A meta-analytical evaluation of the dual-hormone hypothesis: Does cortisol moderate the relationship between testosterone and status, dominance, risk taking, aggression, and psychopathy?

Dekkers, T.J.; Agelink van Rentergem, J.A.; Meijer, B.; Popma, A.; Wagemaker, E.; Huizenga, H.M.

Published in:
Neuroscience and Biobehavioral Reviews

DOI:
10.1016/j.neubiorev.2018.12.004

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Review article

A meta-analytical evaluation of the dual-hormone hypothesis: Does cortisol moderate the relationship between testosterone and status, dominance, risk taking, aggression, and psychopathy?

Tycho J. Dekkers\textsuperscript{a,b,*,1}, Joost A. Agelink van Rentergem\textsuperscript{a,c,1}, Bren Meijer\textsuperscript{a}, Arne Popma\textsuperscript{b,d,e}, Eline Wagemaker\textsuperscript{a}, Hilde M. Huizenga\textsuperscript{a,f,g}

\textsuperscript{a} University of Amsterdam, Department of Psychology, Amsterdam, The Netherlands
\textsuperscript{b} Department of Forensic Psychiatry and Complex Behavioral Disorders, De Bascule - Academic Center for Child- and Adolescent Psychiatry, Amsterdam, The Netherlands
\textsuperscript{c} Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
\textsuperscript{d} Department of Child- and Adolescent Psychiatry, Free University Medical Center (VUmc) Amsterdam, Amsterdam, The Netherlands
\textsuperscript{e} Department of Criminology, Leiden University, Leiden, The Netherlands
\textsuperscript{f} Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, The Netherlands
\textsuperscript{g} Research priority area Yield, Amsterdam, University of Amsterdam, The Netherlands

\textbf{ARTICLE INFO}

Keywords:
Cortisol
Testosterone
Interaction
Dual-hormone
Status-relevant behavior
Aggression
Meta-analysis
Simple slopes

\textbf{ABSTRACT}

According to the dual-hormone hypothesis, the relationship between testosterone and status-relevant behavior is moderated by cortisol, suggesting this relationship only exists when cortisol is low. In the current study, a meta-analysis (including 30 papers with 33 studies, 49 effect sizes, \(n = 8538\)) on the interaction effect of testosterone and cortisol on status-relevant behavior (i.e. status, dominance, risk taking, aggression, and psychopathy) was performed. There was only marginal support for the dual-hormone hypothesis: The effect size of the interaction between testosterone and cortisol on status-relevant behavior was significant but very small (\(r = -.061, p = .026\)), which was corroborated by follow-up meta-analyses on simple slopes on low and high cortisol. Effect sizes were largest for direct status measures, although not significantly different from other outcome measures. Similarly, effect sizes seemed larger for men than for women. However, robustness analyses indicated signs of publication bias, enhanced significance due to potential flexibility in data-analysis, and a lack of power of individual studies, emphasizing the need for a large, pre-registered study.

1. Introduction

A substantial body of animal research shows that, in unstable hierarchies, increased testosterone levels are associated with increased status-relevant behavior such as aggression and dominance (Wingfield et al., 1990; Archer, 2006). In research on humans, the relationship between testosterone and status-relevant behavior is less consistent, and a meta-analysis points out that the association between testosterone and aggression, although significant, is small (Archer et al., 2005, see Carré and Archer, 2018 for a review on the relationship between testosterone and aggression).

During the past few years, a new theoretical framework emerged that may partly account for mixed findings regarding the relationship between testosterone and status-relevant behavior. The dual-hormone hypothesis highlights the crucial role of cortisol, stating that testosterone is particularly related to status-relevant behavior when cortisol levels are low, and not when cortisol levels are high (Mehta and Josephs, 2010).

The neural, physiological and psychological effects of testosterone, a hormone released by the hypothalamic-pituitary-gonadal (HPG) axis, and cortisol, a hormone released by the hypothalamic-pituitary-adrenal (HPA) axis, have been studied extensively (Mehta and Josephs, 2010; Viau, 2002; Hamilton et al., 2015; Sollberger and Ehlers, 2016). Many studies investigated the effects of testosterone and cortisol on the striatal dopaminergic reward system (Mehta et al., 2015a). Animal studies indicated that testosterone increased activity of this system (e.g. de Souza Silva et al., 2009; Mitchell and Stewart, 1989; Shemisa et al., 2006; Thiblin et al., 1999; Wang et al., 2016; Zhang et al., 2016), a
finding which was corroborated by studies in humans (Hermans et al., 2010; Lombardo et al., 2012; Op de Macks et al., 2011; see Welker et al., 2015 for a review on testosterone and reward processing). Cortisol also increases dopaminergic striatal output (Marinelli and Piazza, 2002), as reported in both animal (Tsukada et al., 2011) and human studies (Oei et al., 2014; Oswald et al., 2005; Pruessner, 2004; Wand et al., 2007), although one study in humans reported a decrease in striatal activity after cortisol administration (Montoya et al., 2014). In sum, testosterone has a positive effect on the striatal dopaminergic reward system, the same applies to cortisol, although one study reported a negative effect. Testosterone and cortisol might also lead to status-relevant behavior via neural threat systems. For example, testosterone administration is related to increased amygdala reactivity to angry faces (Hermans et al., 2008), but only if they are approached and not if they are avoided (Radke et al., 2015). Cortisol administration has an opposite effect and leads attention away from angry stimuli (Van Honk et al., 1998). Relatedly, amygdala responses to social threat were strongest in participants high on baseline testosterone and low on baseline cortisol (Hermans et al., 2008).

Functional cross-talk has been reported between the HPA- and HPG-axis (Viau, 2002). A first mechanism of cross-talk is that secretion by one axis can be inhibited by the other. For example, stress causes the secretion of corticotropin-releasing hormone (CRH), adrenocorticotropic hormone (ACTH) and glucocorticoids. This in turn suppresses HPG-axis functioning (Johnson et al., 1992), for example by affecting gonadotropin-releasing hormone and gonadotropin secretion (Tilbrook et al., 2000). Vice versa, testosterone inhibits HPA-axis functioning in the hypothalamus by for example reducing arginine vasopressin synthesis (Salvador, 2012; Viau, 2002). A second mechanism of cross-talk is at the receptor level, as activity in one axis can block receptors in the other. For example, increasing glucocorticoids decreases androgen receptor synthesis (Smith et al., 1985; Mehta et al., 2015a). Within axes, effects on receptors have also been reported, e.g. testosterone release related to winning enhances androgen sensitivity (Fuxjager et al., 2010) and HPA-axis functioning might be influenced via epigenetic effects on the glucocorticoid expression (McGowan et al., 2009). These findings have been corroborated by results from rat studies showing gonadal and adrenal cross-talk at the cell- and genome level (see Viau, 2002 for a review).

On the psychological level, cortisol is thought to block the influence of testosterone on status-relevant behavior (Sherman et al., 2016; Prasad et al., 2017), as cortisol is related to submission, social avoidance and withdrawal (Van der Wruisthen and Solms, 2015), whereas HPG-axis activity is related to approach behavior (Platje et al., 2015). Therefore, the stress system might overrule the gonadal functions (Sherman et al., 2016; Carré and Mehta, 2011; Liening and Josephs, 2010). This mechanism might be evolutionary informed, as it is beneficial to suppress reproductive activity in times of stress (Hamilton et al., 2015; Carré and Mehta, 2011). For example, a recent study showed that the relationship between testosterone and aggression in men was moderated by trait anxiety, with stronger effects when anxiety levels were low (Norman et al., 2015).

Empirically, the dual-hormone hypothesis is supported by studies using a broad range of status-relevant measures, such as status itself, dominance, risk taking, aggression and psychopathy (Mehta and Prasad, 2015). Although most published empirical studies provide evidence for the dual-hormone hypothesis, some published studies report null-findings (e.g. Geniole et al., 2013; Mazur and Booth, 2014) and even reversed patterns were reported (i.e., testosterone was related to status-relevant behavior for high but not low cortisol; Denson et al., 2013a, 2013b; Welker et al., 2014).

The current study aimed to test the dual-hormone hypothesis on status-relevant behavior using a meta-analytic approach. Five outcome measures were reviewed in the current study: status, dominance, risk taking, aggression and psychopathy. In the following, we review literature indicating that these outcome measures are all associated with status-relevant behavior, and literature indicating that these outcome measures are also related to testosterone. Therefore, one overarching meta-analysis on all these five outcome measures was performed. However, although there is this conceptual overlap between these outcome measures (Mehta and Prasad, 2015), there might be heterogeneity as well. Therefore, the five outcome measures were also investigated separately in a moderator analysis.

First, and most obvious, status (i.e. the position of an individual within a societal structure; Hollingshead, 1975) itself is associated with testosterone levels as the essential influence of testosterone on human behavior is searching for, and maintaining, social status (see Eisenegger et al., 2011 for a review). However, note that a recent meta-analysis found no relationship between basal testosterone levels and leadership, a concept that is closely related to status (Van der Meij et al., 2016).

Second, dominance is a personality trait that is characterized by an assertive and self-assured behavioral style (Anderson and Kilduff, 2009), mainly aimed at gaining social status (Mazur and Booth, 1998; Mehta and Josephs, 2010; Cheng et al., 2013). Indeed, dominant behavior also results in higher status. That is, dominant individuals are perceived as more competent (Anderson and Kilduff, 2009), which in turn strongly related to social status (Fiske et al., 2002). Although meta-analytical evidence is lacking, several studies indicate that dominance is positively related to testosterone (Mehta and Beer, 2010; Archer, 2006; Turan et al., 2014; Slater et al., 2011; Mazur and Booth, 1998; Van der Meij et al., 2008; Wirth and Schultheiss, 2007, see Mehta and Josephs, 2010 for an overview of studies). However, some exceptions have been reported. For example, one study reported little to no relationship between baseline testosterone and indicators of dominance, namely toughness (Van Bokhoven et al., 2006) and others showed that competitive behavior was not related to baseline testosterone (Carré and McCormick, 2008). Moreover, dominance and testosterone seem to be especially related when status is low (Josephs et al., 2003; Newman et al., 2005; Carré and McCormick, 2008; Mehta et al., 2008; Zypur et al., 2009). More specifically, according to the so-called mismatch effect, individuals with high testosterone levels and low status experience arousal and distress, and try to actively reduce this by increasing their dominant behavior (Josephs et al., 2006; Van der Meij et al., 2008).

Third, risk taking, defined as the engagement in behavior with a possibility of negative outcomes (Boyer, 2006), might be an evolutionary strategy to attain status (Buss, 2009; Ellis et al., 2012; Mehta et al., 2015a). That is, engagement in risky behavior leads to higher positions in social hierarchies and may therefore be advantageous for reproduction (Ellis et al., 2012). In addition, an important goal of risk taking is attaining status, especially in situations in which status is threatened (Scholer et al., 2010; Kwon et al., 2017; Van den Bos et al., 2013). A small but significant correlation was found between risk taking and testosterone (r = .12 in a recent meta-analysis; Kurath and Mata, 2018). This effect seemed robust and was not influenced by gender or type of outcome measure (Kurath and Mata, 2018). However, some studies reported null effects (e.g. Derntl et al., 2014; van der Loos et al., 2013) or even negative associations (van Anders et al., 2012; see Mehta et al., 2015a for an overview). The relationship between risk taking, status and testosterone is similar as described above for dominance: Only in situations of low perceived status, individuals high in testosterone show more risk taking (Ronay and Von Hippel, 2010).

