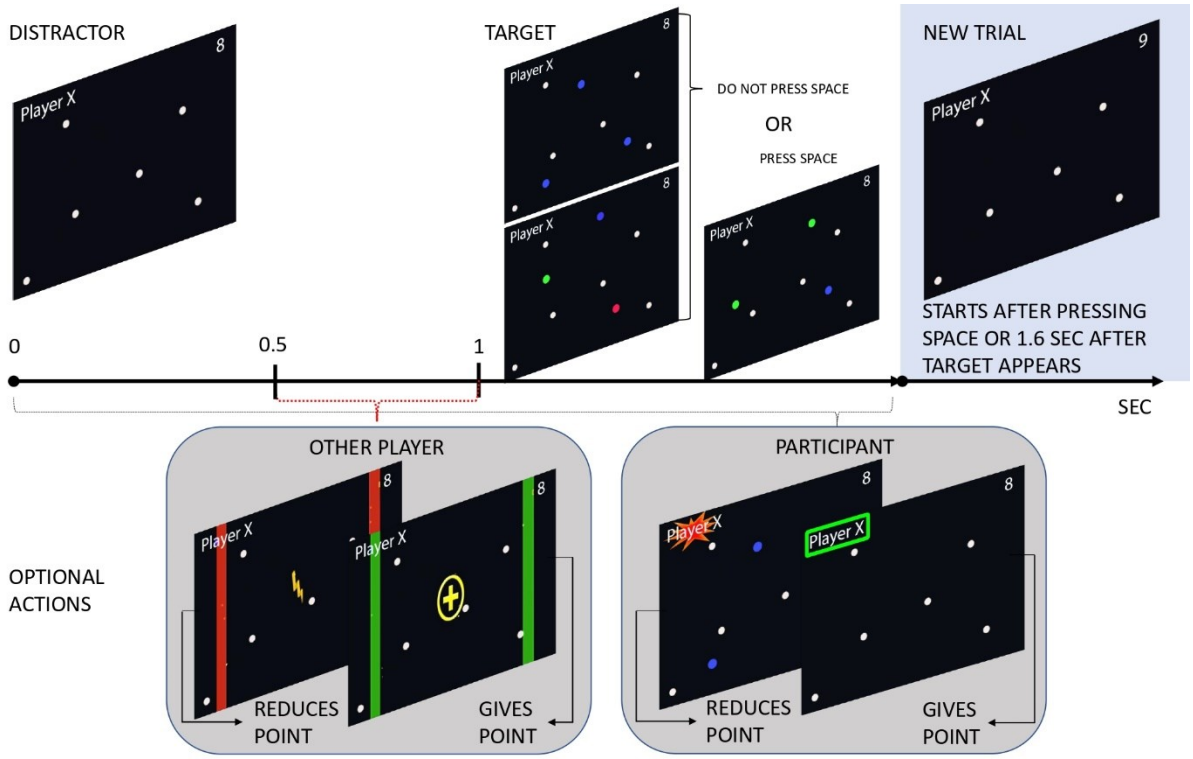


Figure 2. Schematic trial of the modified PSAP task.

The task consisted of 204 trials (see Fig. 2 for a schematic trial). In every trial, participants were asked to observe a black screen with a large number of white, randomly moving distractor dots. To the participants, the subsequent trials appeared as a constant flow of dots, no break occurred between the trials, and the distractor dots were permanently visible. After a randomly determined amount of time, three slightly larger colored dots appeared. These target dots all had either the same color, different colors, or two of the three dots had the same color. If two of the three dots had the same color, the participant was instructed to

press a button (space) to gain one point. This was the case in 120 of 204 trials. If the participant pressed space in the other two cases (where all three dots had the same or different colors), he lost one point. If the participant correctly did not press the space bar in trials where all three target dots had the same or different colors, he also gained one point. Gaining one point in every trial by responding correctly was therefore possible. Each trial ended if the participant pressed space or 1.6 seconds after the target dots appeared if he did not press space. The next trial began with only the white distractor dots on screen. In between the trials, there was neither a black nor a blank screen, and the distractor dots were always visible to establish a flow of the game. To prevent an unnecessarily fast progression of the individual trials, it was ensured that the target dots did only appear after a minimum of one second (on average 1.5 seconds) where only the distractor dots were visible. Each trial lasted, on average, between two and three seconds, depending on how rapidly the participants responded and how long the duration of the randomly determined interval was before the target dots appeared.

A point counter in the right upper corner indicated the number of points that the participant earned. During the instructions (provided in full length in the supplementary materials: Task Instructions), each participant had to complete a practice run of the task (17 trials). In this practice run, a minimum of six points had to be gained to proceed to the main game. If a participant failed to achieve this point threshold, he had to repeat the practice run after having the opportunity to once again read through a synopsis of the instructions. Before playing, each participant had to choose an alias (e.g., “Player M”) that the other players saw during the game. In the instructions, the participants were told that they will be connected over the internet with other players who are taking part in a different study before the game starts. A loading screen, suggesting to the participants that they are being connected to the other players, was shown before the start of the task to increase the credibility of this cover story.

During the task, the alias of the other player, for example “Player X”, was shown in the left upper corner of the screen. If the participant chose to give the other player a point (by pressing the left shift button on the keyboard), a green signal briefly became visible around the other player’s alias. If the participant reduced the other player’s points (by pressing the left control key on the keyboard), a red signal was briefly shown. In case the other player (the program) reduced the points of the participant, a highly visible red signal is shown on screen for 500 ms. If the other player gave the participant a point, an equally visible green signal is shown on screen for 500 ms. An action of the other player always occurred 500 ms after a

trial started, and the signal was visible only until the target dots appeared. However, the participants could give or reduce points at any point during the trial. Importantly, similar to the original task, the options to give the other player a point and to reduce points are costly. If the participant chose one of these options in a trial, he could not gain one point for himself in this trial if he reacted correctly. This detail was clearly explained in the instructions. The more often participants chose to give or reduce points, the less money they would have earned themselves. Additionally, every trial allowed for the selection of only one option. The participant could either choose to give one point to the other player or to reduce the other player's points, but he could not choose both options in one trial. The participants were told in the instructions that they were randomly assigned to the condition of the experiment where they cannot keep the points that they reduced from the other player.

The 204 trials were divided into three blocks (experimental conditions). In each condition, a different player (a preprogrammed type of behavior) was paired with the participant and they were informed on screen when the player changed. Each participant played with three other players (68 trials per player): a neutral player (baseline condition), an aggressive player (aggressive condition), and a prosocial player (prosocial condition). All the participants were first paired with the neutral player. This neutral player took exactly two points away from the participants (in trial 18 and 25) and gave them two points (in trial 8 and 38). This condition was the same for all the participants. The neutral player was deployed to increase the credibility of the task and provide a baseline measurement of the participants' response. After the neutral player, each participant was confronted with the prosocial and aggressive player, but the sequence of these two players was counterbalanced across participants. In randomly defined trials during the corresponding block, the prosocial player exclusively gave points to the participant, and the aggressive player exclusively took away points from the participant. If a point was given or taken away in a trial by the program, no such action would have occurred in the succeeding trial. When the player changed, the point counter was reset to zero to ensure that regardless of the order of the conditions, the participant always started with a point count of zero in each condition. This aspect was also explained to the participants in the instructions. At the end of the task, the participants saw the total number of points that they had earned.

Prior to the study, this modified PSAP task was evaluated with a small sample of psychology students ($N = 13$) to assess the difficulty of the signal detection task, check for the comprehensibility of the instructions, and screen if the task can even elicit the relevant behavior. According to the results from this pilot study, slight adjustments were made (e.g.,

adapting the timing of the trials to avoid making gaining points too difficult), and the amount of money the participants could earn in the task was set based on the pilot study results (1 point = 2 Euro cents). Theoretically, the most profitable strategy for both players would be to only give points to each other. Doing so would eliminate the need to gain points by correctly reacting in the task. However, choosing this strategy would be of high risk because of the lack of a possibility to communicate with the other player. Only one person decided to follow this strategy that resulted in a count of points given twice as high as the next highest count. This one participant was excluded from the analysis for non-technical reasons, as including such data would most likely have biased the results, and the behavior was not comparable to that of the other participants. Due to programming issues and pressure of time, the first 41 participants completed the task with instructions printed out on paper, and the remaining 171 participants were instructed directly on screen (the instructions certainly remained the same). In total, the task (excluding the instructions and the practice trials) took approximately 10 minutes to complete. The source files to run the task using PsychoPy, the data from the study, and all the R-scripts to replicate the results reported here will be made publicly available in an Open Science Framework project.

Results

Questionnaires

Inspection of the data revealed several missing values in the questionnaire data. We decided to replace the missing values. The PSS-10 data contained six missing values (0.3% missing data), which were replaced with the median rating of the corresponding items. The BIS-15 data contained seven missing values (0.2% missing data), which were also replaced with the median rating. In the mood questionnaire (MDBF), 17 missing values (0.3% missing data) have been replaced using the neutral value 3. In the post-task questionnaire, one missing value (0.05% missing data) has been replaced with the median rating. No missing values were present in the DPS data. Robust independent samples trimmed mean tests (Yuen's test) were then used to control for possible group differences in these personality and mood measurements (Mair & Wilcox, 2018). We decided to use robust methods because (a) for most scales, visual inspection of the data revealed single outlying values that might distort results from non-robust statistics and (b) Shapiro-Wilk tests for normality suggested significant deviations from a normal distribution. The resulting mean values for all personality and mood questionnaires for each group, the calculated reliability of all scales, and the results from the robust tests on group differences can be found in the supplement (Questionnaire Results). For most scales, Yuen's test revealed no significant difference

between the groups. However, on the good–bad scale of the MDBF mood questionnaire, Yuen's test was significant, $T_y(107.5) = 2.75$, $p = .007$, $\xi = 0.32$, 95% CI [0.09, 0.49]. This result indicates that the testosterone group reported to feel better than the placebo group, and the robust effect size measure ξ qualifies this effect as medium sized. The standardized scores from the DPS and the BIS-15 have been averaged to form a composite measure of personality traits associated with the behavior in question (risk score). The resulting mean scores were -0.09, 95% CI [-0.20, 0.02] in the placebo group and 0.07, 95% CI = [-0.03, 0.17] in the testosterone group. Yuen's test revealed that this difference between the groups is significant but small, $T_y(125.7) = 2.37$, $p = .019$, $\xi = 0.23$, 95% CI [0.03, 0.41].

Ratings from 211 participants (one participant had to leave the study directly after the task and could not answer the questionnaire) were collected with the post-task questionnaire. Several items were used to assess the positive characteristics of the task (e.g., how funny or exciting was the task). We formed a scale with a range of 1 (*not at all*) to 8 (*very much*) by averaging these six positive items (two had to be reversed). The resulting measure of reliability for this scale was good, Cronbach's $\alpha = .83$, 95% CI [.79, .87], and the resulting mean rating was 4.99, 95% CI [4.70, 5.28] in the placebo group and 5.21, 95% CI [4.94, 5.48] in the testosterone group. Similarly, a scale was formed by averaging four items asking for demanding characteristics (e.g., how exhausting or difficult was the task). The resulting scale had an acceptable reliability, Cronbach's $\alpha = .73$, 95% CI [.66, .79], and a range from 1 to 8. The mean rating was 4.42, 95% CI [4.12, 4.71] in the placebo and 4.34, 95% CI [4.11, 4.57] in the testosterone group. These results indicate that, on average, the participants did rate the task not only as a moderately positive experience but also as a moderately demanding and difficult task. One item of the post-task questionnaire asked whether the participants thought the task was credible or not ("Altogether, did you find the game non-credible?"). Although this item is also included in the positive scale (reverse coded), we also decided to separately report it. The mean rating of this item in the placebo group was 3.35, 95% CI [2.91, 3.79] and 3.35, 95% CI [2.95, 3.76] in the testosterone group. Again, the range was 1 (*not at all*) to 8 (*very much*). In total, 38 of the 212 participants stated explicit concerns about the reality of the other players and the setup of the task.

Additionally, we assessed how fair the participants rated the other players in the task with a separate question in the post-task questionnaire ("Do you think that your teammates have played fair?"). Figure 3 depicts the resulting mean ratings per group and risk score levels (split at the 33rd and 66th percentile). The prosocial player was rated the fairest, and the aggressive player received a very low fairness rating. Interestingly, in participants with either

a high or a low risk score, differences between the treatment groups can be seen in the ratings of the prosocial and aggressive player. By contrast, in participants with normal risk score levels, no or only much smaller differences in the ratings are visible. Through robust tests, we found that participants from the testosterone group who scored low on the risk score demonstrated a tendency to rate the prosocial player fairer than participants with a low risk score from the placebo group (T_y (30.49) = 1.85, p = .074), and they also rated the aggressive player as more unfair (T_y (28.18) = 1.71, p = .098). On the contrary, participants with a high risk score from the testosterone group did not tend to rate the prosocial player differently (T_y (32.37) = 0.95, p = .35) but rated the aggressive player fairer than participants with a high risk score from the placebo group (T_y (39.61) = 3.33, p = .002, ξ = 0.43, 95% CI [0.07, 0.76]). Further information on the post-task questionnaire and additional plots can be found in the supplement (Additional Tables and Figures A).

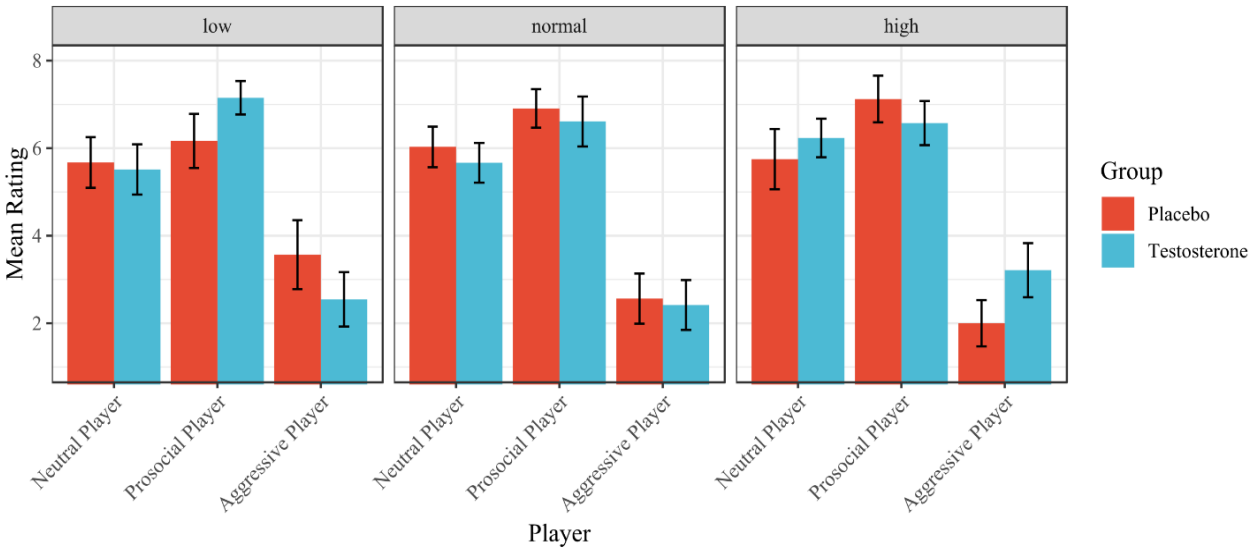


Figure 3. Mean ratings from the post-task questionnaire per group and for low (n=70), normal (n=70), and high (n=72) risk score levels on the question "Do you think that your teammates have played fair?". Scale range was 1 (not at all) to 8 (very much). Error bars indicate the 95% CI for the mean.

Asked for their belief about receiving testosterone or placebo, 159 participants (75%) indicated that they had received the placebo and 47 participants (22%) thought they had received testosterone. Six participants (3%) did not give a meaningful answer on the questionnaire. Of the 47 participants who believed to have received testosterone, roughly half (24) were correct. Fisher's exact test on the 2 × 2 contingency table resulted in a statistically non-significant result, indicating that the participants' belief about the treatment was independent from receiving placebo or testosterone, OR = 0.83, 95% CI [0.40, 1.66]. The participants did not know which gel they were receiving.

Hormone Levels

To control if the gel administration increased the testosterone concentration in the participants, we analyzed the data from the saliva samples. For the analysis reported here, data from five of six samples were available (sample 3 was still being analyzed). Inspection of the data revealed that several participants (predominantly from the testosterone group) had unnaturally high testosterone levels already at the baseline saliva sample (before the drug was applied). For example, 38 participants in the testosterone group and two in the placebo group differed from the placebo group baseline mean by more than two standard deviations (> 1,300 pg/ml). We assume that this result was most likely caused by contamination that occurred while handling the samples. For this reason, we decided to exclude participants with baseline testosterone levels higher than 700 pg/ml (cutoff based on the visual inspection of the distribution) from the analysis of the hormone levels, resulting in a dataset including 165 participants with clean baseline saliva samples. A 2 (group: placebo or testosterone) × 5 (sample: 1 to 6) linear mixed model (including a random intercept per participant to account for the repeated measures) revealed a significant interaction between group and sample, $F(4, 604.14) = 12.40$, $p < .001$. Post-hoc pairwise comparisons revealed that the two groups significantly differed in all the samples except the baseline sample (see also Figure 4).