Fourth, aggression, behavior with the intention to harm another (Baron and Richardson, 1994), is potentially aimed at attaining status (Baumeister et al., 1996; Faris and Felmlee, 2011; Li and Wright, 2014; Warburton et al., 2006). Whether aggression is actually a successful strategy to obtain status is subject to debate, as effects of aggression on status depend on factors like type of population (e.g. children or adults; students or delinquents) and type of aggression (e.g. physical or relational) (Ahn and Rodkin, 2014; Cillessen and Mayeux, 2004; Rose et al., 2004). Aggression is modestly correlated to testosterone (see Archer et al., 2005 for a meta-analysis, r = .08), and the correlation is
moderated by factors like gender, type of population (normal or delinquent), age and source of testosterone (saliva or blood). For example, the relationship between testosterone and aggression only occurred in adolescent boys with deviant peers and not in boys with nondeviant peers (Rowe et al., 2004).

Interestingly, these two lines of research focusing on (1) the relationship between dominance/risk taking/aggression and status and (2) the relationship between dominance/risk taking/aggression and testosterone are linked by theory on the challenge hypothesis. According to this hypothesis, originating in the animal literature (Wingfield et al., 1990), people respond with an increase in testosterone if their status is challenged. This increased testosterone facilitates status-relevant behavior like dominance, risk taking and aggression which in its turn increases the likelihood of regaining or enhancing status (Archer, 2006; Carré and Archer, 2018).

Fifth, psychopathy, which is a multifactorial construct characterized by callous-unemotional and grandiose-manipulative personality traits, often accompanied by immoral behavior and low levels of empathy (Zijlmans et al., 2018), is related to severe forms of aggression (see Frick and White, 2008 for a review). Aggression is, as described before, related to status and testosterone. Furthermore, psychopathy itself is moderately correlated to status-driven risk taking (i.e. risk taking that is driven by social aims; Visser et al., 2014). The relationship with testosterone is less clear for psychopathy than for other outcome measures, as mixed results are reported (Welker et al., 2014; Glenn et al., 2011; Stalenheim et al., 1998). However, note that antisocial behavior, aggression and low levels of empathy, which have been related to psychopathy (e.g. Zijlmans et al., 2018), are related to testosterone (Van Honk et al., 2011; Hermans et al., 2006). For these reasons, the overarching meta-analysis was performed twice, with and without studies on psychopathy.

In order to test the dual hormone hypothesis, it is crucial to test the interaction effect of testosterone and cortisol on behavior, simple slope analyses do not suffice (Nieuwenhuis et al., 2011; Glenn and Stern, 2006). However, many dual-hormone studies do present simple slope analyses. Therefore we do not only report the crucial meta-analytic test on the interaction effect, but also follow up simple slope meta-analyses: one investigating the relationship between testosterone and outcome measures when cortisol is low and one investigating the relationship between testosterone and outcome measures when cortisol is high.

In addition to these overall meta-analyses, three participant characteristics (i.e. sex, age, population type) were tested on their moderating influence on the effect size. First, sex differences might occur, as hormonal systems fundamentally differ across males and females. The specific role of sex as a moderator in the dual-hormone effect is yet unclear. Some studies suggest a more pronounced effect in males than in females (e.g. Welker et al., 2014), however, many other studies found no sex differences (see Mehta and Prasad, 2015 for an overview). Second, age differences might occur, as hormone levels change across development, e.g. testosterone rises in puberty (Grumbach, 2002), and from adulthood both testosterone and cortisol levels decline with age (Feldman et al., 2002; Sharma et al., 1989). However, we have no specific expectation how this might influence the dual-hormone effect. Third, the majority of studies investigated typical populations, whereas other studies tested delinquents. As psychopathology might influence the dual-hormone effect (Mehta and Prasad, 2015), the type of population is taken into account as a potential moderator.

To sum up, the current study investigated the dual-hormone hypothesis by performing a meta-analysis on the interaction effect of testosterone and cortisol on status, dominance, risk taking, aggression, and psychopathy. In line with the dual-hormone hypothesis, it was expected that meta-analyzing the interaction would yield a significant effect size. Largest effects were expected on measures of status, as this is the most direct operationalization of status-relevant behavior.

2. Methods

This meta-analytic review was conducted according to standards in Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Liberati et al., 2009).

2.1. Search strategy

A systematic literature search was conducted independently by two authors (BM, TJD). PsychINFO, Medline and Embase were searched for full English articles from 1946 until June 2017. Additionally, the first 200 relevant hits in Google Scholar were screened. Key search terms were: cortisol, testosterone, and different types of status-relevant behavior. The entire search syntax for each database can be found in the supplementary materials. Three very recent articles, published after the completion of the search, were added as suggested by the reviewers.

2.2. In- and exclusion criteria

There were three inclusion criteria for articles. First, the dependent variable had to be status-relevant behavior. We included studies on the following constructs: status, dominance, risk taking, aggression, and psychopathy. These constructs could be measured with tasks or with self- or other reported questionnaires. For status, measures included mainly status or popularity ratings by others (e.g. Mazur et al., 2015), but also more indirect measures like the number of subordinates (Sherman et al., 2016). For dominance, measures included questionnaires such as the Affective Neuroscience Personality Scale (Van der Westhuizen and Solms, 2015) as well as tasks such as the Dictator Game (Pfisterheicher, 2016b). For risk taking, measures included questionnaires such as the Self-Reported Risk Attitude Questionnaire (Barel et al., 2017) as well as tasks such as the Balloon Analogue Risk Task (Mehta et al., 2015a). For aggression, measures included aggression paradigms such as the Social Threat Aggression Paradigm (Buades-Rotger et al., 2016) and the Taylor Aggression Paradigm (Denson et al., 2013a, 2013b) and records of real-life aggression and aggression questionnaires such as the Buss-Durkee Hostility Inventory (Popma et al., 2007). For psychopathy, measures included self- and other-reported questionnaires such as the Hare Psychopathy Checklist (Glenn et al., 2011) or the Self-Report Psychopathy (Welker et al., 2014).

Second, the interaction between testosterone and cortisol had to be tested, with regard to the outcome measure, therefore articles that only used testosterone-cortisol ratio were excluded. Third, the interaction had to be between basal testosterone and basal cortisol, therefore studies that only investigated hormonal reactivity were excluded. Basal hormone measurements refer to measurements in rest, e.g. without the influence of an experimental manipulation or without the influence of a previous sports competition.

2.3. Study selection

A two-step strategy was used in assessing full text articles. First, titles and abstracts were screened for eligibility by two independent coders (BM, TJD). Second, the remaining articles were screened more thoroughly on the inclusion criteria, by the same independent coders. Articles not meeting the criteria were removed. The remaining articles were coded by one author (BM) and checked by another (TJD, JAvR,

---

2 Please note that there are similarities in the Methods section with a previous meta-analysis by the same authors on another topic (Dekkers et al., 2016).

3 Note that questionnaires measuring externalizing behavior were also coded as “aggression”, as many of the items of these broad scales measure aggressive behavior (e.g. Bongers et al., 2004). Ideally, only the questions on aggression would have been taken into account, but these raw data were not always available.
EW, HMH). Disagreement in any of the steps was solved by debate. If the test statistics of interest were not reported, these were requested from the authors of the included studies. This was done for 27 papers (4 papers reported all necessary information), with 26 positive responses regarding interaction and simple slopes effects, and 25 positive responses regarding main effects of testosterone and cortisol. Therefore, only one study was excluded because data were not available (Scerbo and Kolk, 1994).

2.4. Statistical analysis

A crucial test of the dual-hormone hypothesis requires a test of the interaction between testosterone and cortisol on the outcome measure (cf. Nieuwenhuis et al., 2011; Gelman and Stern, 2006). In this case, a negative interaction effect indicates evidence for the dual-hormone hypothesis. Usually, the dual-hormone hypothesis is also tested with simple slopes (i.e. the regression slope of Y regressed on X, conditioned on a single value of a moderator; Aiken et al., 1991). This means that status-relevant behavior is regressed on testosterone, conditioned on a single value of cortisol. In the current case, this single value is either one standard deviation above or one standard deviation below the mean cortisol level. A positive effect of testosterone on the outcome measure for low cortisol and the absence of such an effect for high cortisol is then taken as evidence for the dual-hormone hypothesis (Mehta et al., 2017).

In the current study, the interaction between testosterone and cortisol on the outcome measures was meta-analyzed. Furthermore, follow-up simple slope meta-analyses were performed: one for low cortisol levels and one for high cortisol levels. A meta-analysis of simple slopes has the disadvantage that what constitutes low cortisol and what constitutes high cortisol can differ substantially over studies. However, more sophisticated analyses that do not share this disadvantage would require raw data from every single study, to which we did not have access. Finally, main effects for testosterone and cortisol were also meta-analyzed.

For all analyses, the t-values and degrees of freedom (df) of effects were transformed into a correlation (Zakzanis, 2001):

$$ r = \frac{t}{\sqrt{t^2 + df}} $$  \hspace{1cm} (1)

However, because the variance of r depends strongly on r itself (Borenstein et al., 2009), it was transformed into Fisher’s z (Alexander et al., 1989; Lipsey and Wilson, 2001):

$$ z = 0.5 \times \ln(\frac{1 + r}{1 - r}) $$  \hspace{1cm} (2)

The variance of Fisher’s z is calculated using Eq. (3) (Field, 2001):

$$ \text{v}_z = \frac{1}{N - 3} $$  \hspace{1cm} (3)

Because Fisher’s z is harder to interpret than the correlation, it was transformed back into r for reporting (Borenstein et al., 2009). Thus, Fisher’s z was used in the analyses whereas the correlation is reported. Cohen (1992) was followed for the interpretation of the correlation effect sizes, with values of 0.10 indicating small, values of 0.30 indicating medium, and values of 0.50 indicating large effect sizes.

In order to account for between study variance, a random effects meta-regression was performed. As some studies contained multiple relevant effect sizes from multiple samples (e.g. men and women separately), the random effects meta-regression consisted of three levels: participants, effect sizes, and studies (see Assink and Wibbelink, 2016; Konstantopouls, 2011; van den Noortgate et al., 2013; Cheung, 2014 for a more elaborative background on multilevel meta-analysis). In order to estimate the overall effect size, the random effects meta-regression was first performed without moderators. Between study heterogeneity is reported by means of the QE test for residual heterogeneity, and by I^2, an estimate of the “proportion of total variation in study estimates that is due to heterogeneity” (Higgins and Thompson, 2002; cf. Geniole et al., 2017). Typical I^2 analyses result in one I^2 statistic; the current study has two I^2 statistics (between and within study heterogeneity) because of the three-level structure (also see metafor R-package; Viechtbauer, 2010). For the moderator analyses, each potential moderator was added separately. For the moderator “outcome measure”, a moderation analysis yields estimates of the effect size for a reference outcome measure (β̂) and for the difference in effect size between the reference outcome measure and another outcome measure (Δβ̂). For example, it yields estimates of the effect size for status (β̂, reported and interpreted as r) and for the difference between status and dominance (Δβ̂, reported as Δr). By repeatedly running this analysis with different reference outcome measures, we obtained estimators for all outcome measures, and for differences between all outcome measures. A similar approach was used for the other moderators.

The analyses were performed in R (metafor package; Viechtbauer, 2010). The entire code, as well as the input file, can be found on OSF (https://osf.io/v5bhj/).