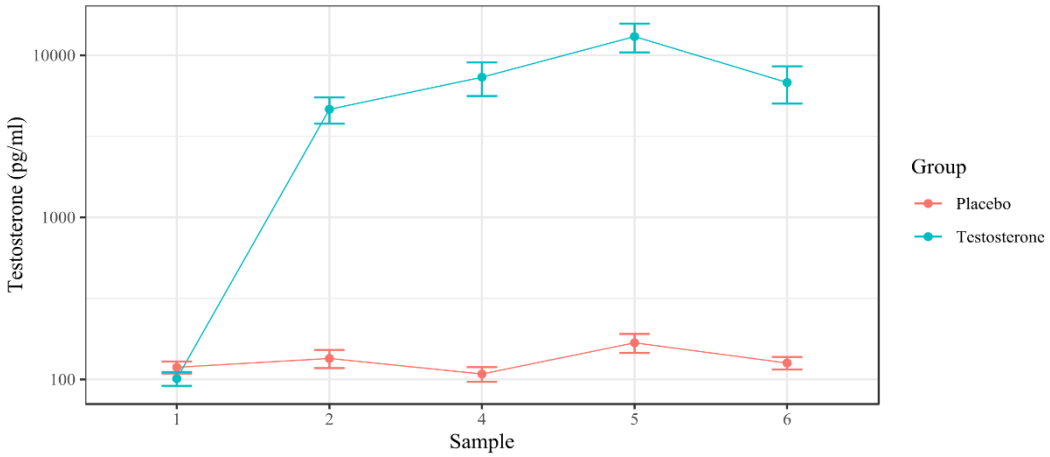


Figure 4. Resulting testosterone levels from participants with uncontaminated baseline samples (n=165). Points depict mean values per group at each sample point. Error bars indicate the standard error of the mean. The scale on the y-axis is log transformed to better display the range of values.

This result indicated that the drug administration significantly increased the testosterone levels in the participants from the treatment group and that the participants did not differ in their testosterone levels before the drug administration. We acknowledge that the contamination of some of the samples poses a threat to the validity of the hormone measurements. However, we also assessed the levels of testosterone metabolites (estradiol and dihydrotestosterone) and found no baseline differences between the placebo and

testosterone group in these measurements even if the contaminated baseline samples were included in the analysis. We therefore assume that the spurious baseline testosterone measurements were caused by contaminated saliva caps and that the participants in both groups did not systematically differ at the baseline measurement in their endogenous hormone levels. The detailed results from the post-hoc comparisons and additional analysis on the estradiol and dihydrotestosterone levels are available in the supplementary materials.

Modified Point Subtraction Aggression Paradigm

Descriptive statistics. To control if the task was able to elicit the expected behavior, we first analyzed the count of prosocial and aggressive actions taken by the participants in each condition of the task. Importantly, as one would expect if the task was able to induce the behavior in question, more points were given to the player in the prosocial condition of the task (10% trimmed $M = 8.06$, 95% CI [6.72, 9.42]) than in the neutral (10% trimmed $M = 3.31$, 95% CI [2.45, 4.18]) or aggressive condition (10% trimmed $M = 1.94$, 95% CI [1.24, 2.65]). Similarly, in the aggressive condition, more points were reduced (10% trimmed $M = 4.5$, 95% CI [3.54, 5.46]) than in the neutral (10% trimmed $M = 0.92$, 95% CI [0.68, 1.15]) or the prosocial condition (10% trimmed $M = 0.16$, 95% CI [0.00, 0.32]) of the task. This result indicates that the options to give or reduce points were not used randomly or regularly throughout the task but were rather a response to the behavior of the other player. Notably, the range of the actions taken by the participants is quite large (see Figure 5). Some participants gave numerous points to the other player regardless of the condition; moreover, some participants decided not to give or reduce points at all. In total, 20 participants (9.4%) did not give or reduce even one point throughout the entire task, 35 participants (16.5%) never chose the option to give points to the other players, and 61 participants (28.7%) never reduced the points of the other player. Additional descriptive information can be found in the supplement (Additional Descriptive Statistics).

Modeling the data. Due to the binary structure of the dependent variable and to take account of the mixed design of the study with a two-level between-subject factor (testosterone/placebo) and a three-level within-subject factor (neutral, prosocial, and aggressive condition), we used generalized linear mixed models (GLMMs) based on a logistic regression. We constructed two main model sets: one for the prosocial actions (points given) and one for the aggressive actions (points reduced). For both analyses, the data consist of 204 trials per participant (68 per condition); in each trial; a 0/1 coded variable recorded if a point was given (or reduced) by the participant. Prior to analyzing the data, we checked for trials in

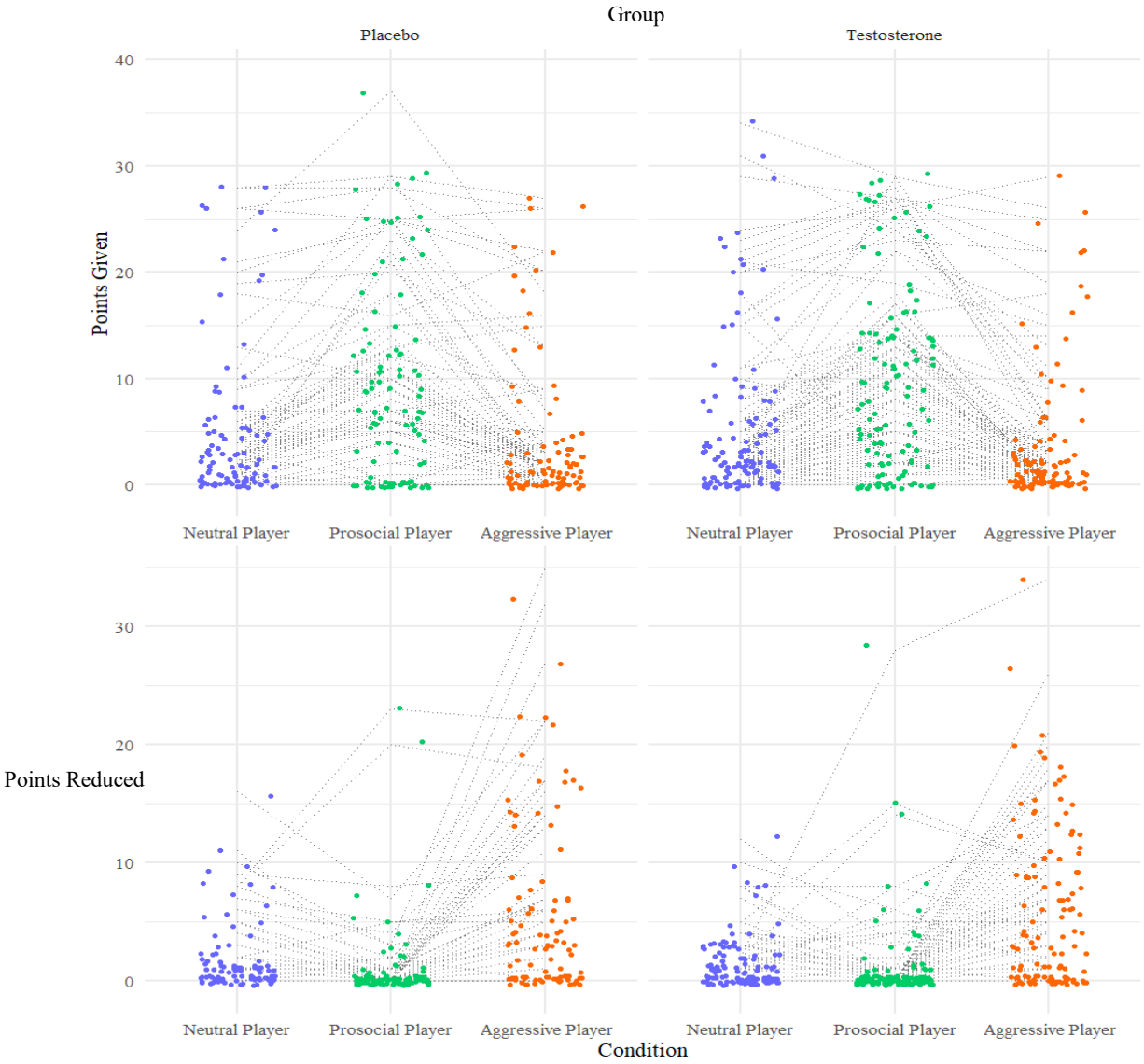


Figure 5. Plot of the actions taken by each participant. Each dot represents the sum of the points given or points reduced by one participant in one condition and the dotted lines connect the summed actions of each participant over the different conditions. The dots have been jittered to enhance the visibility of the distribution. Data from the placebo group (n=95) is dotted on the left side, from the testosterone group (n=117) on the right. The different sequences are not accounted for.

which participants pressed the buttons for simultaneously reducing and giving points. Although the participant would only get feedback on the first button pressed, the program would still record the second button press, rendering the impossibility of discerning the intended action. Both buttons were pressed in 21 trials (from a total of 43,248 recorded trials), and we excluded these trials from the analyses.

Prosocial actions. Based on a series of model comparisons that are described and reported in detail in the supplementary materials, we found that a model including the three fully crossed predictors condition (neutral, prosocial, aggressive), group (placebo, testosterone) and risk score (continuous covariate) has the lowest deviance of all the

Table 1

	Reference = Neutral Condition			Reference = Prosocial Condition			Reference = Aggressive Condition		
Predictors	**OR**	**95% CI**	***p***	**OR**	**95% CI**	***p***	**OR**	**95% CI**	***p***
(Intercept)	0.03	0.02 – 0.04	**<.001**	0.08	0.06 – 0.11	**<.001**	0.01	0.01 – 0.02	**<.001**
Neutral Condition				0.31	0.24 – 0.40	**<.001**	2.22	1.72 – 2.85	**<.001**
Prosocial Condition	3.21	2.51 – 4.11	**<.001**				5.93	4.23 – 8.33	**<.001**
Aggressive Condition	0.54	0.39 – 0.74	**<.001**	0.17	0.12 – 0.24	**<.001**			
Testosterone	1.26	0.76 – 2.11	.370	1.05	0.68 – 1.63	.810	0.88	0.47 – 1.64	.681
Risk Score	0.43	0.21 – 0.88	**.020**	0.86	0.47 – 1.57	.629	0.66	0.28 – 1.54	.339
Neutral Condition × Testosterone				1.20	0.90 – 1.59	.209	1.44	1.01 – 2.05	**.042**
Prosocial Condition × Testosterone	0.83	0.63 – 1.11	.209				1.20	0.80 – 1.80	.367
Aggressive Condition × Testosterone	0.69	0.49 – 0.99	**.043**	0.83	0.56 – 1.24	.369			
Neutral Condition × Risk Score				0.50	0.33 – 0.76	**.001**	0.65	0.39 – 1.09	.104
Prosocial Condition × Risk Score	2.01	1.31 – 3.07	**.001**				1.30	0.73 – 2.31	.365
Aggressive Condition × Risk Score	1.54	0.90 – 2.62	.112	0.77	0.42 – 1.39	.384			
Testosterone × Risk Score	1.26	0.49 – 3.25	.632	0.54	0.24 – 1.21	.134	0.60	0.20 – 1.86	.378
Neutral Condition × Testosterone × Risk Score				2.33	1.34 – 4.05	**.003**	2.09	1.05 – 4.15	**.035**
Prosocial Condition × Testosterone × Risk Score	0.43	0.25 – 0.74	**.003**				0.90	0.42 – 1.92	.781
Aggressive Condition × Testosterone × Risk Score	0.48	0.24 – 0.97	**.040**	1.11	0.50 – 2.47	.790			

Random Effects			
σ^2	3.29	3.29	3.29
τ_{00}	2.94 Participant	2.24 Participant	4.33 Participant
τ_{11}	0.44 Participant.Prosocial Condition	0.44 Participant.Neutral Condition	0.57 Participant.Neutral Condition
	0.57 Participant.Aggressive Condition	1.12 Participant.Aggressive Condition	1.12 Participant.Prosocial Condition
ρ_{01}	-0.50 Participant.Prosocial Condition	0.13 Participant.Neutral Condition	-0.62 Participant.Neutral Condition
	0.31 Participant.Aggressive Condition	0.31 Participant.Aggressive Condition	-0.73 Participant.Prosocial Condition
ICC	.49	**Observations**	43227
Marginal R^2	.102	**Conditional R^2**	.543

Note. The table shows the calculated comparisons from the model with all three task conditions as the reference level. σ^2 = within-group (participant) residual variance, τ_{00} = variation between individual intercepts and average intercept, τ_{11} = random slope variance, ρ_{01} = random-intercept-slope correlation, ICC = intraclass-correlation coefficient (adjusted for random effect variance), marginal R^2 = variance explained by fixed effects only, condition R^2 = variance explained by fixed and random effects.

compared models that included our predictors of interest and offers the best model fit. Type-III Wald chi-square tests on the fixed effects of the model indicated a statistically significant three-way Condition × Group × Risk Score interaction, χ^2 (2) = 11.12, p = .004. To clarify the nature of the interaction, we conducted a simple slope analysis (Preacher, Curran, & Bauer, 2006). In the prosocial condition, we found a negative association between the risk score and the number of points given in the testosterone group (B = -0.76, SE =0.27, z = -2.82, p = .005); however, we found no such association between the risk score and the number of points given in the placebo group (B = -0.15, SE = 0.31, z = -0.48, p = .63). This result indicated that the participants in the testosterone group who were high in dominance and low in self-controlgave less points, whereas the participants who were low in dominance and high in self-control gave more points. Meanwhile, in the placebo group, participants at all risk score levels gave approximately the same number of points to the friendly player. In the aggressive condition, we found a similar negative association between the risk score and the number of points given in the testosterone group (B = -0.92, SE = 0.27, z = -2.32, p = .02), and a weaker negative association between the risk score and the number of points given in the placebo group (B = -0.41, SE = 0.44, z = -0.94, p = .347). However, the estimates from the model (provided in Table 1) indicate a clear effect of the different conditions, denoting that in the aggressive condition, much less points were given than in the prosocial (OR = 0.17, 95% CI [0.12, 0.24], p <.001) or in the neutral condition (OR = 0.54, 95% CI [0.39, 0.74], p <.001). Therefore, the effect in the aggressive condition most likely is of little practical importance. In the neutral condition, a negative association between the risk score and the number of points given was present in both groups (placebo group: B = -0.84, SE = 0.37, z = -2.29, p = .022; testosterone group: B = -0.61, SE = 0.31, z = -1.95, p = .051), such that participants who described themselves to be dominant and low in self-control gave less points than participants low in dominance and high in self-control. We report the model estimates using treatment contrasts for the factors condition and group in Table 1. For a detailed overview, we report the outputs in Table 1 with each condition set as the reference level. For the factor group, the placebo group is always set as reference. To gain a better understanding of the model results, we also plotted the estimated effects of the model (Figure 6). In the neutral condition of the task, we see that a higher risk score is related with a lower tendency to give points; by contrast, a cross-over interaction is visible in the prosocial condition. In the prosocial condition of the task, testosterone increased the tendency to give points to the friendly other player in participants with a low risk score compared to the placebo group, whereas participants with a high risk score became less inclined to give points compared to

the placebo group. A similar cross-over interaction pattern is present in the aggressive condition (although less pronounced and on an overall lower level).

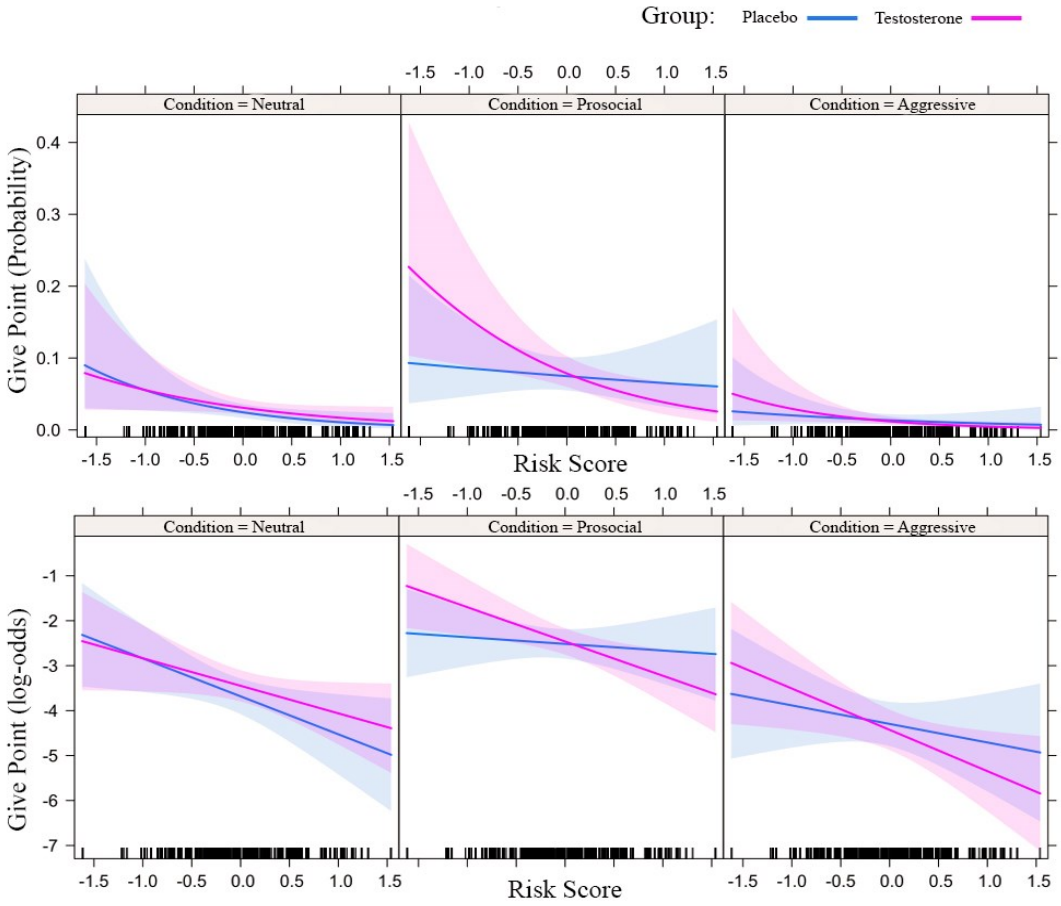


Figure 6. Plot of the estimated effects modeling the points given using GLMMs. For better visibility of the effects, the top row shows the effects plotted on the scale of the response (probability to give points), whereas the bottom row uses the link scale (log-odds) of the model.

Effects of sequence. Some participants played first with the prosocial player after the neutral player (who was the first for everyone), whereas others played first with the aggressive player after the neutral player; thus, we decided to test for the possible effects of the two sequences on the results. To perform this step, we added sequence (a 0/1 coded factor) either as a main effect or as an interaction term to the model reported above. None of these models including sequence provided a significant improvement. Including sequence in the model as a main effect (χ^2 (1) = 1.69, p = .19) or as an interaction term (χ^2 (12) = 18.51, p = .10) did not significantly improve the model. We therefore conclude that the different sequences did not have a systematic effect large enough to explain any differences in points given during the task.

Effects of mood. We found that participants who received testosterone reported to be in a better mood than the participants from the placebo group; thus, we had to control for a

possible confounding influence of the participants' mood on the results from the task. For this step, we added the standardized score of the MDBF good–bad mood scale as a covariate to the model. Including the mood score in the model either as a main effect (χ^2 (1) = 0.03, p = .86) or as an interaction term (χ^2 (12) = 16.8, p = .16) did not significantly improve the model. We can conclude that the mood difference in the treatment groups did not have a systematic effect on the participants' behavior during the task large enough to influence the results from the main analysis.

Effects of stress. We then tested if perceived stress measured with the PSS-10 does influence the participants' tendency to give points to the other player. Again, we find no indication that including the PSS-10 score improves the models. Including the standardized PSS-10 score in the model as a main effect (χ^2 (1) = 0.89 p = .34) or as an interaction term (χ^2 (12) = 11.59, p = .48) did not significantly improve the model. We conclude that perceived stress measured with the PSS-10 does not influence the relationship between testosterone and prosocial actions based on the data reported here.

Effects of points received. The number of points the participants received from the other player was randomized in the task, and some participants received more points from the friendly player in the prosocial condition than others. However, no systematic difference emerged between the placebo and testosterone group in the average number of points that the participants received (using Yuen's Test as a robust measurement: T_y (119.2) = 0.33, p = .74, 95% CI [-0.58, 0.83]). Nonetheless, we conducted an additional analysis that controls for the number of points the participants received in the prosocial condition of the task. Including the mean-centered count of points received by the participants in the model significantly improves the model fit (χ^2 (1) = 9.62, p = .002). However, there is no indication that including the amount of points received as an interaction term in the model reported significantly improves the fit (χ^2 (12) = 17.75, p = .12). The results of the model including the number of points the participants received as a covariate do not change the significance or the direction of the Condition × Risk Score × Testosterone interaction; they instead indicate that the more points the participants received from the other player, the more likely they were to give points to the other player in return, and this effect was the same in both treatment groups. Model results from the model including the number of points received for each participant as a covariate can be found in the supplementary materials.

Effects of the number of actions taken. Additionally, we decided to control for the total number of actions taken by each individual participant. To perform this step, we calculated the sum of points given and points reduced per participant throughout the entire task (across

all conditions) and then standardized the resulting variable. This measurement was then included as a covariate in the model. The model including the total amount of actions as a covariate significantly improved the overall fit (χ^2 (1) = 190.68, $p < .001$); however, compared with this model, including the covariate as an interaction term did not improve the model further (χ^2 (11) = 18.48, p =.07). The results from the model including the total amount of actions as a covariate again revealed the significant Condition × Group × Risk Score three-way interaction (χ^2 (2) = 11.99, p = .002). A table with the results of the model and a plot of the estimated effects can be found in the supplementary materials. Compared to the result from the main analysis reported above, we now find a significant Group × Risk Score interaction in the prosocial condition, such that the difference between the placebo and testosterone group significantly varied across the risk score of the participants (OR = 0.45, 95% CI [0.25, 0.79], p = .006); that is, among the participants with a low risk score, the participants in the testosterone group gave more points than the participants who received a placebo. However, with increasing risk score, this pattern inversed. In participants with a high risk score, the testosterone group gave less points than participants from the placebo group. Using the neutral condition as reference for the comparisons, we again find the significant Prosocial Condition × Testosterone × Risk Score contrast (OR = 0.39, 95% CI [0.23, 0.68], p = .001) that indicates that the treatment groups in the prosocial condition significantly differed in their tendency to give points, depending on the risk score when compared with the neutral condition. However, we no longer find any indication for such a difference between the treatment groups in the aggressive condition (OR = 0.55, 95% CI [0.28, 1.08], p = .08). Taken together, the results from the model including the count of actions throughout the entire task as a covariate clarify the results from the main analysis. We found a more pronounced effect in the prosocial condition of the task and no longer an indication for a similar effect in the aggressive condition.

Aggressive actions. To analyze the aggressive actions (points reduced) of the participants, we followed the same procedure as in the previous analysis of the prosocial actions. First, using generalized linear mixed models, we built and compared a series of models (reported in detail in the supplementary materials). The most complex model including all three fully crossed predictors of interest (Condition × Group × Risk Score) did not prove to be significantly better than a simpler model including only condition as a predictor. We therefore find no evidence for a significant testosterone effect or interaction, and no significant effect or interaction of the risk score regarding the aggressive actions.

Exploratory analysis. To provide additional information for the interpretation of the results, we decided to explore the data and describe now several aspects of the measured behavioral responses that might aid the understanding of the study results. An important point to remember is that these additional analyses were only conducted to illustrate alternative perspectives on the data and offer additional approaches to discuss the results of the main analyses. We did not run any statistical tests because we did not specify any a priori hypothesis concerning these measurements. Therefore, this section is purely descriptive, and the presented results have to be interpreted with caution.

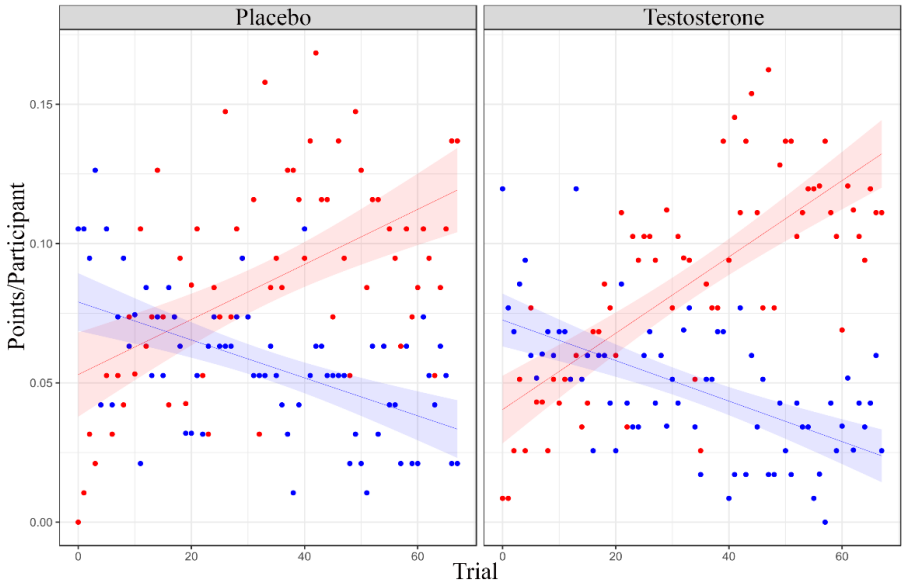


Figure 7. Average number of points given in each trial (in blue) and average number of points reduced in each trial (in red) during the aggressive condition of the task.

Chronological distribution of actions. First, we were interested in the chronological distribution of the participants' actions (points taken or reduced) especially during the aggressive condition. Many participants gave points to the aggressive player instead of retaliating, and we consider that the participants probably adapted their strategy while they were confronted with the aggressive player. We initially calculated the average number of points given and points reduced by the participants in each of the 68 trials of the aggressive condition (sum of points given or taken per trial/number of participants) and separately plotted the values for each group. Figure 7 illustrates a gradual increase in aggressive actions and a decrease in prosocial actions during the course of the aggressive condition – when the participants were provoked by point subtractions. During the first half of the aggressive condition, a mixed behavioral response to the provocations can be seen with roughly the same number of points given and points reduced. In the second half, the number of points reduced is higher than the number of points given. No differences between the placebo and

testosterone group are visible in the plot. In the prosocial condition, during which the participants received points from the other player, we can see a slight increase in points given with no obvious group differences. The corresponding plot for the prosocial condition can be found in the supplement (Additional Tables and Figures).

Reactive vs. proactive actions. As a second exploratory approach, we decided to investigate if participants exhibited a more reactive or proactive type of behavioral response. To perform this step, we generated additional variables that registered if each point given from the participant to the other player was given before or after the participant received a point from the other player. Exactly the same procedure was conducted for the point reductions. We then calculated the average number of points given before a point was received and after a point was received for each trial and plotted the resulting data to visually inspect the results (see supplement: Additional Tables and Figures). We performed the same procedure for the point reductions (see supplement: Additional Tables and Figures). However, we did not conduct any statistical tests because we did not specify any a priori hypothesis regarding these variables, and the data are split into subsets with a reduced number of participants limiting the power of any test. The plots revealed no indication for any meaningful differences between the groups.

Reciprocity. To assess the extent to which the participants reciprocated the behavior of the other players, we calculated and plotted (see supplement: Additional Tables and Figures) the difference between the points they gave and received and the points they reduced and were taken from them. The plots demonstrate that, on average, the participants did give less points in the prosocial condition than they received and also reduced less points than the other player in the aggressive condition. In the prosocial condition, participants with a lower risk score apparently reciprocated slightly more than those with a higher risk score; furthermore, small differences between the placebo and the testosterone group can be seen at low and high levels of the risk score. In the aggressive condition, participants who scored high on the risk score seem to retaliate the point reductions more often in the placebo group; on the contrary, in the testosterone group, the number of point reductions remains constant regardless of the risk score level.

Difference measure. During the entire task, the participants were able to use both options (give or reduce points), and we already have seen that many participants indeed made use of both options in one condition (especially in the aggressive condition). To gain a better understanding of the relation between giving and taking points, we calculated the difference of points given and points reduced per participant in the prosocial and aggressive conditions

and plotted the resulting data (see supplement: Additional Tables and Figures I). In the prosocial condition, points were predominantly given (positive difference), but we find the expected ambivalent result that most participants gave and reduced points roughly to the same extent in the aggressive condition (difference close to zero). Again, we find that higher risk score levels in the placebo group indicate more point reductions (a greater negative difference); by contrast, the distribution of the difference measurement in the testosterone group appears fairly similar regardless of the risk score level.

Discussion

The results of this study indicate that testosterone influences prosocial behavior in men and that personality traits (dominance and self-control) moderate this relationship. We found that the behavioral responses of participants who received testosterone differed from those who received a placebo when they were confronted with a friendly player who gave them points with monetary value in a modified Point Subtraction Aggression Paradigm. In participants with low dominance or impulsivity, testosterone increased the willingness to give points to a friendly other player at a cost to themselves, whereas in participants with highly dominant or impulsive personalities, testosterone decreased the willingness to give points. This finding supports, at least partially, our first hypothesis that testosterone increases prosocial behavior and goes well in line with previous research indicating that testosterone can influence prosocial behavior (Boksem et al., 2013; Diekhof, Wittmer, & Reimers, 2014; Dreher et al., 2016; Eisenegger et al., 2010). Regarding our second hypothesis that testosterone increases the aggressive reactions of men to provocation, no supporting evidence was found. Although previous research has established a clear association between testosterone and aggressive reactions to provocations (Carré et al., 2017; Dreher et al., 2016; Geniole et al., 2019; Wagels et al., 2018), we did not find a testosterone effect when analyzing the aggressive option to reduce points from the other player in any condition of the task. Nonetheless, we found that the rate of point reductions increased when the participants interacted with the aggressive player compared with the other two players. Two questions now arise: (a) Why did we fail to replicate a well-established effect? (b) Why did testosterone cause a difference in prosocial behavior only in men who scored high or low on the risk score?

In our opinion, two main reasons underlie the null-result regarding the aggressive reactions to the point subtractions. The first reason pertains to the availability of the friendly option to give points to the other player. This option is not available in the standard PSAP

(Geniole, MacDonell et al., 2017), and many participants decided to use this non-aggressive option to give points to the aggressive player instead of retaliating by reducing the aggressive player's points. Breaking social norms (not harming another person on purpose), even as a response to provocation, is not considered acceptable in most societies (Gelfand et al., 2011); thus, the participants might have resorted to the aggressive option only after they had realized that the other player was determined to take away their points and that his intentions could not be altered by giving him points. Our exploratory analysis provided some support for this assumption. We found that during the first half of the aggressive condition, a mixed pattern of behavioral responses was recorded with roughly the same number of points given and taken; the number of point reductions by the participants increased and surpassed the number of points given only in the second half of the aggressive condition.

Second, the duration of the aggressive condition is likely to be too short to register any meaningful differences between the groups because of this familiarization process with the other player's attitude. After all, the participants were engaged approximately 10 minutes with the entire task, resulting in a duration of a little more than three minutes per condition. However, the duration of the standard PSAP usually lasts around 10 minutes (Carré et al., 2017). If the duration of the aggressive condition would have been longer, the aggressive reactions might have been more pronounced. This inference raises the question of whether the effects in the prosocial condition could also be enhanced by a longer duration; similarly in the prosocial condition, we must assume that the participants first had to comprehend the friendly intentions of the other player before settling on their strategy to respond. In our exploratory analysis, we found a slight increase in points given from the first to the last trial of the prosocial condition, thereby providing some support for this idea. We can conclude that, especially in the aggressive condition, a longer duration of the task would provide the opportunity for a more precise measurement of the participants' reactions. Therefore, we propose that future studies consider this temporal aspect when adapting a comparable design.

Moreover, we consider an alternative, hypothetical explanation for the null-result regarding the aggressive reactions. We probably observed an effect of testosterone only on prosocial but not on aggressive reactions, as long-term (genomic) testosterone effects might differ from short-term (non-genomic) testosterone effects in their behavioral consequences. Due to its lipophilic qualities, testosterone can pass the extracellular membrane of the cell and can bind to, and activate, intracellular androgen receptors, thereby modifying the gene expression and protein synthesis in the cell (Foradori, Weiser, & Handa, 2008). This genomic pathway of testosterone effects takes time and is assumed to peak several hours after

testosterone exposure. By contrast, the non-genomic effects of testosterone operate much faster by modifying intracellular calcium levels or activating second-messenger pathways (Foradori et al., 2008). Steroids such as testosterone are also able to pass the blood-brain barrier, and they can directly influence modulatory neurotransmitter systems. These effects are also described as neuroactive effects (Höfer, Lanzenberger, & Kasper, 2013). Geniole et al. (2019) discussed the possibility that studies involving longer time periods between testosterone application and behavioral measurements (> 4 h) tend to find effects of testosterone also on prosocial actions, whereas studies using shorter time periods predominantly find an association between testosterone and aggressive reactions. However, the genomic and non-genomic effects of steroids interact (Wilkenfeld, Lin, & Frigo, 2018), and how each pathway contributes to human behavior is far from clear. Nonetheless, the different functional mechanisms of testosterone could provide an alternative, although highly speculative, explanation of why our study failed to replicate the effect of testosterone on aggressive reactions to provocation. In our study, the participants completed the modified PSAP almost four hours after the testosterone administration; the long-term genomic effects probably attenuated the association between testosterone and aggressive behavior that we attempted to capture with the aggressive condition of the task.

In the prosocial condition of the task – where the participants were confronted with a friendly player who gave them points – testosterone caused a difference in the number of points participants gave to the friendly player depending on the risk score (a composite measurement of personality traits including dominance and self-control). Geniole et al. (2019) already discussed the possibility that this risk score is associated with the effects of testosterone on prosocial behavior, such that individuals with a lower risk score demonstrate increased prosocial behavior when treated with testosterone. We found exactly this association and observed a reduced tendency to give points in men who scored high on the risk score. Geniole et al. (2019) explained the connection between testosterone, personality traits included in the risk score, and aggressive behavior by a mediating effect of the dopaminergic system. They found that in males who scored high on the risk score, more aggressive actions were associated with increased feelings of reward when treated with testosterone. Following this idea proposed by Geniole et al. (2019), we now discuss if non-genomic neuroactive testosterone effects (e.g., mediating effects of modulatory neurotransmitter systems) could also provide an explanation for the observed effects of testosterone on prosocial behavior.