2.5. Multiple effect sizes

Several studies in the meta-analysis used multiple relevant dependent variables and therefore contain more than one effect size based on the same participants. To account for dependency between those effect sizes, we used the average t-value of the different outcome variables as input for Eqs. (1)–(3) (Cheung, 2014). As a robustness check, analyses on all original effect sizes, that thus did not account for dependency, were also reported.

One study (Denson et al., 2013a, 2013b) contained an effect size for measures of aggression and dominance, using the same participants. In the overall meta-analysis this was averaged into one effect size, however for the moderator analysis on the type of outcome measures, the effect size of dominance and the effect size of aggression were separately taken into account, and therefore k = 50 instead of k = 49 for these analyses.

2.6. Robustness checks

Studies with significant effects might be more likely to get published than studies with null results (Easterbrook et al., 1991). In case of such a publication bias, the estimated overall effect size will be an overestimation of the true effect size. Flexibility in data-analysis may also lead to an overestimation of the true effect (Simmons et al., 2011). Therefore, several analyses were performed to check robustness of the effects:

- Egger’s test was used to assess potential funnel plot asymmetry, which might be indicative of publication bias (Egger et al., 1997).
- The trim and fill method was used to see how many studies are needed to counter potential funnel plot asymmetry brought by publication bias (Duval and Tweedie, 2000).
- The leave-one-out method was used to check influences of single effect sizes and single studies. These analyses were performed twice, once leaving out single effect sizes and once leaving out single studies, the latter potentially containing multiple effect sizes. The minimum and maximum effect sizes were reported, as well as the effect size when leaving out one particularly large study (Mazur and Booth, 2014).
- A p-curve was calculated to check whether significant findings indicate a true effect, that is, contain evidential value (Simonsohn et al., 2015). In a p-curve, p-values below .05 are plotted on the x-axis and the percentage of studies yielding such a p-value are plotted on the y-axis. Right skewed p-curves are indicative of evidential value, as true effects tend to be highly significant. Flat p-curves...
indicate the absence of evidential value and left skewed p-curves are indicative of flexibility in data-analysis, as these contain many p-values just below .05. P-curves were created using the p-curve app 4.06 (http://www.p-curve.com/app4, 2018http://www.p-curve.com/app4/). Simonsohn et al., (2015) argued that evidential value in a set of studies requires that the curve for p-values lower than .025, the so-called half p-curve, is right skewed (p-value test for right skew < .05), or that the p-value of this test of right skew is < .10 for both the full and half p-curve. If the conclusion is that the set of studies does not contain evidential value, a follow-up test is performed to test whether the set of studies has insufficient power to detect evidential value (Simonsohn et al., 2014). If this test (again a test on right skew of the p-curve) is significant (p < .05), the conclusion is that the set of studies lacks evidential value. If not, the conclusion is that the number of significant p-values is too low to make inferences about evidential value. The p-curve disclosure table can be found in the supplementary materials.

- The test of excessive significance (Ioannidis and Trikalinos, 2007) is used as another measure of potential publication bias and/or flexibility in data-analysis. In this test, the observed number of significant studies is compared to the expected number of significant studies, in which this expected number is based on the estimated overall effect-size and the power to detect this effect in the individual studies.

- PET/PEESE (Precision Effect Test/Precision Effect Estimation with Standard Error; Stanley and Doucouliagos, 2014) analyses were used to check the robustness of the effect. PET/PEESE regresses effect sizes on standard error/variance, and predicts the effect size for a standard error/variance of zero. Whether the PET or PEESE result should be reported depends on whether the PET result is significant (Carter et al., 2015; Stanley and Doucouliagos, 2014): If PET is not significant PET should be reported. Note that the PET/PEESE method has been criticized, especially in situations with small samples and high heterogeneity between studies (Stanley, 2017; see http://datacolada.org/59, 2018http://datacolada.org/59 for criticism and response).

- To check whether there was more evidence for the null hypothesis (i.e. no interaction effect) than for the alternative hypothesis of an interaction effect, a Bayesian random effects meta-analysis was also performed in R, using the metaBMA package (Heck et al., 2017). The effect was tested two-sided, the default prior on the between study variance of the effect sizes was used, multiple priors on the Ha were used as a prior sensitivity check, and Bayes factors (BFs) were computed.

- As psychopathy might be less strongly related to status than the other outcome measures, the meta-analysis was also performed without effect sizes on psychopathy.

- Dependency between effect sizes was accounted for by taking the average t-value of the outcome measures if these outcome measures were administered in the same participants. However, the meta-analysis was also performed without taking potential dependency into account, i.e. by meta-analyzing all 58 effect sizes.

- In the current meta-analysis, some t-values of the interaction were already reported in the papers or were calculated directly from reported values, whereas other t-values were provided by the authors. Also, some models contained additional covariates, whereas others did not. The potential influence of (a) whether t-values were reported or provided by the authors and (b) the addition of covariates to the model, was checked in categorial moderator analyses (also see Table 1). A study was coded as “with covariates” if the difference between N and df was larger than 4, as a model without covariates would yield df = N – k – 1, with k = 3 (testosterone, cortisol, interaction).

- If there is less variation in the independent variables, main or interaction effects may become less pronounced. To check for this possibility, we computed the coefficient of variation (i.e. the standard deviation divided by the mean) of both testosterone and cortisol for each study. We then included these two coefficient of variation variables in two separate moderator analyses.

2.7. Moderators

Potentially relevant participant characteristics that were taken into account were (1) Sex (categorically coded; in case of mixed sex samples, separate data for males and females were requested from the authors); (2) Average age (continuously coded) and (3) Type of population (typical population or delinquents; categorically coded). Finally, the type of measurement (tasks, self-report questionnaires, other-report questionnaires) was taken into account as moderator.

3. Results

3.1. Literature search

Using aforementioned criteria, we found 30 papers, with 33 studies, which contain 58 relevant effect sizes regarding the testosterone by cortisol interaction, containing 8538 participants (see Table 1 for study characteristics). For the corresponding PRISMA flowchart (Moher et al., 2009), see Fig. 1.

3.2. Main analysis: meta-analysis on testosterone x cortisol interaction

The effect of the testosterone x cortisol interaction on all outcome measures together (k = 49, n = 8538) was significant: r = .061, p = .026, 95% CI [-.11, -.01] (see Fig. 2 and Table 2). According to Cohen (1992), an effect of .10 is considered small. Conforming to Higgins et al., (2003) heterogeneity was high (total I2 = 71.6%, between-study F = 41.1%, within-study F = 67.4%).

Moderator analyses testing the effect of type of outcome measure yielded no significant results (also see Fig. 2 at the bottom and Table 3): there were no significant differences between the effect sizes of the different outcome measures. Analyses on the different outcome measures yielded no significant results: Status, k = 5, r = -.158, p = .125, 95% CI [-.35, .04]; Dominance, k = 9, r = -.031, p = .697, 95% CI [-.19, .13]; Risk Taking, k = 13, r = -.099, p = .129, 95% CI [-.22, .03]; Aggression, k = 16, r = -.043, p = .430, 95% CI [-.15, .06]; Psychopathy, k = 7, r = .015, p = .859, 95% CI [-.15, .18]. Note that, as expected, the effect size for the direct measures of status was largest, although it did not deviate significantly from any of the other outcome measures. 4

3.3. Robustness checks for the interaction effect

3.3.1. Funnel plot asymmetry 5

Egger’s test (Egger et al., 1997) indicated no funnel plot asymmetry (z = .139, p = .889), and the trim and fill method (Duval and Tweedie, 2000) indicated that one study was missing on the left side of the funnel plot (see Fig. 3). Correcting the effect size for this hypothetical study resulted in a very similar effect size: r = -.067, p = .016. These analyses thus do not indicate publication bias.

3.3.2. Sensitivity analyses

Sensitivity analyses using the leave-one-out method were performed in order to investigate influences of single effect sizes. With a minimum

4 Note that for separate meta-analyses on the different outcome variables, heterogeneity is still significant, as the QE test for residual heterogeneity was significant for all 5 outcome variables.

5 Note that it is currently not possible to perform Egger’s test and the trim and fill method while modeling dependency of effect sizes. Therefore, we performed these analyses assuming independence of effect sizes.
Table 1
Characteristics, moderators and effect sizes of included studies. Note that if df was not reported, this was calculated using $df = n - k - 1$; with $k = 3$ (cortisol, testosterone, interaction). Footnotes attached to the first author’s name indicate whether statistics were reported in the paper (“P”) or provided by the authors (“A”), and whether the model contained additional covariates above the intercept, $T$, $C$ and $T \times C$ (“Y”) or not (“N”). See Methods section for more explanation of effect size measures. Abbreviations: $C =$ Cortisol; $df =$ degrees of freedom; $F =$ female; $M =$ male; $\mu =$ mean; $T =$ Testosterone.