Although we did not directly measure the pleasure or reward associated by the participants with either giving points to the other player or reducing points from the other player, we did assess how fair the participants rated the other players' behavior. We found indications for testosterone-induced differences in the fairness ratings from participants with low or high risk score levels. Furthermore, we found weak (mostly insignificant) evidence that in participants with low risk score levels, testosterone increased the fairness rating of the friendly player and decreased the fairness rating of the aggressive player compared with participants from the placebo group. In participants with high risk score levels, the data indicated a reversed pattern (i.e., decreased fairness rating of the friendly but increased fairness rating of the aggressive player compared to the placebo group). Perhaps the differences in the fairness ratings are associated with how rewarding the participants experienced the interactions with the other players; studies have found that fair behavior activates brain regions associated with reward, which indicates that experiencing fair behavior is rewarding (Gradin et al., 2015; Tabibnia & Lieberman, 2007; Tabibnia, Satpute, & Lieberman, 2008). How rewarding a certain type of behavior (or interacting with another person behaving in such a way) was perceived by the participants could then explain why testosterone caused an effect on prosocial behavior depending on the risk score if this association is mediated by the dopamine system depending on the personality traits included in the risk score and if testosterone also affects the dopamine system.

Indeed, studies have revealed that testosterone administration influences the activity in brain areas associated with the dopamine system in rodents and humans. Aubele and Kritzer (2012) found that in male rats, the connections between the prefrontal cortex and the ventral tegmental area are enriched with androgen receptors and that cell firing and dopamine levels in these areas are modulated by these connections and hence by testosterone as well. In humans, Hermans et al. (2010) were able to show that testosterone administration influences the activity in the ventral striatum, a brain area associated with dopamine-related reward processing, in response to cues signaling monetary reward. Additionally, studies have found the personality traits included in the risk score to be associated with the dopamine system. Trait impulsivity (self-control) was found to negatively correlate with dopamine autoreceptor availability in the substantia nigra and the ventral tegmental area and positively correlate with dopamine release in the striatum (Buckholtz et al., 2010). A review by Qu, Ligneul, van der Henst, and Dreher (2017) shows how several studies linked the dopaminergic system to dominance behavior, indicating that dopamine plays a role in status-relevant behaviors and is linked to trait dominance. Finally, a causal link between dopamine and prosocial behavior in

humans was established. Sáez, Zhu, Set, Kayser, and Hsu (2015) found that experimentally increasing dopamine levels promoted egalitarian, inequity-avoidant behavior. Taken together, these results suggest that dopamine is associated with prosocial behavior, that personality traits such as dominance and self-control can influence this relationship, and that testosterone influences the dopamine system. Assuming that the fairness rating and the rewarding experience from interacting with the other players in the task are related, a mediating effect from the dopamine system based on the associations described above could provide one possible explanation for the observed results of our study.

In addition to this proposed link between testosterone and dopamine-regulated reward systems that offer one possible explanation for the observed effects of testosterone on prosocial behavior, other neurotransmitter systems might play a role as well. Specifically, experimental testosterone administration might induce non-genomic effects on the serotonergic system, causing differences in prosocial behavior depending on the risk score. This outcome seems possible because studies have revealed that both personality traits included in the risk score (dominance and self-control) are associated with the serotonergic system. In both humans and primates, pronounced dominance behavior is associated with increased serotonin levels (Qu et al., 2017; Weinberg-Wolf & Chang, 2019), and certain genetic polymorphisms of the serotonin receptor gene are associated with high levels of impulsivity (Tomson, Vaht, Laas, Veidebaum, & Harro, 2016). Furthermore, several studies have established a functional link between prosocial behavior and serotonin. Low serotonin levels in humans are associated with increased rejection of unfair offers in the ultimatum game (Crockett, Clark, Tabibnia, Lieberman, & Robbins, 2008), which can be seen as prosocial behavior in the form of altruistic punishment (i.e., enforcing the social norm of not being unfair at a cost to oneself, Crockett, 2009). In a similar study, Crockett, Clark, Lieberman, Tabibnia, and Robbins (2010) revealed that experimentally inducing lower levels of serotonin caused an increase in altruistic punishment in the ultimatum game and also in impulsive behavior (reduced self-control). Additionally, a functional link between testosterone and serotonin is well established (Höfer et al., 2013), and in human males, high testosterone levels negatively correlate with serotonin levels (Perfalk et al., 2017). Therefore, similar to the mediating dopamine mechanism proposed by Geniole et al. (2019), testosterone might exert its effects on human behavior by modulating the serotonin system, which in turn could explain why testosterone influences prosocial behavior only in males with specific personality traits.

To summarize, two possible non-genomic pathways could link testosterone and prosocial behavior in a way that could account for differences based on personality traits included in the risk score; a mediating effect of the dopamine or the serotonin system. However, the experimental evidence for this proposed functional associations between testosterone, dopamine, and possibly serotonin remains week. Many studies investigating involved components (e.g., the link between personality traits and neurotransmitter systems, and many imaging studies) suffer from low powered study designs due to small sample sizes, and several studies provide only correlational evidence. Therefore, these issues need to be further clarified: if the effect of testosterone is truly mediated by the serotonergic or dopaminergic system, and if an interaction between these mediating mechanisms exists, which seems likely (Avery & Krichmar, 2017). These issues could be addressed by studies experimentally manipulating hormone levels and simultaneously controlling for the activity of mediating neurotransmitter systems. Specifically, empirical evidence for the presumed direction of the three-way association between certain personality traits, a mediating neuromodulatory system, and the behavior in question (e.g., that high trait level dominance and reduced prosocial behavior are correlated with activity in the dopamine system and that this association can be influenced by administering testosterone) is scarce. In our opinion, for now the explanatory power from this assumed mediating influence of neuromodulatory systems remains limited, as evidence is still largely preliminary and certainly requires further investigations.

Several limitations of the study have to be mentioned. We unexpectedly found that the results from the mood questionnaire indicated a significant group difference in positive mood states. The question of why the testosterone group reported to feel better and if this mood state has any implications for the results of the task needs to be discussed. The participants completed the mood questionnaire after the two-hour waiting period. This means that the testosterone administration already caused an increase in testosterone levels, and this factor might be the cause for the mood difference between the groups. After all, previous studies reported increased positive mood states in hypogonadal men during a long-term testosterone treatment (Wang et al., 2004), or even an increase in maniac symptoms when administering supraphysiological doses of testosterone (Pope, Kouri, & Hudson, 2000), providing some evidence that exogenous testosterone can increase positive mood states during long-term use. However, other studies do not report positive mood changes in long-term testosterone administration studies in healthy eugonadal men (O'Connor, Archer, Hair, & Wu, 2002), and some studies even suggest a positive association of higher salivary testosterone levels with

negative mood states such as anger and tension (van Honk et al., 1999). The issue of whether testosterone can cause short-term elevated positive mood states in young, eugonadal men also remains unclear. Here, we found increased positive mood states in young men who received testosterone and conclude that exogenous testosterone can cause elevated positive mood states relatively quickly. However, we think that this difference in mood assessed before the participants began with the tasks is unlikely to have a significant impact on the results from the modified PSAP simply because approximately 90 minutes passed, and a variety of different tasks had transpired before the start of the modified PSAP. This aspect certainly affected the mood states of the participants depending, for example, on their perceived success in the tasks or due to fatigue; the baseline mood difference remaining stable until and during the modified PSAP is therefore unlikely. In the end, our analysis also provided no indication for any mood-related effects on the measured behavior.

Similar to this baseline difference in mood, the analysis also revealed a baseline difference in the calculated, composite personality measurement (risk score). No differences between the groups were present in any measurement included (dominance, prestige and impulsivity), but the testosterone group scored higher in the resulting risk score. We performed a median split on the risk score variable and compared the resulting means in both groups separately. For participants with low risk score values, no significant difference between the groups was found, $T_y\,(63.14) = 0.86$, $p = .39$. For participants with high risk score values, the result of the robust test revealed an almost significant difference, $T_y\,(61.26) = 1.96$, $p = .055$, indicating that the testosterone group scored slightly higher (0.1 points on the standardized variable) on the risk score than the placebo group. Although unexpected, this outcome likely does not pose a threat to the validity of the results. First, the difference is small and therefore unlikely to cause a large bias, such that the effect we found is due only to this baseline difference in the risk score. Second, the personality measurements were taken immediately after the testosterone application at the beginning of the waiting period; therefore, this slight difference seems highly unlikely to be due to the testosterone administration (which would pose a much greater threat to the results). Therefore, this slight difference seems likely to be just a random effect with little to no significance.

The possible carry-over effects from the other tasks in the study constitute another threat to the reliability and validity of the results presented here. The participants completed four different tasks before they played the modified PSAP, and some of the tasks included winning and losing or explicit social status manipulations. All of these experimental paradigms are well known to rapidly influence testosterone levels and affect subsequent

behavior (Geniole, Bird, Ruddick, & Carré, 2017; Hamilton, Carré, Mehta, Olmstead, & Whitaker, 2015). This could have caused systematic influences on the results from the modified PSAP. For example, we cannot rule out the possibility that certain participants experienced a specific constellation of events in the precedent tasks that might have strongly influenced the behavior exhibited in the modified PSAP. However, due to the high number of possible combinations of events in the previous tasks, we cannot control for the task experiences that each participant had prior to the modified PSAP and assume that the randomization of the different task conditions prevented any strong systematic bias.

Some aspects regarding our analysis could be criticized, and we address the most obvious issues here. First, we separately focused on each response (giving points or reducing points) and conducted two separate analyses for each type of behavior (prosocial and aggressive). This solution might not be optimal because both response options were available throughout the entire task and, especially in the aggressive condition, both options were used to some extent. Therefore, statistical methods that are able to simultaneously model both responses could help to further clarify the observed behavior and might be able to disentangle the mixed responses in the aggressive condition. However, such methods have to be able to not only deal with multiple response variables but also provide the option to account for the mixed design of the task (within and between subject factors) and accordingly model the random effect structure. For now, the correct application and interpretation of such methods (e.g., Bayesian generalized linear mixed models, see Hadfield; 2010) is beyond the skills of the authors. Second, we did not report standardized effect sizes for our results from the mixed-effect models. The reason is that the standardized effect sizes for mixed-effect models are not easily available; moreover, the best means of computing such measurements, and if this step even makes sense, is still debated (Pek & Flora, 2018; Rights & Sterba, 2019).

The task used in this study was based on a well-established experimental paradigm (the PSAP) but was considerably modified to allow for the different conditions of the task and the various options (give and reduce points), and this merits a discussion of the advantages and disadvantages of the task itself. We initially were worried about the believability of the cover story (playing with participants from another study with which they were connected over the internet), but we found that only a few (38) participants complained to some extent about the believability of the task and that the rating on the unbelievability scale of the post-task questionnaire was relatively low (and equally so in both groups). Additionally, most participants responded to the behavior of the other players, and only a few (20) participants did not use the option to give or reduce points. Taken together, these results indicate that the

majority of the participants believed they were playing with real people and acted accordingly.

Although the participants never received any feedback on the performance of the other player, we noticed that a few participants wrote down in the post-task questionnaire that they were interested in the number of points that the other player gained or stated that they were actively focusing on playing better than their adversary. This indicates that at least a few participants perceived the task as a competitive situation. We propose that future studies emphasize in the instructions that such a task is not a competitive situation to prevent any unwanted effects due to the participants' motivation to win, which is strongly influenced by testosterone (Geniole, Bird et al., 2017; Losecaat Vermeer, Riečanský, & Eisenegger, 2016). Importantly, this unlikely threatens the results from this study because only a few participants reported this motivational aspect.

In the post-task questionnaire, the participants indicated that playing the task was a moderately funny and demanding experience. This is important for two reasons. The task was not overly difficult; therefore, frustration is unlikely to be the main motive for any actions taken during the task, and simultaneously gaining points was not too easy. Hence, each point gained should have a subjective value to the participants, and the action of giving or reducing a point should therefore be a meaningful act on purpose. The participants gained on average 122 points in the task, corresponding to approximately 40 points per condition. The number of points the participants received from the prosocial player and the number of point reductions the participants experienced were randomized, and slight (albeit insignificant) variations emerged between the groups. This might have caused additional bias in the data if, for example a certain minimum number of points must be received in the prosocial condition (or reduced in the aggressive condition) before the participant realized the intentions of the other player; additionally, some participants might have experienced an insufficient number of actions of the other player to trigger the relevant behavior. We would therefore advise to keep the number of point reductions and points received constant across participants to increase comparability in future studies.

To summarize, this study contributes to the ongoing discussion on the behavioral effects of testosterone on human social behavior. We aimed to clarify the relationship between testosterone, aggressive and prosocial behavior, and related personality traits. In a large sample of healthy men, using a modified Point Subtraction Aggression Paradigm, we found evidence for a testosterone effect on prosocial behavior. To date, this work is the first testosterone administration study that established a moderating link between specific

personality traits and prosocial behavior. Trait dominance and impulsivity influenced the effect of testosterone on prosocial behavior, such that an effect of testosterone was only present in men with high or low dominant or impulsive traits. By demonstrating this moderating effect, the results promote our understanding of the effects of testosterone on prosocial behavior. As the explanatory power of our study is evidently limited to men, future research will have to investigate if a similar link between trait dominance or impulsivity and prosocial behavior exists in women.

Connecting to ideas proposed in recent work from Geniole et al. (2019), we discussed hypothetical neurophysiological pathways that could help to clarify how testosterone influences human behavior and why certain personality traits in men moderate this association. However, additional experimental evidence is needed to clarify the physiological mechanisms involved in how personality traits moderate the effect of testosterone on prosocial behavior. In our study, we found indications that testosterone affected not only the participants' behavior but also their appraisal of the behavior of the three other players. Future studies should consider measuring the participants' opinion about their own and other people's behavior in interactive experimental paradigms to a greater extent. Such research theme could aid our understanding of the psychological and functional mechanisms involved in the complex inter-linked network of hormonal functions, personality traits, different neurophysiological pathways, and social behavior.

References

Allee, W. C., Collias, N. E., & Lutherman, C. Z. (1939). Modification of the social order in flocks of hens by the injection of testosterone propionate. *Physiological Zoology*, *12*, 412–440. https://doi.org/10.1086/physzool.12.4.30151519

Almeida, R. M. M. de, Cabral, J. C. C., & Narvaes, R. (2015). Behavioural, hormonal and neurobiological mechanisms of aggressive behaviour in human and nonhuman primates. *Physiology & Behavior*, *143*, 121–135. https://doi.org/10.1016/j.physbeh.2015.02.053

Anderson, C., & Kilduff, G. J. (2009). The pursuit of status in social groups. *Current Directions in Psychological Science*, *18*, 295–298. https://doi.org/10.1111/j.1467-8721.2009.01655.x

Archer, J. (1988). *Cambridge studies in behavioural biology: The behavioural biology of aggression*. New York, NY: Cambridge University Press.

Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience and Biobehavioral Reviews*, *30*, 319–345. https://doi.org/10.1016/j.neubiorev.2004.12.007

Archer, J., Graham-Kevan, N., & Davies, M. (2005). Testosterone and aggression: A reanalysis of Book, Starzyk, and Quinsey's (2001) study. *Aggression and Violent Behavior*, *10*, 241–261. https://doi.org/10.1016/j.avb.2004.01.001

Aubele, T., & Kritzer, M. F. (2012). Androgen influence on prefrontal dopamine systems in adult male rats: Localization of cognate intracellular receptors in medial prefrontal projections to the ventral tegmental area and effects of gonadectomy and hormone replacement on glutamate-stimulated extracellular dopamine level. *Cerebral Cortex*, *22*, 1799–1812. https://doi.org/10.1093/cercor/bhr258

Avery, M. C., & Krichmar, J. L. (2017). Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits*, *11*, 108. https://doi.org/10.3389/fncir.2017.00108

Boksem, M. A. S., Mehta, P. H., van den Bergh, B., van Son, V., Trautmann, S. T., Roelofs, K., . . . Sanfey, A. G. (2013). Testosterone inhibits trust but promotes reciprocity. *Psychological Science*, *24*, 2306–2314. https://doi.org/10.1177/0956797613495063

Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., . . . Zald, D. H. (2010). Dopaminergic network differences in human impulsivity. *Science*, *329*, 532. https://doi.org/10.1126/science.1185778

Carré, J. M., & Archer, J. (2018). Testosterone and human behavior: The role of individual and contextual variables. *Current Opinion in Psychology*, *19*, 149–153. https://doi.org/10.1016/j.copsyc.2017.03.021

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. https://doi.org/10.1016/j.biopsych.2016.06.009

Cheng, J. T., Tracy, J. L., & Henrich, J. (2010). Pride, personality, and the evolutionary foundations of human social status. *Evolution and Human Behavior*, *31*, 334–347. https://doi.org/10.1016/j.evolhumbehav.2010.02.004

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*, 385–396. https://doi.org/10.2307/2136404

Crockett, M. J. (2009). The neurochemistry of fairness: Clarifying the link between serotonin and prosocial behavior. *Annals of the New York Academy of Sciences*, *1167*, 76–86. https://doi.org/10.1111/j.1749-6632.2009.04506.x

Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion*, *10*, 855–862. https://doi.org/10.1037/a0019861

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, *320*, 1739. https://doi.org/10.1126/science.1155577

Dabbs, J. M., Carr, T. S. [Timothy], Frady, R. L. [Robert], & Riad, J. K. (1995). Testosterone, crime, and misbehavior among 692 male prison inmates. *Personality and Individual Differences*, *18*, 627–633. https://doi.org/10.1016/0191-8869(94)00177-T

Dabbs, J. M., Frady, R. L., Carr, T. S., & Besch, N. F. (1987). Saliva testosterone and criminal violence in young adult prison inmates. *Psychosomatic Medicine*, *49*, 174–182. https://doi.org/10.1097/00006842-198703000-00007

Diekhof, E. K., Wittmer, S., & Reimers, L. (2014). Does competition really bring out the worst? Testosterone, social distance and inter-male competition shape parochial altruism in human males. *PLoS ONE*, *9*(7), e98977. https://doi.org/10.1371/journal.pone.0098977

Dreher, J.-C., Dunne, S., Pazderska, A., Frodl, T., Nolan, J. J., & O'Doherty, J. P. (2016). Testosterone causes both prosocial and antisocial status-enhancing behaviors in

human males. *Proceedings of the National Academy of Sciences*, *113*, 11633–11638. https://doi.org/10.1073/pnas.1608085113

Edwards, D. A. (1969). Early androgen stimulation and aggressive behavior in male and female mice. *Physiology & Behavior*, *4*, 333–338. https://doi.org/10.1016/0031-9384(69)90185-1

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature*, *463*, 356–359. https://doi.org/10.1038/nature08711

Foradori, C. D., Weiser, M. J., & Handa, R. J. (2008). Non-genomic actions of androgens. *Frontiers in Neuroendocrinology*, *29*, 169–181. https://doi.org/10.1016/j.yfrne.2007.10.005

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., . . . Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*, 1100–1104. https://doi.org/10.1126/science.1197754

Geniole, S. N., Bird, B. M., Ruddick, E. L., & Carré, J. M. (2017). Effects of competition outcome on testosterone concentrations in humans: An updated meta-analysis. *Hormones and Behavior*, *92*, 37–50. https://doi.org/10.1016/j.yhbeh.2016.10.002

Geniole, S. N., & Carré, J. M. (2018). Human social neuroendocrinology: Review of the rapid effects of testosterone. *Hormones and Behavior*, *104*, 192–205. https://doi.org/10.1016/j.yhbeh.2018.06.001

Geniole, S. N., MacDonell, E. T., & McCormick, C. M. (2017). The Point Subtraction Aggression Paradigm as a laboratory tool for investigating the neuroendocrinology of aggression and competition. *Hormones and Behavior*, *92*, 103–116. https://doi.org/10.1016/j.yhbeh.2016.04.006

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. https://doi.org/10.1177/0956797619826970

Gradin, V. B., Pérez, A., MacFarlane, J. A., Cavin, I., Waiter, G., Engelmann, J., . . . Steele, J. D. (2015). Abnormal brain responses to social fairness in depression: An fMRI study using the Ultimatum Game. *Psychological Medicine*, *45*, 1241–1251. https://doi.org/10.1017/S0033291714002347

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*, 1–22. Retrieved from http://www.jstatsoft.org/v33/i02/

Hamilton, L. D., Carré, J. M., Mehta, P. H., Olmstead, N., & Whitaker, J. D. (2015). Social neuroendocrinology of status: A review and future directions. *Adaptive Human Behavior and Physiology*, *1*, 202–230. https://doi.org/10.1007/s40750-015-0025-5

Hartig, F. (2009). DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.2.4. Retrieved from https://CRAN.R-project.org/package=DHARMa

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*, 277–283. https://doi.org/10.1016/j.neuroimage.2010.04.019

Höfer, P., Lanzenberger, R., & Kasper, S. (2013). Testosterone in the brain: Neuroimaging findings and the potential role for neuropsychopharmacology. *European Neuropsychopharmacology*, *23*, 79–88. https://doi.org/10.1016/j.euroneuro.2012.04.013

Losecaat Vermeer, A. B., Riečanský, I., & Eisenegger, C. (2016). Competition, testosterone, and adult neurobehavioral plasticity. *Progress in Brain Research*, *229*, 213–238. https://doi.org/10.1016/bs.pbr.2016.05.004

Mair, P., & Wilcox, R. (2018). Robust statistical methods using the WRS2 package. Retrieved from https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.pdf

Mazur, A., & Booth, A. (1998). Testosterone and dominance in men. *Behavioral and Brain Sciences*, *21*, 353–363.

Meule, A., Vögele, C., & Kübler, A. (2011). Psychometrische Evaluation der deutschen Barratt Impulsiveness Scale – Kurzversion (BIS-15) [Psychometric evaluation of the German Barratt Impulsiveness Scale - short version (BIS-15)]. *Diagnostica*, *57*, 126–133. https://doi.org/10.1026/0012-1924/a000042

O'Connor, D. B., Archer, J., Hair, W. M., & Wu, F. C. W. (2002). Exogenous testosterone, aggression, and mood in eugonadal and hypogonadal men. *Physiology & Behavior*, *75*, 557–566. https://doi.org/10.1016/S0031-9384(02)00647-9

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*, 208–225. https://doi.org/10.1037/met0000126

Perfalk, E., Cunha-Bang, S. d., Holst, K. K., Keller, S., Svarer, C., Knudsen, G. M., & Frokjaer, V. G. (2017). Testosterone levels in healthy men correlate negatively with serotonin 4 receptor binding. *Psychoneuroendocrinology*, *81*, 22–28. https://doi.org/10.1016/j.psyneuen.2017.03.018

Pope, H. G., Kouri, E. M., & Hudson, J. I. (2000). Effects of supraphysiologic doses of testosterone on mood and aggression in normal men. *Archives of General Psychiatry*, *57*, 133. https://doi.org/10.1001/archpsyc.57.2.133

Prasad, S., Narayanan, J., Lim, V. K. G., Koh, G. C. H., Koh, D. S. Q., & Mehta, P. H. (2017). Preliminary evidence that acute stress moderates basal testosterone's association with retaliatory behavior. *Hormones and Behavior*, *92*, 128–140. https://doi.org/10.1016/j.yhbeh.2016.10.020

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448. https://doi.org/10.3102/10769986031004437

Qu, C., Ligneul, R., van der Henst, J.-B., & Dreher, J.-C. (2017). An integrative interdisciplinary perspective on social dominance hierarchies. *Trends in Cognitive Sciences*, *21*, 893–908. https://doi.org/10.1016/j.tics.2017.08.004

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*, 309–338. https://doi.org/10.1037/met0000184

Sáez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology*, *25*, 912–919. https://doi.org/10.1016/j.cub.2015.01.071

Steyer, R. (1997). *Der mehrdimensionale Befindlichkeitsfragebogen (MDBF): Handanweisung* [The Multidimensional Mood State Questionnaire (MDMQ): Manual]. Göttingen: Hogrefe Verl. für Psychologie.

Tabibnia, G., & Lieberman, M. D. (2007). Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1118*, 90–101. https://doi.org/10.1196/annals.1412.001

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*, 339–347. https://doi.org/10.1111/j.1467-9280.2008.02091.x

Tomson, K., Vaht, M., Laas, K., Veidebaum, T., & Harro, J. (2016). Effect of a human serotonin 5-HT2A receptor gene polymorphism on impulsivity: Dependence on cholesterol levels. *Journal of Affective Disorders*, *206*, 23–30. https://doi.org/10.1016/j.jad.2016.07.036

Van Honk, J., Tuiten, A., Verbaten, R., van den Hout, M., Koppeschaar, H., Thijssen, J., & Haan, E. de (1999). Correlations among salivary testosterone, mood, and selective attention to threat in humans. *Hormones and Behavior*, *36*, 17–24. https://doi.org/10.1006/hbeh.1999.1521

Wagels, L., Votinov, M., Kellermann, T., Eisert, A., Beyer, C., & Habel, U. (2018). Exogenous testosterone enhances the reactivity to social provocation in males. *Frontiers in Behavioral Neuroscience*, *12*, 37. https://doi.org/10.3389/fnbeh.2018.00037

Wang, C., Cunningham, G., Dobs, A., Iranmanesh, A., Matsumoto, A. M., Snyder, P. J., . . . Swerdloff, R. S. (2004). Long-term testosterone gel (AndroGel) treatment maintains beneficial effects on sexual function and mood, lean and fat mass, and bone mineral density in hypogonadal men. *The Journal of Clinical Endocrinology and Metabolism*, *89*, 2085–2098. https://doi.org/10.1210/jc.2003-032006

Weinberg-Wolf, H., & Chang, S. W. C. (2019). Differences in how macaques monitor others: Does serotonin play a central role? *Wiley Interdisciplinary Reviews. Cognitive Science*, *10*, e1494. https://doi.org/10.1002/wcs.1494

Wilkenfeld, S. R., Lin, C., & Frigo, D. E. (2018). Communication between genomic and non-genomic signaling events coordinate steroid hormone actions. *Steroids*, *133*, 2–7. https://doi.org/10.1016/j.steroids.2017.11.005

Wingfield, J. C., Hegner, R. E., Dufty, A. M., & Ball, G. F. (1990). The "Challenge Hypothesis": Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *The American Naturalist*, *136*, 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*, 239–251. https://doi.org/10.1159/000047243

List of Figures

List of Tables

Supplementary Materials

Abstract

The steroid hormone testosterone plays an important role in human social behavior. It influences both prosocial and aggressive behavior. Recent research has indicated that testosterone can increase aggressive reactions in men with certain personality traits. Here, we intended to investigate if these personality traits also moderate the effect of testosterone on prosocial behavior. A modified version of the Point Subtraction Aggression Paradigm was developed and implemented in a testosterone administration study. The task allowed for both aggressive and prosocial actions of the participants and included the confrontation with three differently behaving other players as a context manipulation. In the study, 212 healthy males were randomly assigned to receive a placebo or exogenous testosterone via topical gel in a double-blind procedure. The findings indicated that testosterone affected the frequency of prosocial actions of the participants in response to a friendly player who gave them points with monetary value, and personality traits moderated this effect. We found that testosterone promoted prosocial actions in men with low dominance and high self-control; by contrast, testosterone attenuated prosocial actions in men with high dominance and low self-control. However, no effect of testosterone on the aggressive actions was found. In addition, participants were asked to rate how fair the other players were behaving during the task. The results provided indications that testosterone had an effect on the participants' fairness ratings. Implications for future studies are discussed.

Keywords: testosterone administration, aggression, prosocial behavior, PSAP

Abstract (German)

Das Steroidhormon Testosteron spielt eine wichtige Rolle im menschlichen Sozialverhalten. Es beeinflusst sowohl prosoziales, als auch aggressives Verhalten. Jüngste Untersuchungen haben gezeigt, dass Testosteron bei Männern mit bestimmten Persönlichkeitsmerkmalen die Häufigkeit von aggressiven Reaktionen erhöhen kann. In dieser Studie wollten wir untersuchen, ob diese Persönlichkeitsmerkmale mit der Wirkung von Testosteron auf das prosoziale Verhalten in Zusammenhang stehen. Eine modifizierte Version des Point-Subtraction-Aggression-Paradigm wurde entwickelt und in einer Testosteron-Applikationsstudie vorgegeben. Der Task erlaubte sowohl aggressive als auch prosoziale Handlungen der Studienteilnehmer und beinhaltete eine Kontextmanipulation. Die Studienteilnehmer waren mit drei, sich unterschiedlich verhaltenden Mitspielern konfrontiert. In der Studie erhielten 212 gesunde Männer im Doppelblindverfahren randomisiert entweder ein Placebo, oder Testosteron über ein Gel. Die Ergebnisse zeigten, dass Testosteron die Häufigkeit prosozialer Handlungen der Teilnehmer in Reaktion auf einen freundlichen Spieler, der ihnen Punkte mit monetärem Wert gab, beeinflusste. Persönlichkeitsmerkmale beeinflussten diesen Effekt. Bei Männern mit niedriger Dominanz und hoher Selbstkontrolle förderte Testosteron prosoziale Aktionen, während bei Männern mit hoher Dominanz und niedriger Selbstkontrolle prosoziale Aktionen durch Testosteron abgeschwächt wurden. Es wurde jedoch keine Wirkung von Testosteron auf aggressive Reaktionen festgestellt. Zusätzlich wurden die Teilnehmer gebeten, zu bewerten, wie fair sich die anderen Spieler während der Aufgabe verhielten. Hier lieferten die Ergebnisse Hinweise darauf, dass Testosteron die Fairness-Bewertungen der Teilnehmer beeinflusst hat. Implikationen für zukünftige Studien werden diskutiert.

Stichworte: Testosteron Applikation, Aggression, Prosoziales Verhalten, PSAP

Questionnaire Results

In this table we provide the resulting measurements from all questionnaires used in the study. Mean ratings are provided separately by group. The reliability, of course, was calculated on the whole sample data. Robust tests on mean differences were performed to test for baseline differences between the groups.

Personality and Mood Questionnaire Results.

Scale	Group[a]	*M* [95% CI]	Reliability[b]	Yuen's Test
DPS				
Dominance	P	3.89 [3.72, 4.07]	α=.77 [.72, .82]	T_y (125.9) = 0.69, p = .489
	T	4.00 [3.81, 4.18]		
Prestige	P	5.12 [4.97, 5.26]	α=.71 [.63, .76]	T_y (111.4) = 1.47, p = .145
	T	5.31 [5.20, 5.42]		
BIS-15	P	31.22 [30.01, 32.43]	α=.82 [.77, .84]	T_y (123.1) = 0.69, p = .493
	T	31.78 [30.53, 33.02]		
PSS-10	P	24.00 [22.75, 25.25]	α=.85 [.81, .98]	T_y (126) = 1.95, p = .054
	T	25.29 [24.18, 26.40]		
MDBF				
Good - Bad	P	33.51 [32.72, 34.29]	α=.84 [.78, .87]	T_y (106.7) = 2.84, p = .005, ξ =0.32, 95% CI [0.08, 0.50]
	T	35.15 [34.54, 35.75]		
Awake - Tired	P	27.45 [26.20, 28.71]	α=.91 [.88 .92]	T_y (114.9) = 0.83, p = .410
	T	27.86 [26.74, 28.99]		
Calm - Nervous	P	33.31 [32.41, 34.20]	α=.83 [.77 .87]	T_y (121.9) = 1.44, p = .153
	T	34.18 [33.38, 34.98]		

Note. Point range for scales are: DPS: 1 to 7, BIS-15: 15 to 45, PSS-10: 10 to 50 and MDBF: 8 to 40. High values on the mood scales represent positive mood states (feeling good, awake or calm). T_y = test statistic for Yuen's Test, ξ = robust measure of effect size (interpret like Cohen's d).
[a] P = placebo group (n= 95), T = testosterone group (n = 117). [b] α = Cronbach's α, bootstrapped 95% CI in brackets.

Task Instructions

The following text passage contains the translated (from German) task instructions. The instructions and practice run took about 4 Minutes to complete.

> Please wait for the instructions of the experimenter. [Press a button told by experimenter]
>
> Please read the following instructions carefully!
>
> Your task will be to play a game with other people.
>
> You will first play an exercise to learn how the game works.
>
> When ready, press the space bar to begin the exercise instruction.
>
> In this game, you can increase your score if you respond correctly. The higher your score, the more money you can earn.
>
> You will see a lot of moving white dots on the screen. It looks like this…
>
> [Example]
>
> Your score will be displayed here [the counter is circled].
>
> From time to time, three colored dots will appear that move as well. If two of these three points are the same color, as they are now [the participants see an example on screen], you should press the space bar to increase your score by one. [A suitable example is displayed and they have to press the space bar; the counter on the screen increases by one]
>
> Your score now increased by one!
>
> There will also be cases where all three colored dots have different colors, like now. [Matching example displayed] or all three colored dots have the same color, just like now [a suitable example is displayed]. In such cases, you should **not** press the spacebar to increase your score by one.
>
> If you press the spacebar, if all three colored dots have different colors, or all three have the same color, then you lose a point of your score.
>
> You will play several rounds.
>
> At the beginning of a round you will only see the white dots. After a while, the colored ones will pop up and you should make the right reaction. If you've pressed the space bar, or if you did not press the space bar and a certain amount of time has passed, the colored dots will disappear and a new round will start.

Important: You can increase your score by one in each round if you respond correctly.

Summary [They see it on screen before the exercise starts]:
You will play several rounds in the exercise. In order to increase your score, you must react correctly in each round. If two of the three colored dots are the same color, you must press the space bar to increase your score. If all colored dots are the same color, or all colored dots have different colors, you should not press the space bar to increase your score. If you still press the spacebar in such cases, your score will decrease by one. Important: You can increase your score by one in each round if you get the right reaction! If anything is still unclear, you can now ask the experimenter (please raise your hands). Otherwise you can press the space bar to start the exercise!

[They play the exercise].

You now know how the game works and how you can increase your score to earn money.

Now follow the instructions for the game. [Press the spacebar]

The game will be played together with other players. Before the game starts, you will be connected to your teammates via the internet. [Press the spacebar]

Now please choose an alias, which will be displayed to your teammates. Just enter a letter under which you want to play.

Press the corresponding key on the keyboard (without special characters). [They choose an alias]

Here you can see the code name of your fellow player. [Example will be displayed]. You will not be playing with the same player throughout the game. You will be informed on screen if that happens. If you change players, your score will be set to 0. That does not mean that your earned points are gone. At the end of the game you will see how much you have earned in total.

During the game you have additional options. You can give your opponent a point for his score. To do this, you must press the left shift key. The key is highlighted in red on the keyboard displayed [The participants will see an image of a keyboard where the corresponding key is highlighted in red]. [They must now press this button to continue]

If you give your opponent a point for his score you will see a green signal, as now. [The participants will see an example on the screen]. You can choose this option anytime

during the game. However, in this round you cannot earn a point for your score by making the right reaction.

You can also take away one point from your score. To do this, you must press the left control key (Ctrl). The key is highlighted in red on the keyboard displayed [you will see an image of a keyboard where the corresponding key is highlighted in red]. (You must now press this button to continue).

If you take one point away from your score, you will see a red signal, as it is now. [You will see an example on the screen]. You can choose this option anytime during the game. However, in this round you cannot earn a point for your score by making the right reaction.

Today, you were randomly assigned to the condition of the experiment, in which you cannot keep the points you take away from your opponent.

Your teammates can also give you a point for their score or take away one point!