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>N</th>
<th>Moderators</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M/F</td>
<td>$\mu$ Age (sd)</td>
<td>Population</td>
</tr>
<tr>
<td>Barel et al. (2017)</td>
<td>37</td>
<td>M 26.2 (2.1)</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>F 25.4 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Busdes-Rotger et al. (2016)</td>
<td>39</td>
<td>F 23.2 (3.2)</td>
<td>Typical</td>
</tr>
<tr>
<td>Cueva et al. (2015)</td>
<td>66</td>
<td>M 21.9 (2.9)*</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Dabbs et al. (1991)</td>
<td>113</td>
<td>M 17-18 (NA)</td>
<td>Delinquent</td>
</tr>
<tr>
<td>Denton et al. (2013a)</td>
<td>49</td>
<td>F 20.7 (6.1)</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>F 18-23 (NA)</td>
<td>Typical</td>
</tr>
<tr>
<td>Gleno et al. (2011)</td>
<td>33</td>
<td>M 20.0 (2.4)*</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Geniole et al. (2011)</td>
<td>104</td>
<td>M 20.0 (2.5)*</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Glenn et al. (2011)</td>
<td>150</td>
<td>M 36.5 (8.8)*</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Grotzinger et al. (2018)</td>
<td>134</td>
<td>M</td>
<td>Typical</td>
</tr>
<tr>
<td></td>
<td>93</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Mazur and Booth (2014)</td>
<td>4462</td>
<td>M 38.3 (2.5)</td>
<td>Veterans</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
| First author (year) | N | Measure | DV | Moderators Effects | T main effect Testosterone (df) | Fishers z main effect Testosterone (se) | T main effect Cortisol (df) | Fishers z main effect Cortisol (se) | T Interaction (df) | Fishers z Interaction (se) | T Low C (df) | Fishers z Low C (se) | T High C (df) | Fishers z High C (se) |
|---------------------|---|---------|----|---------------------|-------------------------------|-------------------------------------|-------------------------------|-----------------------------------|----------------||-----------------------|-------------|------------------|-------------|------------------|
| Mazur et al. (2015) study 1<sup>st</sup> | 30 | M 18-26 (NA) | Typical | Status rank | .93 (21.2) | .20 (.19) | −1.37 (23.89) | −.28 (.19) | 1.96 (24.7) | .38 (.19) | −.80 (24.7) | −.16 (19) | 2.28 | (24.7) | .44 (.19) |
| Mazur et al. (2015) study 2<sup>nd</sup> | 15 | M | Status rank | .76 (5.23) | .33 (.29) | −.16 (9.84) | −.05 (.29) | .46 (8.8) | .15 (.29) | −.03 (8.8) | .01 (29) | .62 | (8.8) | .21 (.29) |
| Mehta and Josephs (2010) study 2<sup>nd</sup> | 23 | M | Decision whether to compete again | Typical | .53 (22) | −.11 (21) | −.84 (22) | −.18 (21) | 1.02 (22) | .22 (21) | −.97 (22) | −.21 (21) | .54 | (22) | .11 (21) |
| Mehta et al. (2015a) study 1<sup>st</sup> | 51 | M 21.2 (3.3) | Typical | Sensation Seeking Scale (SSS) | 2.17 (55) | .29 (13) | −.17 (55) | −.23 (13) | −.21 (55) | −.29 (13) | 2.87 (55) | .38 | (13) | .08 | (13) |
| Mehta et al. (2015a) study 2<sup>nd</sup> | 50 | M (20.22) | Risk questions | Informant risk taking | −.14 (37) | −.02 (16) | −.07 (37) | −.16 (16) | −.16 (37) | −.26 (16) | .93 (37) | −.15 | (37) | −.15 | (37) |
| Mehta et al. (2015a) study 2<sup>nd</sup> | 51 | M 20.0 (2.22) | Typical | Balloon Analogue Risk Task (BART) | .65 (156) | .05 (88) | .62 (156) | .05 (88) | .62 (156) | .05 (88) | 1.97 (156) | .16 | (88) | .42 | (88) |
| Mehta et al. (2017) | 42 | M 19.9 (2.9) | Typical | Hawk Dove Game | 1.50 (22.7) | .31 (16) | .95 (23.1) | .20 (16) | .75 (22.3) | .16 (16) | .37 (23.0) | .06 | (16) | .39 | (16) |
| Pfattheicher et al. (2014) | 72 | M 22.6 (2.7) | Typical | Anti-social punishment game | 1.02 | (67) | .12 | (60) | −.07 | (60) | −.19 | (67) | .26 | (60) | .24 | (67) |
| Pfattheicher (2016a) | 128 | M 22.0 (4.0) | Typical | Short Dark Triad (SD3) | −1.68 (124) | −.15 (99) | −1.78 (124) | −.16 (99) | 1.79 (124) | .16 (99) | −1.45 (124) | −.13 | (99) | 1.41 | (99) |
| Pfattheicher (2016b) | 153 | M 22.1 (4.4) | Typical | Dictator game | 1.56 (146) | .13 (86) | −.35 (146) | −.03 (86) | −.29 | (146) | .27 | (86) | 2.79 | (146) | .23 | (86) |
| Platje et al. (2015) | 148 | M 17.0 (4.9) | Typical | Youth Self Report (YSR) | .68 | (144) | .06 | (88) | −.19 | (144) | .07 | (88) | .07 | (144) | .06 | (88) |
| Ponzi et al. (2016) | 40 | M 27.9 (4.9) | Typical | Cognitive Social Structures (CSS) | .91 | (35) | .15 | (35) | .13 | (35) | .23 | (35) | .35 | (35) | .24 | (35) |
| Pomp et al. (2007) | 95 | M 13.7 (7) | Delinquent | Rass-Durkee Hostility Inventory (BDHI) | −.37 | (91) | −.04 | (10) | −.56 | (91) | −.10 | (10) | −.46 | (91) | −.15 | (10) | .52 | (91) |

(continued on next page)
<table>
<thead>
<tr>
<th>First author (year)</th>
<th>N</th>
<th>M/F</th>
<th>μ Age (sd)</th>
<th>Population Measure</th>
<th>DV</th>
<th>T main effect Testosterone (df)</th>
<th>Fisher's z main effect Testosterone (se)</th>
<th>T main effect Cortisol (df)</th>
<th>Fisher's z main effect Cortisol (se)</th>
<th>T Interaction (df)</th>
<th>Fisher's z Interaction (se)</th>
<th>T Low C (df)</th>
<th>Fisher's z Low C (se)</th>
<th>T High C (df)</th>
<th>Fisher's z High C (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ronay et al. (2017)</td>
<td>44</td>
<td>M</td>
<td>22.1 (2.6)</td>
<td>Typical Gambling task</td>
<td>Risk Taking</td>
<td>-0.04 (40)</td>
<td>-0.01 (16)</td>
<td>0.02 (40)</td>
<td>0.003 (16)</td>
<td>0.35 (40)</td>
<td>0.05 (16)</td>
<td>0.27 (40)</td>
<td>0.04 (16)</td>
<td>0.25 (40)</td>
<td>0.04 (16)</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>F</td>
<td>21.1 (2.2)</td>
<td>Typical Balloon Analogue Risk Task (BART)</td>
<td>Risk Taking</td>
<td>-0.67 (60)</td>
<td>-0.09 (13)</td>
<td>-0.14 (60)</td>
<td>-0.02 (13)</td>
<td>-0.22 (60)</td>
<td>-0.03 (13)</td>
<td>-0.27 (60)</td>
<td>-0.04 (13)</td>
<td>-0.60 (60)</td>
<td>-0.08 (13)</td>
</tr>
<tr>
<td>Ronay et al. (2018)</td>
<td>42</td>
<td>M</td>
<td>22.09 (3.12)</td>
<td>Typical Questions about no. of subordinates</td>
<td>Status</td>
<td>2.36 (38)</td>
<td>0.37 (16)</td>
<td>0.55 (38)</td>
<td>0.09 (16)</td>
<td>-0.82 (38)</td>
<td>-0.44 (16)</td>
<td>0.13 (38)</td>
<td>-0.32 (16)</td>
<td>-0.47 (38)</td>
<td>0.16 (16)</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>F</td>
<td>21.86 (2.79)</td>
<td>Typical Child Behavior Checklist (CBCL)</td>
<td>Externalizing behavior Parent report</td>
<td>-1.13 (93)</td>
<td>-0.12 (10)</td>
<td>0.70 (93)</td>
<td>0.07 (10)</td>
<td>1.14 (93)</td>
<td>0.12 (10)</td>
<td>-0.95 (93)</td>
<td>-0.10 (10)</td>
<td>-0.76 (93)</td>
<td>0.08 (10)</td>
</tr>
<tr>
<td>Sherman et al. (2016)</td>
<td>78</td>
<td>M</td>
<td>48.7 (5.9)</td>
<td>Typical Questions about no. of subordinates</td>
<td>Status</td>
<td>1.42 (74)</td>
<td>0.16 (12)</td>
<td>0.40 (74)</td>
<td>0.04 (12)</td>
<td>-2.89 (74)</td>
<td>-0.33 (12)</td>
<td>3.20 (74)</td>
<td>0.36 (12)</td>
<td>-0.54 (74)</td>
<td>-0.12 (12)</td>
</tr>
<tr>
<td>Smith and Apicella (2017)</td>
<td>247</td>
<td>M</td>
<td>22.0 (5.6)</td>
<td>Typical Multiple Price List (MPL)</td>
<td>Economic risk taking</td>
<td>-0.16 (243)</td>
<td>-0.01 (6)</td>
<td>0.32 (243)</td>
<td>0.02 (6)</td>
<td>0.61 (243)</td>
<td>0.04 (6)</td>
<td>-0.47 (243)</td>
<td>0.03 (6)</td>
<td>-0.22 (243)</td>
<td>0.01 (6)</td>
</tr>
<tr>
<td>Tackett et al. (2014)</td>
<td>47</td>
<td>M</td>
<td>16.0 (1.3)</td>
<td>Typical Child Behavior Checklist (CBCL)</td>
<td>Externalizing behavior Parent report</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-1.22 (41)</td>
<td>-0.19 (15)</td>
<td>1.55 (41)</td>
<td>0.24 (15)</td>
<td>-0.33 (41)</td>
<td>-0.05 (15)</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>F</td>
<td>16.0 (1.3)</td>
<td>Typical Child Behavior Checklist (CBCL)</td>
<td>Externalizing behavior Parent report</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-0.75 (53)</td>
<td>-0.05 (13)</td>
<td>-0.01 (53)</td>
<td>-0.001 (13)</td>
<td>-0.59 (53)</td>
<td>-0.08 (13)</td>
</tr>
<tr>
<td>Van den Bos et al. (2013)</td>
<td>23</td>
<td>M</td>
<td>24.11 (10.4)</td>
<td>Typical Auction Overbidding (Risk Taking)</td>
<td>Overbidding</td>
<td>2.16 (18)</td>
<td>0.49 (22)</td>
<td>-1.78 (18)</td>
<td>-0.41 (22)</td>
<td>-3.19 (18)</td>
<td>-0.69 (22)</td>
<td>2.61 (18)</td>
<td>0.72 (18)</td>
<td>1.52 (10)</td>
<td>0.46 (35)</td>
</tr>
<tr>
<td>Van der Westhuizen and Solms (2015)</td>
<td>36</td>
<td>M</td>
<td>21.6 (2.9)</td>
<td>Typical Affective Neuroscience Personality Scales (ANPS) – adjusted</td>
<td>Dominance</td>
<td>-0.80 (32)</td>
<td>-0.14 (17)</td>
<td>-0.44 (32)</td>
<td>-0.08 (17)</td>
<td>0.56 (32)</td>
<td>0.10 (17)</td>
<td>-1.18 (32)</td>
<td>0.21 (17)</td>
<td>-0.03 (32)</td>
<td>-0.005 (17)</td>
</tr>
<tr>
<td>Welker et al. (2014)</td>
<td>95</td>
<td>M</td>
<td>21.7 (4.7)</td>
<td>Typical Self-Report Psychopathy (SRP-SF)</td>
<td>Psychopathy</td>
<td>0.01 (90)</td>
<td>0.001 (10)</td>
<td>2.39 (90)</td>
<td>0.25 (10)</td>
<td>2.19 (90)</td>
<td>0.23 (10)</td>
<td>-1.21 (90)</td>
<td>0.13 (10)</td>
<td>1.87 (90)</td>
<td>0.20 (10)</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>F</td>
<td>24.11 (10.4)</td>
<td>Typical Self-Report Psychopathy (SRP-SF)</td>
<td>Psychopathy</td>
<td>1.58 (85)</td>
<td>0.17 (111)</td>
<td>-0.11 (85)</td>
<td>-0.01 (111)</td>
<td>-0.70 (85)</td>
<td>-0.08 (111)</td>
<td>1.61 (85)</td>
<td>0.17 (85)</td>
<td>0.87 (10)</td>
<td>0.09 (111)</td>
</tr>
<tr>
<td>Zillioli et al. (2015)</td>
<td>315</td>
<td>M</td>
<td>28.73 (0.13)</td>
<td>Typical Davis's Interpersonal Reactivity Index</td>
<td>Trait empathy (rescored)</td>
<td>-0.33 (311)</td>
<td>-0.02 (06)</td>
<td>-1.81 (311)</td>
<td>-0.10 (06)</td>
<td>-2.40 (311)</td>
<td>-0.14 (06)</td>
<td>1.36 (311)</td>
<td>0.08 (06)</td>
<td>-2.07 (311)</td>
<td>0.12 (06)</td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>F</td>
<td>27.42 (0.2)</td>
<td>Typical Davis's Interpersonal Reactivity Index</td>
<td>Trait empathy (rescored)</td>
<td>0.72 (138)</td>
<td>0.06 (08)</td>
<td>-1.16 (138)</td>
<td>-0.10 (08)</td>
<td>-1.00 (138)</td>
<td>-0.01 (08)</td>
<td>0.48 (138)</td>
<td>0.04 (08)</td>
<td>0.50 (138)</td>
<td>0.04 (08)</td>
</tr>
</tbody>
</table>
of $r = -0.054$ ($p = 0.040$), and a maximum of $r = -0.071$ ($p = 0.023$), the influence of single effect sizes seemed moderate. The influence of single studies, potentially containing multiple effect sizes, was similar, with a minimum of $r = -0.052$ ($p = 0.057$) and a maximum of $r = -0.070$ ($p = 0.011$). As can be seen from these analyses, the significance of the interaction effect can change based on a single study, which warrants caution regarding the results.

As the meta-analysis contained one study with a particularly large sample size (Mazur and Booth, 2014), analyses were also performed without this study. The overall interaction effect was similar: $r = -0.063$ ($p = 0.028$). Thus, omitting this study from the analysis does not change the results.

3.3.3. P-curve

To check the evidential value of the study and to check for flexibility in data-analysis, a p-curve was created (see Fig. 4; Simonsohn et al., 2015). Out of the 49 effect sizes, 14 were statistically significant. From these 14 significant effect sizes, 9 had p-values < .025. The p-curve analysis on the interaction effect did indicate that no evidential value was present (full p-curve $z = -1.32, p = .09$; half p-curve $z = -3.5, p = .36$). However, this was likely due to a lack of power to detect evidential value (full p-curve $z = -.89, p = .19$).

3.3.4. Test of excessive significance

The exploratory test for excess of significance (Ioannidis and Trikalinos, 2007) was used to check if the number of significant results differed from the expected number. Based on the overall effect and the sample size in each study, 5.1 significant effect sizes were expected, and 14 were found in the current meta-analysis. This difference was significant: $X^2 (1) = 17.18, p < .001$. Furthermore, this analysis showed that, except for one very large study (power = .98), all single studies included in the meta-analysis were underpowered given the overall effect size of the meta-analysis and sample size in each study (.05 < power < .20).

3.3.5. Precision Effect Test/Precision Effect Estimation with Standard Error (PET/PEESE)

According to Stanley and Doucouliagos (2014), we report the PET result, as it was not significant: the estimated PET effect size was $r = -0.022, p = .407$. This indicates that the true effect size of the dual-hormone effect might be lower than reported in the main analyses of this study.

3.3.6. Bayesian meta-analysis

The current meta-analysis was also conducted using Bayesian
methods. A two-sided Bayesian meta-analysis showed that there was more support for the null hypothesis (i.e. there is no interaction effect of testosterone × cortisol), BF$_{01} = 2.91$, than for the alternative hypothesis, BF$_{10} = .34$. Changing the prior (with a Cauchy prior width of $r = .25$ instead of the default of $r = 1/\sqrt{2}$) leads to BF$_{01} = 1.09$, BF$_{10} = .92$. In both situations, the BF is thus in favor of the null hypothesis.

Fig. 2. Forest plot of effect sizes on the interaction between testosterone and cortisol on status-relevant behavior. Negative effect sizes indicate evidence for the dual-hormone hypothesis. Note that these effect sizes are in terms of Fisher’s $z$, which slightly deviate from the reported correlations ($r$). Colors match corresponding outcome measures. * $p < .05$; ** $p < .01$.

Table 2

Test statistics for overall meta-analysis on the interaction effect of testosterone and cortisol on status-relevant behavior. $r$ is the effect size (correlation), $\sigma^2_{\text{estimate}}$ is the variance between effect sizes, $\sigma^2_{\text{study}}$ is the variance between studies, QE is a test for residual heterogeneity and $I^2$ is “the proportion of total variation in study estimates that is due to heterogeneity” (Higgins and Thompson, 2002). Abbreviations: $k =$ number of effect sizes, 95% CI = confidence interval of mean effect size over studies, $df =$ degrees of freedom, dep. = dependency. * $p < .10$; * * $p < .05$; ** $p < .01$; *** $p < .001$. 

<table>
<thead>
<tr>
<th>Effect size</th>
<th>$z$ [95% CI]</th>
<th>$r$</th>
<th>$\sigma^2_{\text{study}}$</th>
<th>$\sigma^2_{\text{estimate}}$</th>
<th>$\text{QE (df)}$</th>
<th>$I^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>-0.30 [-0.54, -0.07]</td>
<td>.38</td>
<td>.01</td>
<td>.76</td>
<td>.01</td>
<td>.17</td>
</tr>
<tr>
<td>Interaction (without psychopathy)</td>
<td>-0.15 [-0.41, 0.72]</td>
<td>.35</td>
<td>.68</td>
<td>-.03</td>
<td>.35</td>
<td>.68</td>
</tr>
<tr>
<td>Interaction (with controlling dep)</td>
<td>-0.33 [-0.56, -0.10]</td>
<td>.35</td>
<td>.00</td>
<td>.64</td>
<td>.35</td>
<td>.00</td>
</tr>
<tr>
<td>Low Cortisol</td>
<td>-0.24 [-0.53, 0.05]</td>
<td>.25</td>
<td>.05</td>
<td>.05</td>
<td>.25</td>
<td>.05</td>
</tr>
<tr>
<td>Low Cortisol (without psychopathy)</td>
<td>-0.21 [-0.37, 0.06]</td>
<td>.26</td>
<td>.04</td>
<td>.02</td>
<td>.26</td>
<td>.04</td>
</tr>
<tr>
<td>High Cortisol</td>
<td>-0.20 [-0.07, 0.47]</td>
<td>.18</td>
<td>.05</td>
<td>.02</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td>High Cortisol (without psychopathy)</td>
<td>-0.18 [-0.36, 0.01]</td>
<td>.16</td>
<td>.00</td>
<td>.31</td>
<td>.16</td>
<td>.00</td>
</tr>
<tr>
<td>High Cortisol (without controlling dep.)</td>
<td>-0.17 [-0.36, 0.01]</td>
<td>.16</td>
<td>.00</td>
<td>.31</td>
<td>.16</td>
<td>.00</td>
</tr>
</tbody>
</table>
hypothesis. According to the interpretation guidelines of Bayes factors (Jarosz and Wiley, 2014), the evidence for the null hypothesis is anecdotal (this applies to both priors).

### 3.3.7. Analyses without psychopathy-studies

As it might be argued that psychopathy is not related to status-relevant behavior, we checked its influence by performing the meta-analysis again without all studies using psychopathy as an outcome measure (i.e. \( k = 7 \)). The effect of the testosterone x cortisol interaction on status-relevant behavior (\( k = 42 \)) slightly increased: \( r = -0.076, p = 0.018, 95\% \text{ CI } [-0.13, -0.01] \) (see Table 2).

### 3.3.8. Analyses without controlling for dependency between participants

In the main analysis, dependency between participants (i.e. one sample delivered more than one outcome) was controlled for by taking the average of these outcomes. If all effect sizes were analyzed, without taking this dependency into account, the effect was highly similar, but no longer significant: \( k = 58, r = -0.064, p = 0.050, 95\% \text{ CI } [-0.13, 0.00] \).
3.3.9. Influence of reporting

In a categorical moderator analysis, we tested whether there was a difference between effect sizes reported in papers ($k = 10$) and those provided by authors ($k = 39$). This did not make a difference: $\Delta r = -.048$, $p = .505$.

3.3.10. Influence of additional covariates

The difference between studies with ($k = 15$) and without ($k = 34$) additional covariates was not significant, $\Delta r = -.010$, $p = .870$.

3.3.11. Influence of coefficient of variation

In two separate continuous moderator analyses, we tested whether the coefficient of variation of both testosterone and cortisol influenced the effect size of the interaction. This was not the case (for testosterone, $k = 40$, $\Delta r = .056$, $p = .769$; for cortisol, $k = 38$, $\Delta r = -.102$, $p = .249$).

3.3.12. Summary of tests for robustness

These robustness analyses indicate that caution is warranted in the interpretation, as (a) significance of the overall effect depended on single studies and effect sizes; (b) p-curve analysis did not indicate evidential value because of a lack of power; (c) there are more significant findings than expected; (d) power of the individual studies is low; (e) the true effect might be smaller according to PET/PEESE analyses; (f) the data seem more in favor of the null hypothesis according to a Bayesian meta-analysis. The robustness analyses also indicate that no differences were found between (a) analyses with and without psychopathy as outcome measure; (b) effects that were reported in the original article and effects that were requested from the authors and (c) effects from models with vs. without additional...
covariates.

3.4. Explorative moderator analyses

Four moderators were tested (see Table 4). A categorical moderator analysis showed that the overall effect size for men \( r = -.089, p = .008, 95\% CI [1.15, .02] \) was significant, whereas this was not the case for women \( r = -.008, p = .853, 95\% CI [-.10, .08] \). However, this difference between men and women was not significant \( (\Delta r = .081, p = .153, 95\% CI [-.03, .19]) \).

A continuous moderator analysis found no moderating influence of age on the effect size: \( \Delta r = .003, p = .436, 95\% CI [-.01, .01] \). No difference was found between studies investigating typical populations and studies investigating delinquents \( (\Delta r = .144, p = .248, 95\% CI [.37, .10]) \). However, it must be noted that the number of effect sizes on delinquents was very low.

Finally, no difference was found for different measurement types: comparisons between tasks and self-report questionnaires \( \Delta r = .008, 95\% CI [-.15, -.02] \) was significant, whereas this was not the case for low cortisol \( n = 8469 \) when cortisol is low yield a significant effect size: \( r = .568, 95\% CI [-.15, .08] \) (Table 2). According to the guidelines of Cohen (1992), this is a small effect. This is in line with the dual-hormone hypothesis, stating that there is a relationship between testosterone and status-relevant behavior when cortisol is low. Heterogeneity was high (total \( I^2 = 75.4\% \), between-study \( I^2 = 7.3\% \), within-study \( I^2 = 68.2\% \); Higgins et al., 2003).

Moderator analyses testing the effect of type of outcome measure yielded no significant results (Table 5): there were no significant differences between the effect sizes of the different outcome measures. However, the effect size of studies with Status as outcome measure was significant, \( k = 5, r = .229, p = .024, 95\% CI [.03, .41] \). Effect sizes on Dominance, Risk taking, Aggression and Psychopathy were not significant (see Table 5 for statistics).

3.5. Follow-up meta-analysis 1: Testosterone and status-relevant behavior for low cortisol

The relationship between testosterone and status-relevant behavior \( k = 49, n = 8469 \) when cortisol is low yielded a significant effect size: \( r = .063, p = .035, 95\% CI [-.08, .12] \) (Table 2). According to the guidelines of Cohen (1992), this is a small effect. This is in line with the dual-hormone hypothesis, stating that there is a relationship between testosterone and status-relevant behavior when cortisol is low. Heterogeneity was high (total \( I^2 = 75.4\% \), between-study \( I^2 = 7.3\% \), within-study \( I^2 = 68.2\% \); Higgins et al., 2003).

Moderator analyses testing the effect of type of outcome measure yielded no significant results (Table 5): there were no significant differences between the effect sizes of the different outcome measures. However, the effect size of studies with Status as outcome measure was significant, \( k = 5, r = .229, p = .024, 95\% CI [.03, .41] \). Effect sizes on Dominance, Risk taking, Aggression and Psychopathy were not significant (see Table 5 for statistics).

3.6. Follow-up meta-analysis 2: Testosterone and status-relevant behavior for high cortisol

The relationship between testosterone and status-relevant behavior \( k = 49, n = 8471 \) when cortisol was high yielded a non-significant effect size: \( r = .006, p = .777, 95\% CI [-.03, .05] \) (see Table 2). This is in line with the dual-hormone hypothesis, stating that there is no relationship between testosterone and status-relevant behavior when cortisol is high. Heterogeneity was low to moderate (total \( I^2 = 35.9\% \), between-study \( I^2 = 34.4\% \), within-study \( I^2 = 1.5\% \); Higgins et al., 2003).