If your teammates give you a point for their score, you will see these green icons. [You see an example].

If your teammates take away one point from their score, they will see these red symbols. [You see an example on the screen]

Synopsis [see it on screen before the game starts]: During the game you should increase your score as you learned in the exercise. Press the spacebar when two of the three colored dots have the same color. If all three colored dots are the same color, or all three colored dots have different colors, you should not press the space bar. If you press the spacebar in these cases, you lose one point of your score. During the game, you have the option of giving your teammates a point for the score (by pressing the left shift key) or of taking away one point from their score (by pressing the left control key). Your teammates also have these options and can give them points for their score or take them away. You will now be connected to your fellow players. Just wait for that to finish. Then the game begins.

Show the experimenter that you have finished the instruction (by raising your hand briefly) and then wait for the experimenter to instruct you to begin. [they must press a concealed button to start]

Additional Descriptive Statistics

Different sequences. Of the 212 participants that completed the modified PSAP, 107 played first with the prosocial player and 105 were paired with the aggressive player first. The different sequences were quite balanced within each experimental group. In the placebo group 50 participants played first with the prosocial player and 45 played first with the aggressive player. In the testosterone group 57 participants played first with the prosocial player and 60 played first with the aggressive player.

Points gained. On average, the participants achieved a score of 124.63 points, 95% CI [122.21, 127.05]. The range of values was 22 to 180 and 50% of the participants gained between 98 and 149 points. In the neutral condition the participants gained an average of 43.90 points, 95% CI [43.03, 44.76], in the prosocial condition 58.07 points, 95% CI [56.72,59.42], and in the aggressive condition 23.02 points, 95% CI [22.32, 23.72]. The means were calculated using robust methods (10% trimmed means). No differences between the groups were observed.

Actions of other players. The participants received on average 11.51 points, 95% CI [11.18, 11.85], from the prosocial player. This was randomized and so some participants received more points, others less. The range of received points was 6 to 20, but 50% of the participants received between 10 and 13 points. In the aggressive condition on average 12 points, 95% CI [11.61, 12.37], were reduced from the participants with a range from 5 to 21 and 50% of the participants experienced between 10 and 14 point reductions.

Response times. In the trials where it was required to press the space bar to earn points, the participants needed on average 800ms, 95% CI [795, 805], to do so correctly. No differences between the groups were observed.

Notes on observed behavior. A wide variety of responses to the different conditions of the task were observed. In total 42 participants chose to reduce points while playing with the prosocial player. However, only 12 of those 42 participants exclusively reduced points, the others reduced as well as gave points in this condition. Similarly, 128 participants decided to give points to the player who was reducing their points instead of retaliating by reducing the other player's points too. Again, only 39 of those 128 participants exclusively gave points to the aggressive player while the others chose from both possible actions.

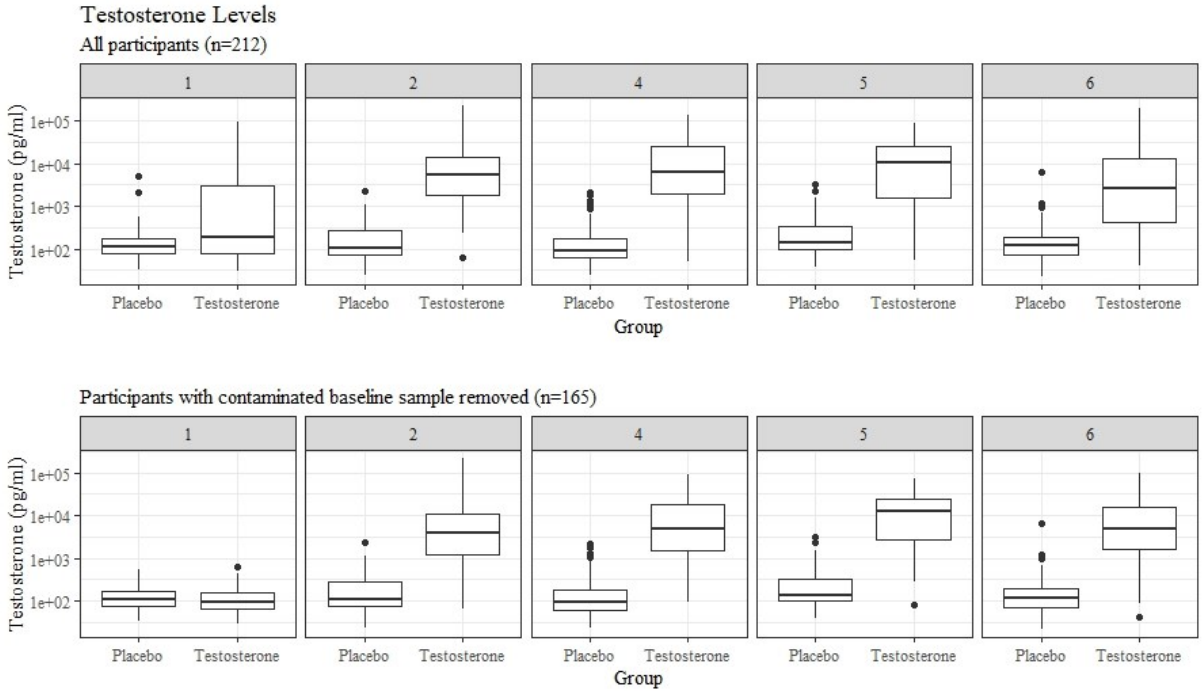
Additional Information on the Hormone Data Analysis

Testosterone. The table below shows the results from the post-hoc pairwise comparisons (Bonferroni corrected for multiple comparisons) between the placebo and the testosterone group, n=165. Sample = timepoint of saliva sample in the study, estimate =

estimated mean difference, SE = standard error of mean difference, LL = lower limit of 95% confidence interval for estimate, UL = upper limit of CI.

Sample	Estimate	SE	*df*	LL	UL	*t*-ratio	*p*
1 (Baseline)	8.06	2038	582.25	-5261.03	5277.15	0	1.00
2	-13029.17	2045	585.07	-18315.54	-7742.8	-6.37	< .001
4	-12037.29	2042	583.94	-17316.71	-6757.87	-5.89	< .001
5	-15976.42	2428	710.57	-22249.25	-9703.6	-6.58	< .001
6	-12604.39	2038	582.25	-17873.48	-7335.3	-6.18	< .001

Boxplots showing the measured testosterone values at each sample, for the whole dataset (above) and the reduced dataset including only participants with a clean baseline sample.



Dihydrotestosterone (DHT). First, we looked at the data from all 212 participants. From these 212 participants, 150 provided 401 samples that allowed the analysis of the DHT levels. A 2 (group: placebo or testosterone) × 5 (sample: 1 to 6) linear mixed model (including a random intercept per participant to account for the repeated measures) revealed a significant main effect of group, $F(1, 193.32) = 14.54$, $p < .001$, but no significant effect of sample, $F(4, 366.05) = 0.58$, $p = .68$, and no significant Group × Sample interaction $F(4, 366.05) = 0.66$, $p = .62$. Even though the interaction was not significant, we calculated the post-hoc pairwise comparisons (reported in the table below) to explore the results. These indicated that the two groups did not differ significantly in their DHT levels at sample point 1, 2, or 4 but did so at sample point 5 and 6. The pairwise comparisons were again Bonferroni corrected for multiple comparisons.

Sample	Estimate	SE	*df*	LL	UL	*t*-ratio	*p*
1	-29.08	35.54	407.97	-121.06	62.9	-0.82	1.000
2	-54.43	28.78	393.62	-128.93	20.07	-1.89	.297
4	-68.37	30.7	400.49	-147.82	11.07	-2.23	.132
5	-89.48	33.45	409.86	-176.05	-2.92	-2.68	.039
6	-90.14	30.26	403.63	-168.45	-11.83	-2.98	.015

We then looked at the data including only the 165 participants with clean baseline samples (regarding the testosterone levels), 105 participants provided 268 samples that allowed for DHT analysis. A 2 (group: placebo or testosterone) × 5 (sample: 1 to 6) linear mixed model (including a random intercept per participant to account for the repeated measures) revealed again a significant main effect of group, $F(1, 169.15) = 11.69$, $p < .001$, but no significant effect of sample, $F(4, 297.94) = 1.07$, $p = .37$, and no significant Group × Sample interaction $F(4, 297.94) = 1.22$, $p = .30$. Exploratory post-hoc pairwise comparisons ((Bonferroni corrected for multiple comparisons), revealed that the two groups did not differ significantly in their DHT levels at sample point 1,2, or 4 but did so at sample point 5 and 6.

Sample	Estimate	SE	*df*	LL	UL	*t*-ratio	*p*
1	-4.98	38.27	277.34	-104.25	94.28	-0.13	1.000
2	-20.18	25.29	267.87	-85.8	45.44	-0.8	1.000
4	-44.64	26.46	270.62	-113.28	24.01	-1.69	.464
5	-80.15	28.76	277.08	-154.75	-5.55	-2.79	.028
6	-87.42	25.67	270.97	-154.01	-20.82	-3.41	.004

Estradiol. Here, we also first looked at the data from all 212 participants (982 samples). A 2 (group: placebo or testosterone) × 5 (sample: 1 to 6) linear mixed model (including a random intercept per participant to account for the repeated measures) revealed a significant Group × Sample interaction, $F(4, 780.34) = 2.61$, $p = .035$. Post-hoc pairwise comparisons (reported in the table below) revealed that the two groups did only differ significantly in their estradiol levels sample point 6. The pairwise comparisons were again Bonferroni corrected for multiple comparisons.

Sample	Estimate	SE	*df*	LL	UL	*t*-ratio	*p*
1	0.04	0.31	742.62	-0.76	0.83	0.12	1.000
2	-0.47	0.31	742.62	-1.26	0.33	-1.51	0.657
4	-0.76	0.31	747.02	-1.55	0.04	-2.44	0.074
5	-0.78	0.37	908.45	-1.73	0.17	-2.12	0.170
6	-1.06	0.31	744.47	-1.86	-0.27	-3.45	0.003

We then also looked at the data including only the 165 participants with clean baseline samples regarding the testosterone levels (770 samples). A 2 (group: placebo or testosterone) × 5 (sample: 1 to 6) linear mixed model (including a random intercept per participant to account for the repeated measures) revealed no significant main effect of group, $F(1, 168.9) = 1.44$, $p = .23$, no significant effect of sample, $F(4, 612.4) = 0.72$, $p = .57$, and no significant Group × Sample interaction $F(4, 612.4) = 1.58$, $p = .18$. The groups did not differ in their estradiol levels.

Model Specifications and Comparisons.

Model Specifications. The table below describes the configuration of the models used in the model comparisons to assess the best fitting model. The models were calculated using the *lme4* package in R using a logit link-function. For the factor *condition* treatment contrasts were used, but for the factor group we used sum-to-zero contrasts. We compared the models using a chi-square likelihood ratio test to determine, which predictors should be included in the model and if using random slopes in addition to using random intercepts improves the model fit.

Model Specifications for the Model Comparisons using a Chi-Square Likelihood Ratio Test.

Model	Predictors included
Baseline	Intercept + Random Intercept per Participant
Model 1	Condition + Random Intercept per Participant
Model 2	Condition + Random Intercept and Slopes (per Condition)
Model 3	Condition + Group + Random Intercept and Slopes
Model 4	Condition × Group + Random Intercept and Slopes
Model 5	Condition + Risk Score + Random Intercept and Slopes
Model 6	Condition × Risk Score + Random Intercept and Slopes
Model 7	Condition + Group + Risk Score + Random Intercept and Slopes
Model 8	Condition × Group + Risk Score + Random Intercept and Slopes
Model 9	Condition + Group × Risk Score + Random Intercept and Slopes
Model 10	Condition × Risk Score + Group + Random Intercept and Slopes
Model 11	Condition × Group × Risk Score + Random Intercept and Slopes

Note. Condition = three level factor specifying the different conditions (neutral, prosocial, aggressive) of the task. Group = either testosterone or placebo treatment group. Risk score = standardized score including dominance and self-control measurements. × = Interaction between predictors is included, + = only main effects of predictors are included.

Model comparisons for the prosocial actions. The results from the model comparison indicate that including the within-subject factor condition greatly improves the model fit, model 1 vs. baseline, χ^2 (2) = 755.47, p = < .001. Including random slopes for the factor condition drastically improves the model too, model2 vs. model1: χ^2 (5) = 269.42, p = < .001. Including the factor group (testosterone/placebo) either only as a main effect, model 3 vs model 2: χ^2 (1) = 0.07, p = .79, or as a main effect and a Condition × Group interaction does not improve the model, model 4 vs. model 2, χ^2 (3) = 2.47, p = .48. Including the risk score as a main effect, however, does significantly improve the model fit, model 5 vs model 2, χ^2 (1) = 5.64, p = .018. Almost a significant improvement occurs if the risk score is included not only as a main effect but also as a Condition × Risk Score interaction, model 6 vs. model 2, χ^2 (3) = 7.40, p = .06. Compared with the model including only condition and the risk score as main effects (model5), including also group either as a main effect (model 7), χ^2 (1) = 0.36, p = .55, or as an interaction with either condition, model 8 vs. model 5, χ^2 (3) = 2.77, p = .43, or the risk score, model 9 vs. model 5, χ^2 (2) = 1.83, p = .40, does not improve the model further.

Also, a model including only the crossed predictors condition and risk score and the main effect of group provides no improvement, model 10 vs. model 5, χ^2 (3) = 2.10, p = .55. The most complex model (model 11) provides an improvement compared with the second-best model (model 5) including only condition and risk score as main effects: χ^2 (8) = 16.31, p = .038 and also when compared with any other model previously included in any model comparison.

However, alternative fit indices provide a mixed result. Schwarz's Bayesian information criterion (BIC), which corrects more conservatively for the number of parameters estimated, is increased in this model compared to simpler models including fewer interactions (e.g. BIC (Condition × Group × Risk Score) = 20686, BIC (condition and group included only as main effects) = 20617), the Akaike information criterion (AIC) has the same value (AIC = 20530) in both models and this also is the lowest value of all models compared. Although slightly ambiguous, these results provide at least some support to further investigate possible effects based on the model including the three-way interaction.

To assess the quality of the model we used the DHARMa package in R (Hartig, 2009) to run a residual analysis. DHARMa provides a simulation-based approach to analyze and interpret residuals from generalized linear mixed models (among others). The distribution of the resulting residuals (based on 1000 simulations) does not show any obvious patterns that would indicate either heteroscedasticity or a misspecified model. Most tests performed on the residuals (e.g. for a significant number of outliers) provide a non-significant result. The only test that results in a significant result is the K-S test on uniformity of the residuals (D = 0.007, *p*=.023) suggesting a possible deviation of the residual distribution from the expected distribution. However, no deviation of the residuals is visible in the corresponding plots (also when plotted against different predictors) and therefore it seems likely that this significant result is due to the large sample size (and the high power of this test).

Model comparisons for the aggressive actions. The models built and compared were the same as in the previous analysis on the prosocial actions except that the 0/1 coded dependent variable now indicates if a point was reduced by the participant (see Table above for the model specifications). Compared to the baseline model including only the intercept, the model including the factor condition as a predictor significantly improves the model fit (χ^2 (2) = 1145.00, *p* = <.001) and again, including random slopes for each condition significantly improves the model (model 2 vs. model 1: χ^2 (5) = 368.90, *p* = <.001). Including the factor group as a main effect (model 3 vs. model 2: χ^2 (1) = 0.03, *p* = .86) or an interaction term (model 4 vs. model 2: χ^2 (3) = 0.90, *p* = .83) does not improve the model. Other than with the

prosocial actions, including the risk score does not improve the model fit (model 5 vs model 2: χ^2 (1) = 1.95, p = .16; model 6 vs. model 2: χ^2 (3) = 5.94, p = .11) and no other model tested provides a better fit than the model including only condition as predictor (we refrain from reporting all the insignificant tests here). The most complex model (model 11) including all three fully crossed predictors of interest does result in the lowest deviance. However, the difference in deviance is not significant compared with any other model (this is the case with the model on the prosocial actions though). E.g. when compared with the model including only the factor condition no improvement occurs, model 11 vs. model 2: χ^2 (9) = 10.09, p = .34. The AIC and BIC are also increased compared to the simpler models. Based on this series of model comparisons, we conclude that there is no evidence for any effect of the factor group (placebo/testosterone) or the risk score on the aggressive actions of the participants.

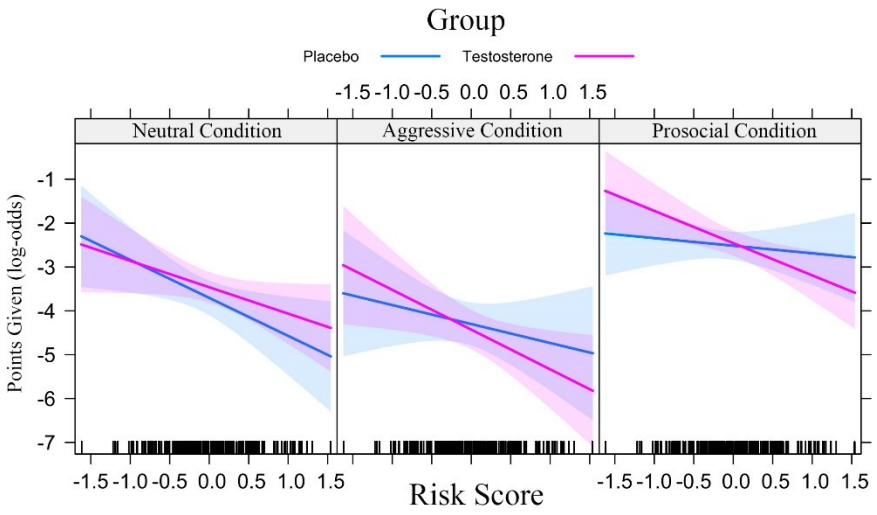
Additional Model Results

Including the amount of points received as covariate. The table below shows the model results (modelling points given) including the amount of points received (from the friendly player) as a covariate.