Moderator analyses testing the effect of type of outcome measure, and analyses on the different outcome variables, yielded no significant effect (see Table 5 for statistics on these moderator analyses).

3.7. Robustness checks for low cortisol

3.7.1. Funnel plot asymmetry

Egger’s test (Egger et al., 1997) indicated no funnel plot asymmetry \( (z = .083, p = .934) \), but the trim and fill method (Duval and Tweedie, 2000) indicated that three studies were missing on the left side of the funnel plot (i.e. three studies not in line with the hypothesis were missing, see Fig. 5, left panel). When these hypothetical studies were added to the funnel plot (Fig. 5), the corrected effect size slightly decreased and was no longer significant: \( r = .044, p = .149 \), which might indicate publication bias.

3.7.2. Sensitivity analyses

Sensitivity analyses using the leave-one-out method were performed in order to investigate influences of single effect sizes. With a minimum of \( r = .056 (p = .059) \) and a maximum of \( r = .076 (p = .002) \), single effect sizes determine whether the overall effect is significant or not. The influence of single studies, potentially containing multiple effect sizes seemed similar, with a minimum of \( r = .053 (p = .072) \) and a maximum of \( r = .079 (p = .002) \). As can be seen from these analyses, the significance of the effect of testosterone on status-relevant behavior for low cortisol can change based on a single study or effect size, which warrants caution regarding the results.

As the meta-analysis contained one study with a particularly large sample size (Mazur and Booth, 2014), analyses for low cortisol were also performed without this study. The effect was similar: \( r = .061 (p = .047) \). Thus, omitting this study from the analysis does not change the results.

3.7.3. P-curve

To check the evidential value of the study and to check for flexibility in data-analysis, a p-curve was created (see Fig. 6; Simonsohn et al., 2015). Out of the 49 effect sizes, 13 were statistically significant. From these 13 significant effect sizes, 9 had p-values < .025. The p-curve analysis on the interaction effect indicated evidential value (full p-curve \( z = -2.78, p = .003 \); half p-curve \( z = -2.84, p = .002 \)), as the p-curve was right skewed (i.e. most significant p-values were low). However, note that in the case of an expected attenuated interaction effect, which is the case for the dual-hormone hypothesis, the p-curve of the interaction itself (see main analyses) should be used to assess evidential value (Simonsohn et al., 2015).

3.7.4. Test of excessive significance

The exploratory test for excess of significance (Ioannidis and Trikalinos, 2007) was used to check if the number of significant results differs from the expected number. Based on the overall effect and the sample size in each study, 5.2 significant effect sizes were expected, and 13 were found in the current meta-analysis. This difference is significant: \( \chi^2 (1) = 13.20, p < .001 \).

3.7.5. PET/PEESE analyses

The PEESE is reported as PET was significant (Stanley and Doucouliagos, 2014), the estimated PEESE effect size was \( r = .085, p < .001 \). This indicates that the true effect size of the effect for low cortisol might be slightly larger than the overall estimated effect for low cortisol \( r = .063 \).

3.7.6. Bayesian meta-analysis

The current meta-analysis was also conducted using Bayesian methods. A two-sided Bayesian meta-analysis using the default prior of the metaBMA R-package showed that there was more support for the null hypothesis (i.e. there is no effect of testosterone for low cortisol), \( BF_{10} = 3.38 \), than for the alternative hypothesis, \( BF_{10} = .30 \). Changing the prior (with a Cauchy prior width of \( r = .25 \) instead of the default of \( r = 1/\sqrt{2} \) leads to \( BF_{10} = 1.27, BF_{10} = .79 \). In both situations, the BF is thus in favor of the null hypothesis (i.e. the effect of testosterone on status-relevant behavior for low cortisol equals zero). According to the interpretation guidelines of Bayes factors (Jarosz and Wiley, 2014), the evidence for the null hypothesis is substantial (using the default prior) to anecdotal (using another prior).

3.7.7. Analyses without psychopathy-studies

As psychopathy might be more ambiguously related to status-
relevant behavior than other outcome variables, we checked its influence by performing the meta-analysis again without all studies using psychopathy as an outcome measure. The effect for low cortisol ($k = 42$) was similar: $r = .072$, $p = .043$ (see Table 2).

### 3.7.8. Analyses without controlling for dependency between participants

In the follow-up analysis for low cortisol, dependency between participants (i.e., one sample delivered more than one outcome) was controlled for by taking the average between these outcomes. However,
when all effect sizes were analyzed, without taking this dependency into account, the effect was highly similar: $k = 58$, $r = .068$, $p = .042$, 95% CI [.00, .13] (see Table 2).

3.7.9. Influence of reporting
In a categorical moderator analysis, we tested whether there was a difference between effect sizes reported in papers ($k = 10$) and those provided by authors ($k = 39$). This did not make a difference: $\Delta r = .012$, $p = .883$.

3.7.10. Influence of additional covariates
The difference between studies with ($k = 15$) and without ($k = 34$) additional covariates was not significant, $\Delta r = .055$, $p = .405$.

3.7.11. Influence of coefficient of variation
In two separate continuous moderator analyses, we tested whether the coefficient of variation of both testosterone and cortisol influenced the effect size for low cortisol. This was not the case (for testosterone, $k = 40$, $\Delta r = -.171$, $p = .318$; for cortisol, $k = 38$, $\Delta r = .010$, $p = .905$).

3.7.12. Summary of tests for robustness
Again, these robustness analyses indicate that caution is warranted in the interpretation, as (a) trim-fill analyses indicated missing studies and the corrected effect size no longer reached significance; (b) significance of the effect for low cortisol depended on single studies and effect sizes; (c) there are more significant findings than expected; (d) the data seem more in favor of the null hypothesis according to a Bayesian meta-analysis. No differences were found between (a) analyses with and without psychopathy as outcome measure; (b) effects that were reported in the original article and effects that were requested from the authors and (c) effects from models with vs. without additional covariates. Finally, the $p$-curve indicated evidential value.

3.8. Robustness checks for high cortisol

3.8.1. Funnel plot asymmetry
Egger’s test (Egger et al., 1997) indicated no funnel plot asymmetry ($z = -.020$, $p = .984$), and the trim and fill method (Duval and Tweedie, 2000) indicated that zero studies were missing on the funnel plot (see Fig. 5, right panel). These analyses do not indicate publication bias.

3.8.2. Sensitivity analyses
Sensitivity analyses using the leave-one-out method were performed in order to investigate influences of single effect sizes. With a minimum of $r = -.001$ ($p = .975$) and a maximum of $r = .012$ ($p = .542$), the influence of single effect sizes seemed small. The influence of single studies, potentially containing multiple effect sizes seemed small as well, with a minimum of $r = .001$ ($p = .998$) and a maximum of $r = .011$ ($p = .602$). Overall, the results for high cortisol seem quite robust.

As the meta-analysis contained one study with a particularly large sample size (Mazur and Booth, 2014), analyses for high cortisol were also performed without this study. The effect was similar: $r = -.001$ ($p = .955$). Thus, omitting this study from the analysis does not change the results.

3.8.3. $p$-curve
As only 2 of the 43 $p$-values were below .05, no $p$-curve has been created for the simple slopes analyses on high cortisol.

3.8.4. Test of excessive significance
The exploratory test for excess of significance (Ioannidis and Trikalinos, 2007) was used to check if the number of significant results differs from the expected number, given the overall effect. Based on the overall effect and the sample size in each study, 2.5 significant effect sizes were expected, and 4 were found in the current meta-analysis. This difference is not significant: $\chi^2 (1) = 1.00$, $p = .317$. This indicates no excess in significant findings for high cortisol.
3.8.5. PET/PEESE analyses

The PEESE is reported as PET was significant (Stanley and Doucouliagos, 2014), the estimated PEESE effect size was $r = .050, p = .004$. This indicates that the true effect size of the effect for high cortisol might be larger than the overall estimated effect.

3.8.6. Bayesian meta-analysis

The current meta-analysis was also conducted using Bayesian methods. A two-sided Bayesian meta-analysis using the default prior of the metaBMA R-package showed that there was more support for the null hypothesis (i.e. there is no effect of testosterone on status-relevant behavior for high cortisol), $BF_{01} = 42.18$, than for the alternative hypothesis, $BF_{10} = .02$. Changing the prior (with a Cauchy prior width of $r = .25$ instead of the default of $r = 1/\sqrt{2}$) leads to $BF_{01} = 15.00$, $BF_{10} = .07$. In both situations, the BF is thus in favor of the null hypothesis (i.e. the effect of testosterone on status-relevant behavior for high cortisol equals zero). According to the interpretation guidelines of Bayes factors (Jarosz and Wiley, 2014), the evidence for the null hypothesis is very strong (using the default prior) to strong (using another prior).

3.8.7. Analyses without psychopathy-studies

As psychopathy might be more ambiguously related to status-relevant behavior than other outcome variables, we checked its influence by performing the meta-analysis again without all studies using psychopathy as an outcome measure. The effect for high cortisol ($k = 42$) was similar: $r = .001, p = .949$.

3.8.8. Analyses without controlling for dependency between participants

In the follow-up analysis for high cortisol, dependency between participants (i.e. one sample delivered more than one outcome) was controlled for by taking the average between these outcomes. However, when all effect sizes were analyzed, without taking this dependency into account, the effect was highly similar: $k = 58, r = .005, p = .836, 95\% CI [-.04, .05]$.

3.8.9. Influence of reporting

In a categorical moderator analysis, we tested whether there was a difference between effect sizes reported in papers ($k = 10$) and those provided by authors ($k = 39$). This did not make a difference: $\Delta r = .036, p = .531$.

3.8.10. Influence of additional covariates

The difference between studies with ($k = 15$) and without ($k = 34$) additional covariates was not significant, $\Delta r = .035, p = .443$.

3.8.11. Influence of coefficient of variation

In two separate continuous moderator analyses, we tested whether the coefficient of variation of both testosterone and cortisol influenced the effect size for high cortisol. This was not the case (for testosterone, $k = 40, \Delta r = -.049, p = .740$; for cortisol, $k = 38, \Delta r = -.135, p = .057$).

3.8.12. Summary of tests for robustness

The effect of high cortisol seemed quite robust.

3.9. Moderator analyses on simple slopes

There were no moderating influences of sex, age and type of population for both low and high cortisol (see $\beta_j$ coefficients in Table 6). For type of measurement, effect sizes for low cortisol were significantly larger for studies that used other-reported questionnaires than for studies that used tasks (see Table 6). As many of these moderator analyses were based on a few studies per cell, these results are likely to be underpowered and should be interpreted with caution.

3.10. Main effects of testosterone

For a better interpretation of the interaction effect, main effects of testosterone were also analyzed. Note that this is not similar as a meta-analysis on the effects of testosterone (e.g. Archer et al., 2005), as we only included studies that also measured cortisol. We analyzed the effects of testosterone by extracting, from each study, the main effect of testosterone in a model that also included cortisol and the testosterone by cortisol interaction. The effect of testosterone on all outcome measures together ($k = 47$) was significant and in the expected direction: $r = .062, p = .001, 95\% CI [.02, .10]$. The effect size was small (Cohen, 1992).

Moderator analyses testing the effect of type of outcome measure yielded no significant results, i.e. the main effect of testosterone did not differ between outcome measures. Moderator analyses on the different outcome measures yielded a significant result for Status ($k = 5, r = .168, p = .047, 95\% CI [.00, .32]$) and non-significant results for Dominance ($k = 9, r = .028, p = .662, 95\% CI [-.10, .15])$, Risk Taking ($k = 13, r = .053, p = .299, 95\% CI [-.05, .15])$, Aggression ($k = 14, r = .063, p = .150, 95\% CI [-.02, .15])$, and Psychopathy ($k = 7, r = .005, p = .937, 95\% CI [-.12, .13]$).