Analysis of Deviance Table (Type III Wald χ^2 tests)

Predictor	χ^2	*df*	*p*
(Intercept)	667.70	1	<.001
Condition	251.70	2	<.001
Group	0.85	1	0.36
Risk Score	9.13	1	0.003
Points Received (centered)	9.86	1	0.002
Condition × Group	4.89	2	0.09
Condition × Risk Score	4.00	2	0.14
Group × Risk Score	0.30	1	0.58
Condition × Group × Risk Score	10.62	2	0.005

Plot of the estimated effects from the model including the amount of points received by each participant as a covariate. The direction of the effects did not change.

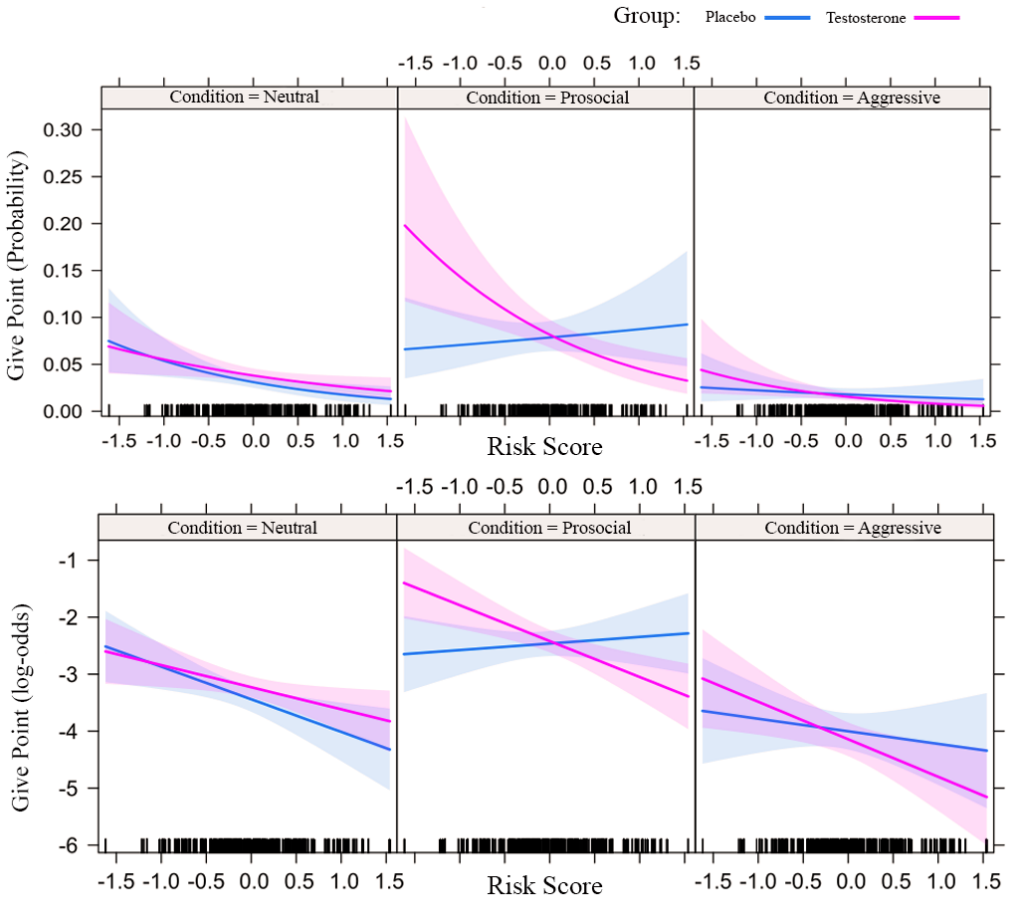


Results from the model including the total amount of actions as a covariate. The table below reports the calculated contrasts from the model including the total amount of actions taken by the participants as covariate. We no longer see an indication for a significant Aggressive Condition × Testosterone × Risk Score effect (contrasted with the neutral condition).

	Reference = Neutral Condition			**Reference = Prosocial Condition**			**Reference = Aggressive Condition**		
Predictors	**OR**	**95% CI**	***p***	**OR**	**95% CI**	***p***	**OR**	**95% CI**	***p***
(Intercept)	0.03	0.03 – 0.04	**<.001**	0.08	0.07 – 0.10	**<.001**	0.02	0.01 – 0.03	**<.001**
Neutral Condition				0.37	0.30 – 0.47	**<.001**	1.66	1.26 – 2.20	**<.001**
Prosocial Condition	2.67	2.13 – 3.35	**<.001**				4.45	3.26 – 6.06	**<.001**
Aggressive Condition	0.60	0.45 – 0.80	**<.001**	0.22	0.16 – 0.31	**<.001**			
Testosterone	1.24	0.94 – 1.64	.130	1.05	0.77 – 1.42	.772	0.87	0.59 – 1.28	.483
Risk Score	0.60	0.40 – 0.90	**.013**	1.19	0.77 – 1.83	.440	0.85	0.46 – 1.48	.558
Amount of Actions	3.25	2.85 – 3.71	**<.001**	3.25	2.85 – 3.71	**<.001**	3.25	2.85 – 3.71	**<.001**
Neutral Condition × Testosterone				1.19	0.89 – 1.58	.242	1.42	1.01 – 2.01	**.044**
Prosocial Condition × Testosterone	0.84	0.63 – 1.12	.244				1.20	0.81 – 1.78	.362
Aggressive Condition × Testosterone	0.70	0.50 – 0.99	**.045**	0.83	0.56 – 1.24	.365			
Neutral Condition × Risk Score				0.50	0.33 – 0.77	**.002**	0.70	0.42 – 1.17	.178
Prosocial Condition × Risk Score	1.99	1.30 – 3.06	**.002**				1.40	0.79 – 2.48	.247
Aggressive Condition × Risk Score	1.42	0.85 – 2.38	.180	0.71	0.40 – 1.27	.254			
Testosterone × Risk Score	1.14	0.67 – 1.93	.630	0.45	0.25 – 0.79	**.006**	0.63	0.30 – 1.31	.214
Neutral Condition × Testosterone × Risk Score				2.55	1.46 – 4.44	**.001**	1.81	0.93 – 3.50	.079
Prosocial Condition × Testosterone × Risk Score	0.39	0.23 – 0.68	**.001**				0.71	0.34 – 1.50	.370
Aggressive Condition × Testosterone × Risk Score	0.55	0.28 – 1.08	.083	1.41	0.66 – 3.03	.379			
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	0.59 $_{\text{Participant}}$			0.95 $_{\text{Participant}}$			1.20 $_{\text{Participant}}$		
τ_{11}	0.47 $_{\text{Participant.Prosocial Condition}}$			0.47 $_{\text{Participant.Neutral Condition}}$			1.06 $_{\text{Participant.Prosocial Condition}}$		
	0.53 $_{\text{Participant.Aggressive Condition}}$			1.06 $_{\text{Participant.Aggressive Condition}}$			0.53 $_{\text{Participant.Neutral Condition}}$		
ρ_{01}	-0.10 $_{\text{Participant.Prosocial Condition}}$			-0.62 $_{\text{Participant.Neutral Condition}}$			-0.58 $_{\text{Participant.Prosocial Condition}}$		
	0.07 $_{\text{Participant.Aggressive Condition}}$			-0.40 $_{\text{Participant.Aggressive Condition}}$			-0.72 $_{\text{Participant.Neutral Condition}}$		
ICC	.22			**Observations**			43227		
Marginal R^2	.312			**Conditional R^2**			.461		

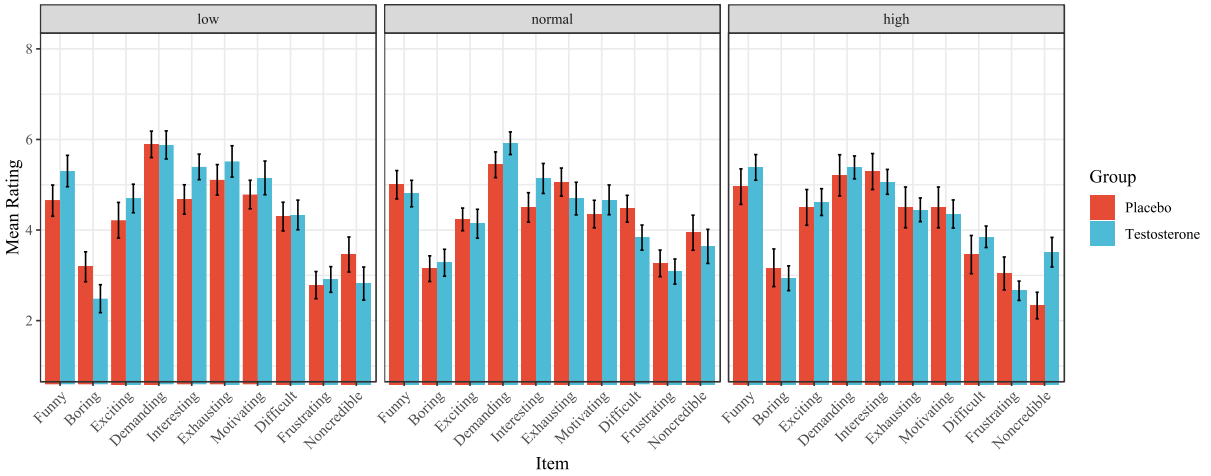
Note. The table shows the calculated comparisons from the model including the sum of all actions taken as an offset. Calculated comparisons are reported for with all three task conditions as the reference level. σ^2 = within-group (participant) residual variance, τ_{00} = variation between individual intercepts and average intercept, τ_{11} = random slope variance, ρ_{01} = random-intercept-slope correlation, ICC = intraclass-correlation coefficient (adjusted for random effect variance), marginal R^2 = variance explained by fixed effects only, condition R^2 = variance explained by fixed and random effects.

We then plotted the estimated effects of this model including the amount of all actions taken by each participant as a covariate. The crossover-interaction in the prosocial condition now is more pronounced.



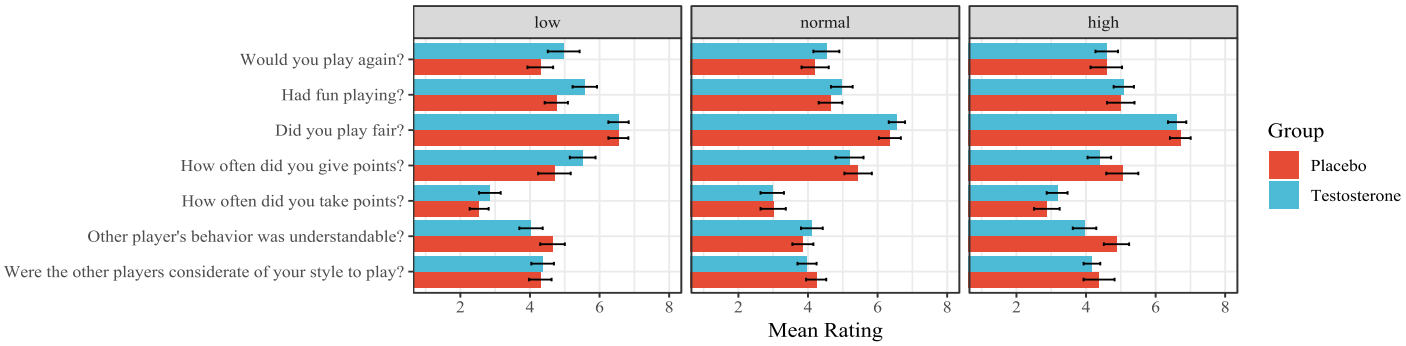
Additional Tables and Figures

Post-task questionnaire rating results. Here, we provide all rating results from the post task questionnaire. First, we provide the mean ratings for the items asking for task characteristics. The mean ratings are provided per group at low, normal and high levels of the risk score. The items asking how, funny, exciting, interesting, motivating, boring (reversed) and noncredible (reversed) the task was were combined into the positive scale reported in the results section of the thesis. The remaining items (demanding, exhausting difficult, frustrating) were combined into the demanding scale (also reported in the results section).



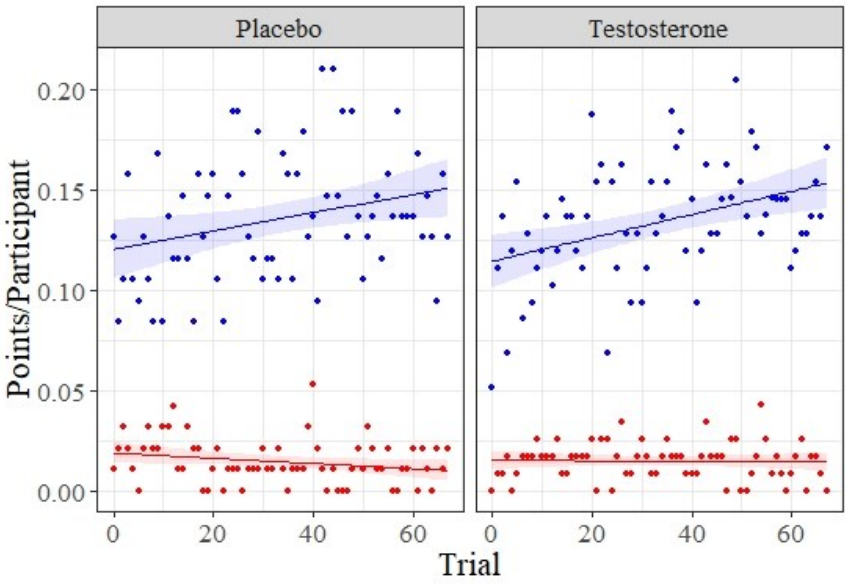
Note. Mean rating results from 10 items asking for characteristics of the task in total (across all conditions). For both groups and low, normal, and high risk score levels separately. The scale of the items ranged from 1 (*not at all*) to 8 (*very much*). Error bars indicate the standard error of the mean.

Second, we provide the results from the additional items asking for an assessment of the participants own behavior, if they enjoyed the task, how many points they took or gave and what they thought about the other player. These items asked for an assessment across all task conditions (not split for the neutral, friendly or aggressive player). Again, the mean ratings are provided split by group and risk score level.



Note. Mean rating results from 7 items asking for several aspects of the participants' and the other players' behavior (across all conditions). Reported for both groups and low, normal, and high risk score levels separately. The scale of the items ranged from 1 (*not at all*) to 8 (*very much*). Error bars indicate the standard

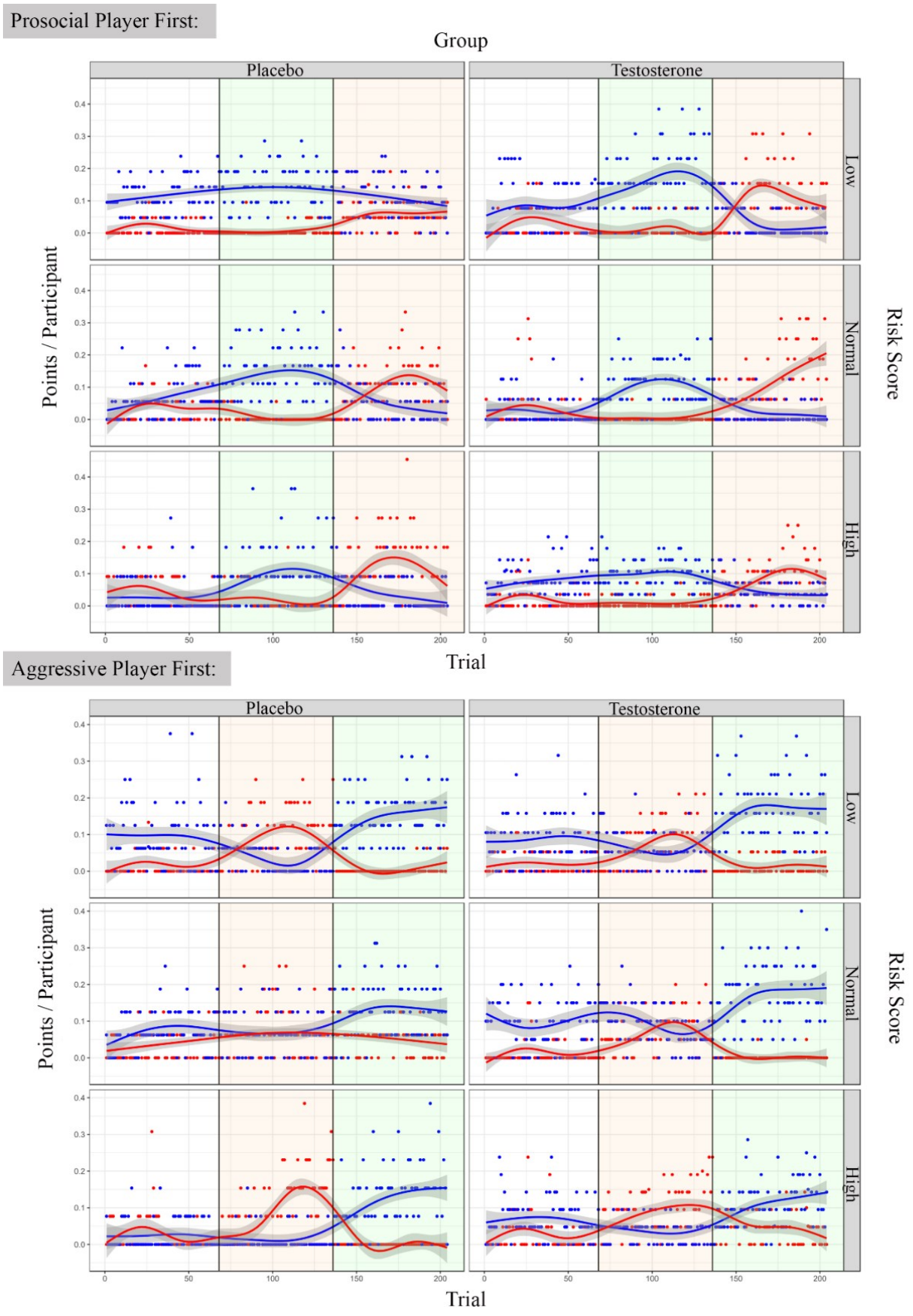
Chronological distribution of points given and taken per trial in the prosocial condition. This plot presents the average amount of points given per participant (in blue) vs. points reduced (in red) during the prosocial condition of the task (not considering the different sequences of the task conditions). We can see a slight increase in the amount of points given per trial as the condition progressed. Almost no points were reduced in this condition.