3.11. Main effects of cortisol

Similarly, the main effect of cortisol on all outcome measures together ($k = 47$) was in the expected direction, significant but small: $r = .036, p = .001, 95\% CI [.06, .01]$. Moderator analyses testing the effect of type of outcome measure yielded no significant results, i.e. the main effect of cortisol did not differ between outcome measures. Moderator analyses on the different outcome measures yielded no significant results: Status ($k = 5, r = .047, p = .521, 95\% CI [-.19, .10])$, Dominance ($k = 9, r = .064, p = .227, 95\% CI [-.17, .04]$), Risk Taking ($k = 13, r = .005, p = .890, 95\% CI [-.08, .07])$, Aggression ($k = 14, r = -.043, p = .155, 95\% CI [-.10, .02])$ and Psychopathy ($k = 7, r = -.046, p = .306, 95\% CI [-.13, .04]$).

4. Discussion

This study aimed to test the dual-hormone hypothesis using a meta-analytic approach. The dual-hormone hypothesis posits that testosterone is particularly related to status-relevant behavior when cortisol levels are low, and not when cortisol levels are high (Mehta and Josephs, 2010). A meta-analysis on the interaction effect between testosterone and cortisol on status-relevant behavior was performed, as well as two follow-up meta-analyses on the relationship between testosterone and status-relevant behavior, one for low and one for high cortisol levels.

Only marginal support for the dual-hormone hypothesis was found: The overall effect size of the testosterone by cortisol interaction on status-relevant behavior was significant but small. All interaction effects that were included were derived from models that also contained main effects of testosterone and cortisol, suggesting that studying the interaction between these hormones is of added value as compared to only studying main effects. Moderator analyses indicated that the interaction effect was not influenced by sex, age and population type, although it should be noted that these null findings could originate in a lack of power. Follow-up meta-analyses on the simple slopes also marginally supported the dual-hormone hypothesis, indicating a significant but small effect size for low cortisol and an effect size close to zero for high cortisol.

---

6 For one study (Tackett et al., 2014), containing two effect sizes, main effects of testosterone and cortisol were not available, therefore $k = 47$ instead of $k = 49$. 

---

Although small, the magnitude of the effects in the current study ($r = .06$ for testosterone × cortisol, $r = .06$ for testosterone and $r = .04$ for cortisol) is comparable to other meta-analytic effects found in hormone-behavior research. That is, testosterone and aggression, $r = .08$ (Archer et al., 2005), testosterone and risk taking, $r = .12$ (Kurath and Mata, 2018), testosterone and leadership, $r = .007$ (Van der Meij et al., 2016), competition outcome and testosterone changes, $d = .20$ ($= r = .10$, Geniole et al., 2017), cortisol and externalizing behavior, $r = -.05$ (Alink et al., 2008).

Robustness analyses suggested that the observed interaction effect might be an overestimation of the true effect, and caution is warranted in the interpretation. That is, significance of the overall effect depended on single studies; the evidential value of the meta-analysis was weak according to the p-curve, because of a lack of power; there are more significant findings than expected; power of all but one of the individual studies is low ($\leq 20\%$); the true effect might be smaller than the observed effect according to PET/PEESE analyses; and the data seem more in favor of the null hypothesis according to a Bayesian meta-analysis.

The results of the current study must be interpreted in the light of at least three limitations. First, although all outcome measures might be related to status and all outcome measures might be related to testosterone, there might be heterogeneity in the magnitude of these relationships. That is, outcome measures on status itself are more direct indicators of status than, for example, measures of aggression or psychopathy. In addition, the link between risk taking and testosterone seems more pronounced than the link between psychopathy and testosterone. Therefore, effect sizes of different outcome measures were compared in a moderator analysis, which indicated that there were no differences between status, dominance, risk taking, aggression and psychopathy. Additionally, the effect size for none of the outcome measures was significant. Although not significant, the effect size for measures of status ($r = -.16$) appeared to be larger than for the other outcome measures, although of a similar magnitude as effect sizes usually reported in hormone behavior research. A similar and slightly stronger pattern of results was observed for the effect of testosterone for low cortisol: only the effect size of status was significant, but direct comparisons with other outcome measures yielded no significant differences.

Second, the absence of significant effects from the moderator analyses might be due to a lack of power to investigate these effects. In general, moderator analyses for meta-analyses with less than 100 studies are likely to be underpowered (Schmidt and Hunter, 2014).

Third, the current study did not investigate studies using testosterone-cortisol ratios, instead of interaction effects. The general outcome of studies investigating the testosterone-cortisol ratio in relation to status-relevant behavior roughly resembles the dual-hormone hypothesis: Higher testosterone-cortisol ratios were associated with violence (Romero-Martínez et al., 2013), psychopathy (Glen et al., 2011), and neural responses to threat and anger (Hermans et al., 2008; Denson et al., 2013a, 2013b; Sollberger and Ehlers, 2016).

Apart from the findings on overall effect sizes, heterogeneity in these effect sizes also warrants discussion, as it was large. Therefore it is important to consider several sources of heterogeneity, and to identify future directions of research into the dual-hormone hypothesis accordingly.

First, heterogeneity in results might be explained by measurement differences. In our analyses, mixed samples were split: Men and women were analyzed separately. With regard to the p-curve, this approach could have led to a loss of power. For two studies (Barel et al., 2017; Mehta and Josephs, 2010), a significant interaction effect for the mixed sample was split into two non-significant effects when men and women were analyzed separately. For two other studies (Mehta et al., 2015a, 2015b, study 1; Grotzinger et al., 2018), a significant interaction effect for the mixed sample was split into a significant effect for women and a non-significant effect for men.

Table 6 Test statistics of moderator analyses on simple slopes. $\hat{\beta}$ denotes the difference in $r$ (i.e. $\Delta r$) between the groups mentioned in the Effects column, $\hat{\beta}$ denotes the estimate of the effect size for the particular outcome measure (see Methods section for further details), $\hat{\beta}^2_{estimates}$ is the variance between effect sizes, $\hat{\beta}^2_{study}$ is the variance between studies, QE is a test for residual heterogeneity, $^*$ $p < .10$, $^{**} p < .05$, $^{***} p < .01$, $^{****} p < .001$. Abbreviations: $k = number of effect sizes, df = degrees of freedom, CI = Confidence Interval, C = Cortisol, DV = Dependent Variable.

<table>
<thead>
<tr>
<th>Moderators (k)</th>
<th>Effects</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_0$</th>
<th>$95%$ CI</th>
<th>$\hat{\beta}^2_{estimates}$</th>
<th>$\hat{\beta}^2_{study}$</th>
<th>QE (df)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Men vs women</td>
<td>.084</td>
<td>.092*</td>
<td>.025</td>
<td>.000</td>
<td>125.3 (47)***</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Women (17 effect sizes)</td>
<td>.008</td>
<td>.000</td>
<td>.005</td>
<td>80.4 (47)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Low C (43)</td>
<td>Age in years</td>
<td>.004</td>
<td>.000</td>
<td>.021</td>
<td>.006</td>
<td>116.4 (41)***</td>
<td></td>
</tr>
<tr>
<td>Age High C (43)</td>
<td>Age in years</td>
<td>.000</td>
<td>.000</td>
<td>.005</td>
<td>65.3 (41)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population type (49)</td>
<td>Typical vs Delinquents</td>
<td>-.168</td>
<td>-.11</td>
<td>.023</td>
<td>.002</td>
<td>125.9 (47)***</td>
<td></td>
</tr>
<tr>
<td>Low C</td>
<td>Delinquents (2 effect sizes)</td>
<td>-.221</td>
<td>-.45</td>
<td>.004</td>
<td>77.9 (47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low C</td>
<td>Typical vs Delinquents</td>
<td>.120</td>
<td>.30</td>
<td>.000</td>
<td>79.9 (47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low C</td>
<td>Delinquents (2 effect sizes)</td>
<td>-.108</td>
<td>-.29</td>
<td>.001</td>
<td>.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement type (49)</td>
<td>Task vs Self-report</td>
<td>.087</td>
<td>.29</td>
<td>.002</td>
<td>.000</td>
<td>116.4 (46)***</td>
<td></td>
</tr>
<tr>
<td>Low C</td>
<td>Task vs Other-report</td>
<td>.177*</td>
<td>.34</td>
<td>.011</td>
<td>.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low C</td>
<td>Self- vs Other-report</td>
<td>.091</td>
<td>.26</td>
<td>-.002</td>
<td>.090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Tasks (21 effect sizes)</td>
<td>-.085*</td>
<td>.16</td>
<td>.006</td>
<td>.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Other-reg. quest. (8 effect sizes)</td>
<td>-.175*</td>
<td>.32</td>
<td>.023</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Measurement type (49)</td>
<td>.002</td>
<td>.09</td>
<td>.004</td>
<td>.004</td>
<td>73.3 (46)***</td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Task vs Self-report</td>
<td>-.087</td>
<td>.22</td>
<td>.002</td>
<td>.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Self- vs Other-report</td>
<td>-.089</td>
<td>.22</td>
<td>.005</td>
<td>.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Tasks (21 effect sizes)</td>
<td>-.014</td>
<td>.05</td>
<td>.016</td>
<td>.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Self-report quest. (20 effect sizes)</td>
<td>-.04</td>
<td>.07</td>
<td>.004</td>
<td>.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High C</td>
<td>Other-reg. quest. (8 effect sizes)</td>
<td>-.073</td>
<td>.19</td>
<td>.004</td>
<td>.004</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In our analyses, mixed samples were split: Men and women were analyzed separately. With regard to the p-curve, this approach could have led to a loss of power. For two studies (Barel et al., 2017; Mehta and Josephs, 2010), a significant interaction effect for the mixed sample was split into two non-significant effects when men and women were analyzed separately. For two other studies (Mehta et al., 2015a, 2015b, study 1; Grotzinger et al., 2018), a significant interaction effect for the mixed sample was split into a significant effect for women and a non-significant effect for men.
immunoassays, especially in women (Granger et al., 2004), and low levels of testosterone in women are inflated by these measures (Welker et al., 2016); (4) differences in testosterone as well as cortisol estimates between saliva kits from different companies have been observed (Welker et al., 2016; Baecher et al., 2013) and (5) single saliva samples are sensitive to fluctuations (Harden et al., 2016).

A promising alternative for immunoassay measurements is the use of mass-spectrometry methods (Taylor et al., 2015). These methods have a higher specificity and a wider range of analysis than immunoassays (Field, 2013). Immunoassays correlate only moderately with mass-spectrometry methods with hair samples only found a testosterone × cortisol interaction using the mass-spectrometry method (Grotzinger et al., 2018), and another study using mass-spectrometry confirmed the dual-hormone hypothesis as well (although only in men, Ronay et al., 2018). Therefore, future studies on the dual-hormone hypothesis are recommended to use mass-spectrometry methods, or at least to acknowledge and, if possible, control for the disadvantages of immunoassays, e.g. by averaging across multiple saliva samples, and by storage at -80 ° Celsius only.