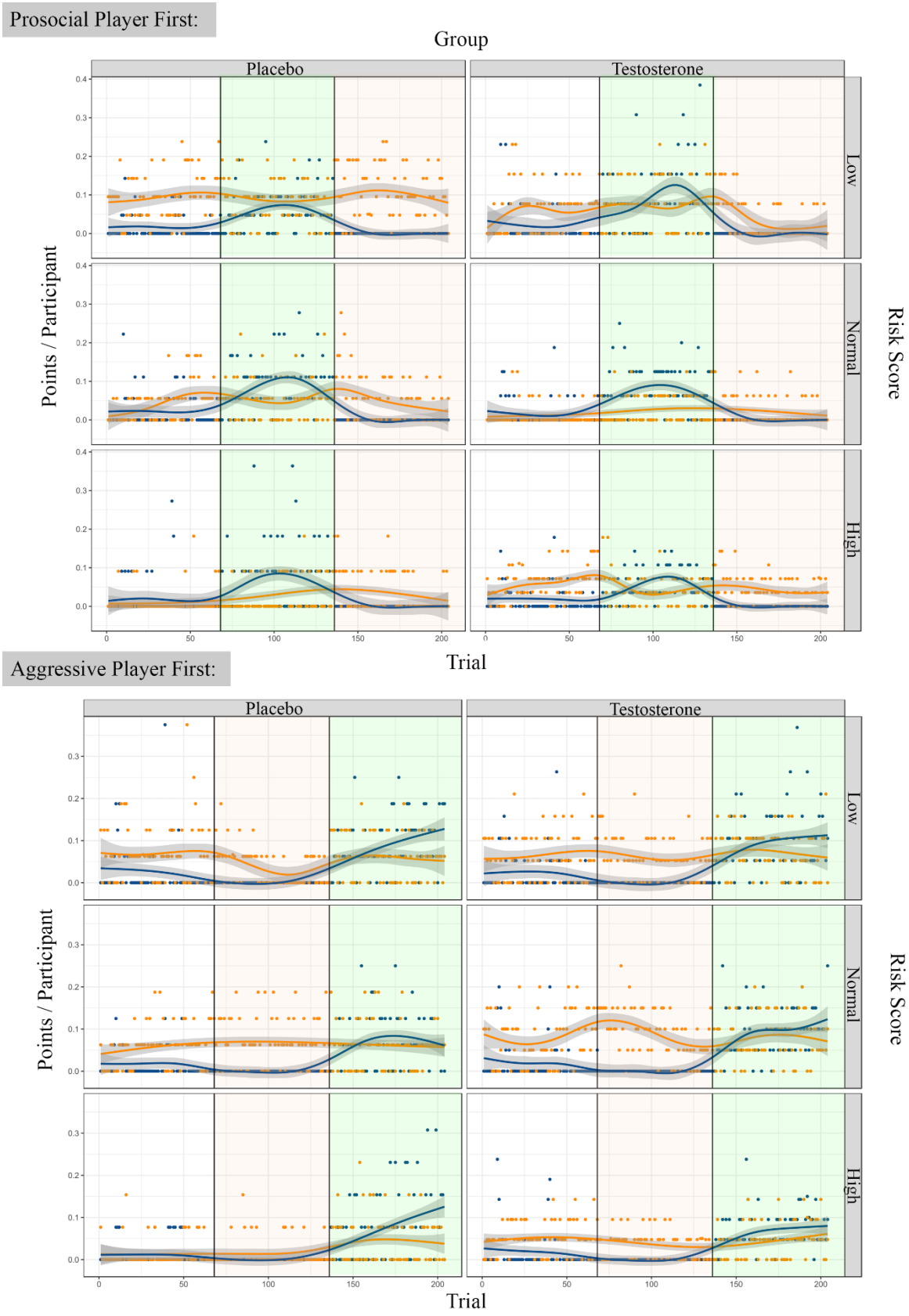
Detailed chronological distribution of points given and taken per trial. On the following pages we provide more detailed plots concerning the chronological distribution of actions taken by the participants across the whole task. To do this, we looked at each sequence and group, as well as low, normal, and high risk score levels separately. First, we looked at the distribution of points given and reduced across the whole task while considering the different sequences (aggressive player first vs. prosocial player first) and the risk score simultaneously (Chronological Plot 1). We calculated the average amount of points given (in blue) per trial. This was done by summing up all points given in a trial and then dividing this value by the number of participants to ensure comparability between groups and the different risk score levels. The same was done for the points taken (in red) per trial. We plotted these measurements and added smoothed conditional means for a trend indication.

We then did the same for the points given before or after the participant received a points from the other player. For this, we first calculated a variable that determined if each points given from the participant to the other player was given before the participant received a point or after. Then we calculated the average amount of points given before and after in each trial across the whole task in the same way as the procedure described above. Again, we considered the different sequences, groups and the risk score. The results are plotted in Chronological Plot 2.

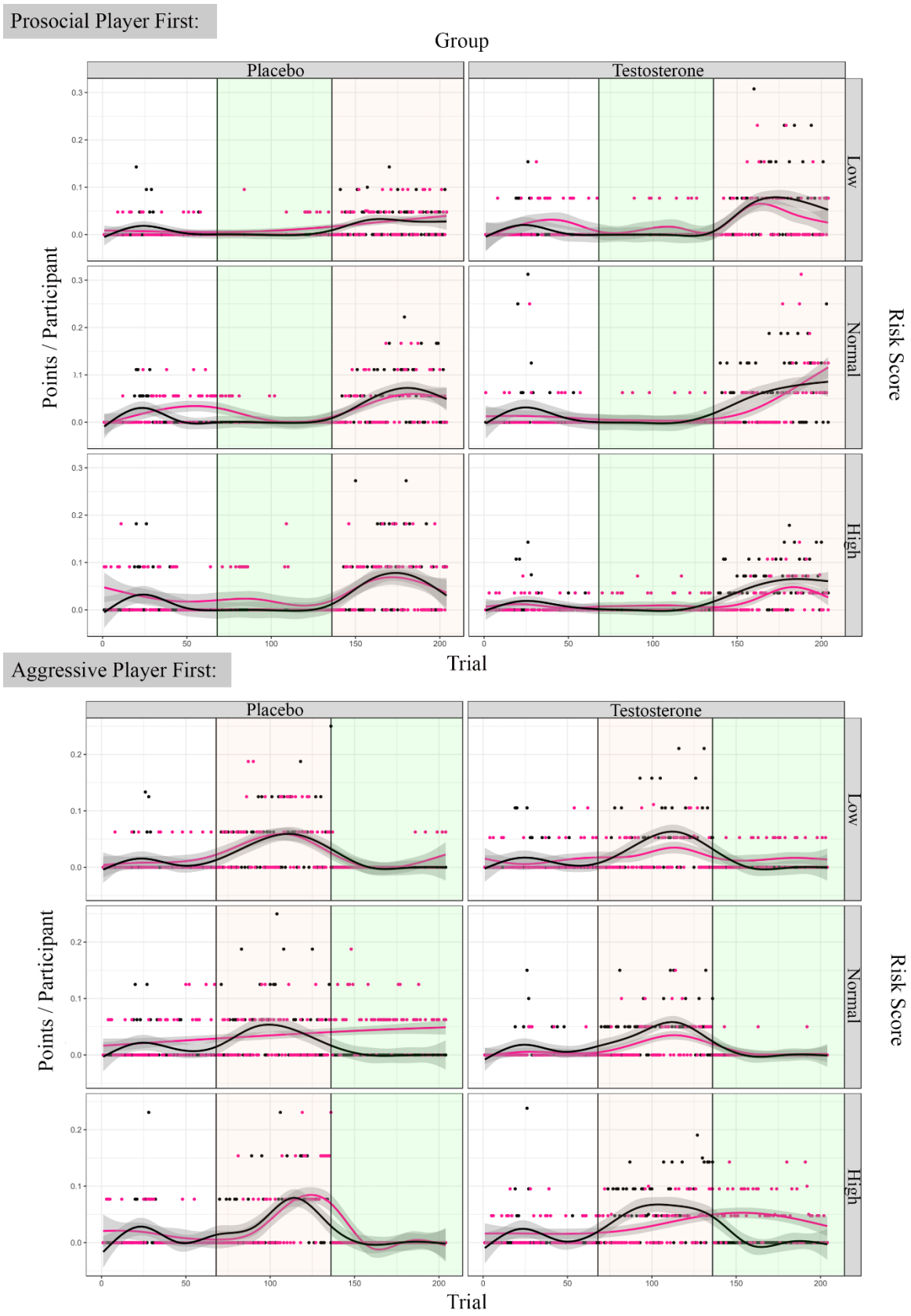
Finally, we did exactly the same as in Plot 2 just with the point reductions. First we calculated if each point reduction occurred before or after the other player reduced the points of the participants. Then we calculated the average amount of points reduced before and after in each trial across the whole task in the same way as the procedure described above. Again, we considered the different sequences, groups and the risk score. The results are plotted in Chronological Plot 3. These plots serve purely exploratory purposes, no inferential analysis was carried out on the measurements involved in these plots (see the main part of the thesis for additional information).



Chronological Plot 1. Average amount of points given per trial in blue. Average amount of points taken per trial in red. These measurements are standardized by the number of participants in each factor combination. Plotted across the whole task, taking the different sequences (aggressive vs. prosocial player first) into account. White background = neural player, red background = aggressive player, green background = prosocial player

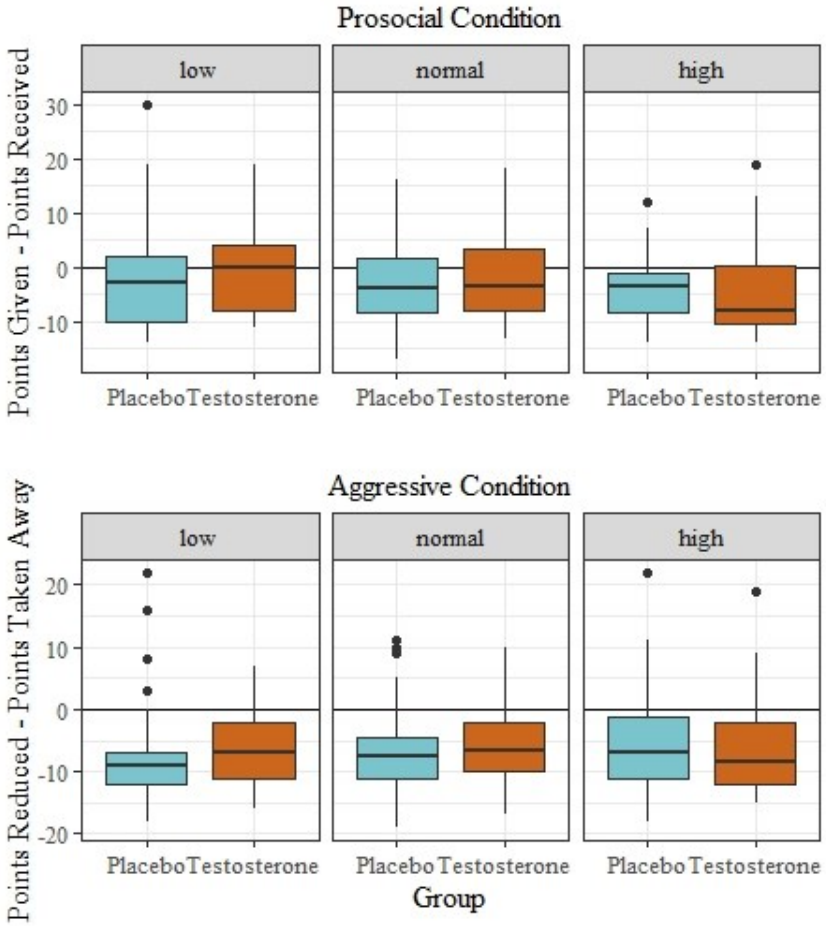


Chronological Plot 2. Plot of the average amount of points given before the participant received a point from the other player (in orange) and after the participants received a point (in green) per trial during the task (at low, normal, and high levels of the risk score; for both sequences separately).White background = neural player, red background = aggressive player, green background = prosocial player



Chronological Plot 3. Plot of the average amount of points taken before the participant experienced a point reduction from the other player (in pink) and after the participants experienced a point reduction (in black) per trial during the task (at low, normal, and high levels of the risk score; for both sequences separately). White background = neural player, red background = aggressive player, green background = prosocial player

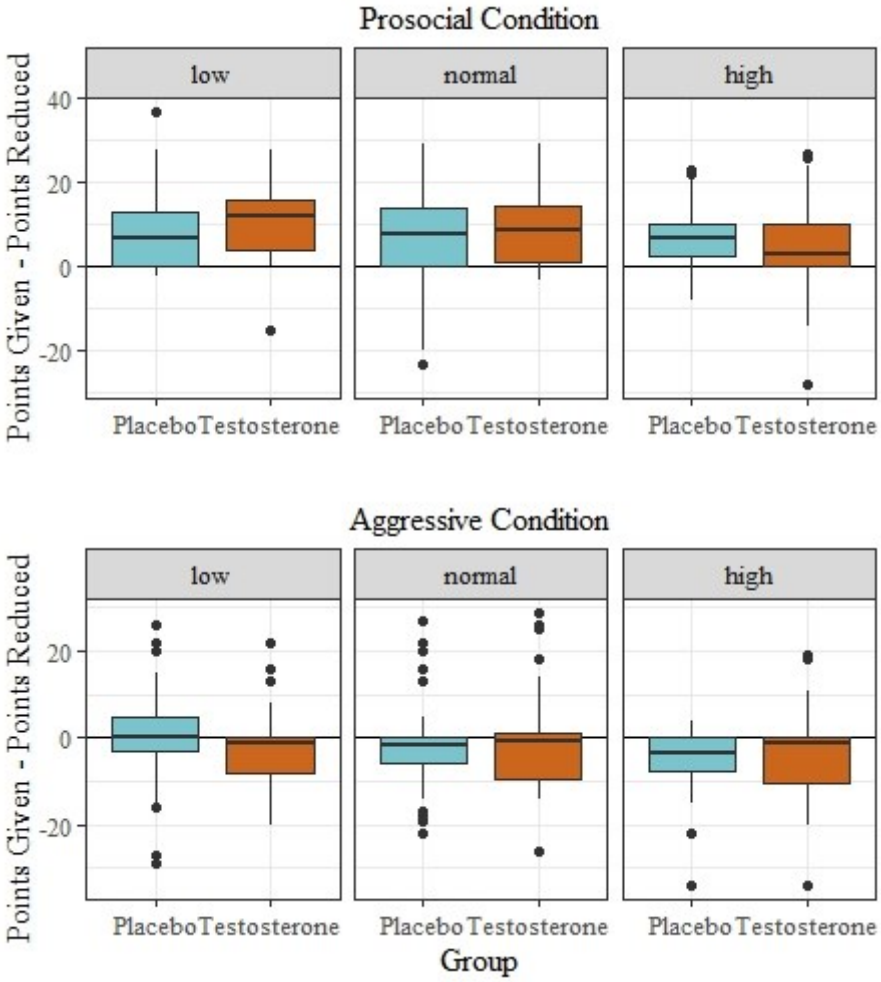
Difference measurements. To illustrate the relationship between the participants' behavior and the behavior of the other player, we provide a plot of the difference between the amount of actions of the participant in the prosocial condition and the amount of points he received from the friendly player. And similarly, the difference between the amount points the participants reduced and the amount of points the aggressive player reduced from them. The results are provided in Difference Plot 1.



Difference Plot 1. Difference between points given to the other player and points received from the other player (in the prosocial condition) plotted in the top row at low, normal, and high levels of the risk score. Difference between points reduced from the other player and points reductions experienced (in the aggressive condition) plotted below.

Because the participants had both options to act (give a point or reduce a point) available all the time throughout the task, we calculated the difference between the points given and points reduced by each participant in the prosocial and the aggressive condition of the task. The result can be seen in Difference Plot 2. Again, these plots serve purely

exploratory purposes, no inferential analysis was carried out on the measurements involved in these plots (see the main part of the thesis for additional information).



Difference Plot 2. Difference between points given and points reduced per participant at low, normal, and high levels of the risk score (0 = same amount of points given and reduced) in the prosocial and aggressive condition of the task. A negative difference indicates that more points were reduced than given by a participant.