Second, it might be important to incorporate individual differences (Carré and Archer, 2018). Accordingly, we compared effects of men and women, and this moderator analysis yielded no significant differences. However, given the number of effect sizes, this analysis was likely to be underpowered (Schmidt and Hunter, 2014). Of note, in the moderator analysis, the interaction effect and the effect of testosterone for low cortisol was significant for men and almost zero for women. These lower effects in women might originate in measurement errors as mentioned above, but could also indicate a true absence of the dual hormone effect in women. Large studies including both men and women and using precise measurements are needed to detect whether there are differences between men and women regarding the testosterone by cortisol interaction on status-related behavior (Mehta and Prasad, 2015).

Other individual differences might relate to personality characteristics. For example, the dual-hormone interaction effect on externalizing problems in youth was only observed for those high on the personality trait emotional instability and disagreeableness (Tackett et al., 2014). Similarly, testosterone was particularly related to aggression and dominant behavior in participants high on trait dominance (Carré et al., 2017; Slater et al., 2011; Pfaatheicher, 2016b). Finally, it should be noted that the majority of the studies was conducted in student samples, which may bias results.

Third, context might also influence the effects, thereby creating heterogeneity. For example, dual-hormone effects on dominance were particularly observed after social defeat (Mehta and Josephs, 2010; see also Henry et al., 2017). Moreover, another study found a testosterone by cortisol interaction on aggression in a condition of social inclusion, but not in a condition of social exclusion (Geniole et al., 2011). Relatedly, a recent meta-analysis on competition related changes in testosterone showed stronger testosterone reactivity in lab studies than in field studies (Geniole et al., 2017). These studies suggest that environment matters. Another issue that may explain inconsistent findings is that it is speculated that the interaction between cortisol and testosterone in relation to status-relevant behaviors differs between situations of perceived threat versus reward (see Mehta et al., 2017 for discussion on threat pathways and the dual-hormone interaction and Welker et al., 2015 for discussion on reward pathways; see Chester, 2017 for an example that aggression might be related to both pathways). Future studies should further determine environmental conditions under which the dual-hormone hypothesis holds.

Fourth, heterogeneity in results might be caused by heterogeneity in instruments that were used. Even within categories of outcome measures, replications with identical instruments are scarce. In this meta-analysis only a few instruments were used more than once: The BART (Mehta et al., 2015a; Ronay et al., 2018), the YSR (Platje et al., 2015 and Tackett et al., 2014) and the PSAP (Geniole et al., 2011, 2013) were used twice.

The moderator analysis comparing types of measurement was not significant (i.e. effects did not differ between tasks, self-report questionnaires and other-report questionnaires). The correlation between these different measurement types is often unknown. Even when measuring the same construct, correlations between tasks and questionnaires are often low (e.g. Toplak et al., 2013). More direct replications, using similar instruments as used in earlier studies, are needed to provide an estimate of the dual-hormone interaction, not affected by variation in instruments.

These potential origins of heterogeneity thus foster various ideas for future research. In addition, there are other interesting venues for future research. First, although the dual-hormone hypothesis yields a small effect size for basal hormone levels, it may show a larger effect size for state-dependent hormonal changes (i.e. ‘state vs trait’ hormones; Mehta and Prasad, 2015). For example, increased testosterone combined with decreased cortisol during bargaining was related to beneficial outcomes (Mehta et al., 2015b). In another study, induction of acute stress, with associated rises in cortisol, suppressed the effect of basal testosterone on retaliation (Prasad et al., 2017). Furthermore, negative associations between basal cortisol and testosterone reactivity after competition or social evaluation were reported (e.g. Bedgood et al., 2014), although this might also be the result of a statistical confound (Welker et al., 2017).

Second, an important consideration for future research is whether behavior that is associated with the high testosterone low cortisol profile is pro- or anti-social, and whether this leads to the actual attainment of social status (Hamilton et al., 2015). An interesting hypothesis is that the route to high status might differ between contexts. For example, the high testosterone low cortisol profile is associated to pro-social constructs like leadership (Mehta and Josephs, 2010; Sherman et al., 2016) and popularity (e.g. being able to inspire others, Edwards and Cato, 2013; likability among teammates, Ponzi et al., 2016) in what were assumed to be prestige-based status hierarchies. Although speculatively, in certain contexts the high testosterone low cortisol profile might even be associated with suppressed anti-social behavior, as this might not be the route towards social status (relatedly, in student populations, anti-social behavior like cheating and psychopathic traits was associated with a high testosterone high cortisol profile, Lee et al., 2015; Welker et al., 2014). However, in other populations anti-social behavior might be helpful to attain status, and therefore this anti-social behavior might be related to the high testosterone low cortisol profile in these populations (e.g. aggression and violent crimes were associated with the high testosterone low cortisol profile in delinquents, Dabbs et al., 1991; Popma et al., 2007). A related question is whether the high testosterone low cortisol profile is beneficial or not. It is suggested that in stable hierarchies, high status provides a buffer towards cortisol and testosterone responses to a stressor, whereas in unstable hierarchies, high status increases testosterone and cortisol responses to stress (Knight and Mehta, 2017). Another study showed that in a leadership position, individuals with a high testosterone low cortisol profile showed more confidence and less stress (Mehta and Josephs, 2010). Furthermore, a recent study suggested that groups with high testosterone and low cortisol performed best (Akinola et al., 2016).

However, the most important recommendation for future research stems from the lack of robustness of the results and the low power of the individual studies in the current meta-analysis. As this might indicate publication bias and/or flexibility in data-analysis, a well-powered pre-registered study is needed to examine the dual-hormone hypothesis. Large-scale pre-registered replications are increasingly often used to test the presence or absence of effects found in meta-analyses (e.g. Carter et al., 2015). Pre-registration is a solution to remove publication bias and flexibility in data-analysis (Simmons et al., 2011). Given the fact that effect sizes were largest for measures of status itself and for males, this replication study should first focus on status as an outcome measure.

To conclude, the dual-hormone hypothesis, which implies that testosterone is particularly related to a broad range of status-relevant...
behavior when cortisol is low, is marginally supported by the current meta-analysis. Given the signs of publication bias and/or flexibility in data-analysis, a future large scale pre-registered study is required, preferably in men using direct measures of status.

Role of funding sources
This research is supported by a VICI grant 453-12-015 (HMH, TJD) and by a MaGW grant 480-12-015 (JAvR) from the Netherlands Organization for Scientific Research (NWO). The funding source had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, nor the decision to submit the paper for publication.

Conflicts of interest
The authors report no potential conflicts of interest.

Acknowledgements
The authors would like to thank Efrat Barel, Macià Buades-Rotger, Kathleen Casto, Carlos Cueva, Thomas Denson, Shawn Geniole, Andrea Glenn, Andrew Grotzinger, Allan Mazur, Pranjal Mehta, Stefan Pfattheicher, Evelien Platje, Richard Ronay, Kristopher Smith, Jennifer Tackett, Keith Welker, Donne van der Westhuizen, Samuele Zilioli for kindly providing the requested data, and Janneke Staaks for help with the literature search. Furthermore, the authors would like to thank Quentin Gronau for assistance with the Bayesian analyses. Finally, the authors would like to thank 4 anonymous reviewers for the useful and constructive comments, and Janneke Staaks for help with the literature search. Furthermore, the authors performed a split on cortisol (selecting the participants with 1/3 lowest and 1/3 highest cortisol sample), instead of a simple median split. Furthermore, the authors used the moderating role of basal cortisol. Psychoneuroendocrinology 47, 107–115.

Appendix A. Supplementary materials
Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.neubiorev.2018.12.004.

References

Ahn, H., Rodkin, P.C., 2014. Classroom-level predictors of the social status of aggression:

Studies marked with an asterisk were included in the meta-analysis.

Non-log converted values for cortisol and testosterone were used.

Note that sample sizes are slightly lower than reported in the original paper, because participants with missing salivary data were omitted. Furthermore, the authors performed a split on cortisol (selecting the participants with 1/3 lowest and 1/3 highest cortisol sample), instead of a simple slope analysis.

Age information was reported for the entire sample.

Note that df for Low and High C are lower than expected based on N, because the reported analyses were based only on one third of the sample. Therefore, df were 19, 21, 24, 21 for men and women and for Low and High C respectively. These data were provided by the authors.

Only r and df were reported, t-values were calculated using $r = \sqrt{(t^2 / (t^2 + df))}$ (Zakzanis, 2001). This study did not use simple slopes, but conducted a median split. A df for the median-split follow-up analyses was deduced from the df. Note that the t-value of the interaction was rescored from positive to negative, as the result was in line with the dual-hormone hypothesis, hence the interaction should be negative.

Submissiveness scores were recoded as this is roughly reversed to the personality trait dominance.

t-values for the interaction and for High C were provided by the authors.

Age information was reported for the entire sample.

Age information was reported for the entire sample.

The raw dataset was provided by the authors. All effects were calculated from there. Note that sample sizes differ slightly from the original paper, because of some missing values.

Age information was reported for the entire sample.

Z-values were interpreted as t-values.


Barel, E., Shahrbani, S., Tzischinsky, O., 2017. Sex hormone/cortisol ratios differen-


11 Age information was reported for the entire sample.
12 Note that df for Low and High C are lower than expected based on N, because the reported analyses were based only on one third of the sample. Therefore, df were 19, 21, 24, 21 for men and women and for Low and High C respectively. These data were provided by the authors.
13 Only r and df were reported, t-values were calculated using $r = \sqrt{(t^2 / (t^2 + df))}$ (Zakzanis, 2001). This study did not use simple slopes, but conducted a median split. A df for the median-split follow-up analyses was deduced from the df. Note that the t-value of the interaction was rescored from positive to negative, as the result was in line with the dual-hormone hypothesis, hence the interaction should be negative.
14 Submissiveness scores were recoded as this is roughly reversed to the personality trait dominance.
15 t-values for the interaction and for High C were provided by the authors.
16 Age information was reported for the entire sample.
17 Age information was reported for the entire sample.
18 The raw dataset was provided by the authors. All effects were calculated from there. Note that sample sizes differ slightly from the original paper, because of some missing values.
19 Age information was reported for the entire sample.
20 Z-values were interpreted as t-values.
21 The raw dataset was provided by the authors. All effects were calculated from there.
22 In the original paper, t = .8 was reported. Because of the direction (i.e. B = -.49), this was coded as t = -.8. This was confirmed by the authors.
23 As the dependent variable was dichotomous, simple slopes analyses used Z-values which were provided by the authors. These were interpreted as t-values. Furthermore, sample sizes were derived from Mehta and Jones (2006).
24 Note that the df were calculated using the APIIM method, which accounts for the non-independence of observations (Mehta et al., 2017).
25 Age was provided by the authors. This information is based on 94 participants, males and females together.
26 All t-values were provided by the authors. The t-values belonging to the OLS regression model for antisocial punishment were chosen.
27 All t-values were calculated based on p-values and N as reported in Table II, model 2, of the original study.
28 All values were derived from the dataset that was provided by the authors. Note that the sample size in boys slightly differs from the original paper, as in the paper four participants were excluded for reasons irrelevant to the current study.
29 Age information was reported for the entire sample.
30 All t-values were either provided by the authors or calculated based on p-values and df as reported in the original paper.
31 N and all t-values on covert aggression were based on the raw dataset that was provided by the authors. Values for overt aggression were derived from the original paper.
32 In the original paper, this was documented as $t = .31$.
33 All values were derived from the dataset that was provided by the authors.
34 All t-values were calculated based on p-values and df as reported in the Results section of the original paper, the direction of the t-value is based on the direction of the beta value.
35 Instead of simple slopes analyses, the authors performed a median split. Furthermore, the model contained one additional parameter (i.e. risk attitude).
36 All values were derived from the dataset that was provided by the authors.
37 Empathy scores were recoded and interpreted as psychopathy.


