Psychoneuroendocrinology 110 (2019) 104435

Contents lists available at ScienceDirect

Psychoneuroendocrinology

journal homepage: www.elsevier.com/locate/psyneuen

Challenging the challenge hypothesis on testosterone in fathers: Limited meta-analytic support

Willemijn M Meijer[a], Marinus H van IJzendoorn[b,c], Marian J Bakermans - Kranenburg[a,d,*]

[a] Clinical Child & Family Studies, Faculty of Behavioral and Movement Sciences, Vrije Universiteit, Amsterdam, the Netherlands
[b] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands
[c] School of Clinical Medicine, University of Cambridge, UK
[d] Leiden Institute for Brain and Cognition, Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Testosterone
Parenting
Fathers
Hormones
Parental status
Challenge hypothesis

ABSTRACT

In fathers testosterone levels are suggested to decrease in the context of caregiving, but results seem inconsistent. In a meta-analysis including 50 study outcomes with N = 7,080 male participants we distinguished three domains of research, relating testosterone levels to parental status (Hedges' g = 0.22, 95% CI: 0.09 to 0.35; N = 4,150), parenting quality (Hedges' g = 0.14, 95% CI: 0.03 to 0.24; N = 2,164), and reactivity after exposure to child stimuli (Hedges' g = 0.19, 95% CI: -0.03 to 0.42; N = 766). The sets of study outcomes on reactivity and on parenting quality were both homogeneous. Parental status and (higher) parenting quality were related to lower levels of testosterone, but according to conventional criteria combined effect sizes were small. Moderators did not significantly modify combined effect sizes. Results suggest that publication bias might have inflated the meta-analytic results, and the large effects of pioneering but small and underpowered studies in the domains of males' parental status and parenting quality have not been consistently replicated. Large studies with sufficient statistical power to detect small testosterone effects and, in particular, the moderating effects of the interplay with other endocrine systems and with contextual determinants are required.

1. Introduction

In many species, lower testosterone levels are linked to increased parenting efforts. In rodents paternal testosterone decreases after the birth of pups (e.g., Brown et al., 1995; Trainor et al., 2003) and higher testosterone levels in fathers are associated with less nurturing behavior (Clark and Galef, 1999). In mice, testosterone treatment has a negative impact on paternal behavior (Okabe et al., 2013). Similar results are found in biparental primates such as marmosets and tamarins. Higher testosterone levels predict less nurturing behavior in marmoset fathers (Nunes et al., 2001), and paternal testosterone drops following the birth of pups in marmosets and cotton–top tamarins (Ziegler et al., 2009, 2004). Such findings are in line with the "challenge hypothesis", originally based on avian research (Wingfield et al., 1990), suggesting that testosterone levels are higher in the context of competition and lower in the context of monogamous relationships (Rosenbaum et al., 2018) and in the context of caregiving (Archer, 2006). In humans, a number of studies support an association between steroid hormones and paternal behavior, but results seem inconsistent. To contribute to the test of the challenge hypothesis, we performed a systematic literature search and meta-analysis on the relation between testosterone and parenting in human fathers.

Several studies in humans show that fathers have lower testosterone levels than non-parents (Barrett et al., 2013; Gettler et al., 2011b; Gray et al., 2006; Kuzawa et al., 2010, 2009; Muller et al., 2009). Moreover, there is limited evidence suggesting that paternal testosterone levels decline during the prenatal period (Edelstein et al., 2015) and during the transition into parenthood (Berg and Wynne-Edwards, 2001; Gettler et al., 2011b; Storey et al., 2000). According to a study among two groups of Tanzanian men this decline in testosterone may not be related to fatherhood itself but to being in the proximity of one's child. In Hadza men (foragers), less paternal care was related to higher testosterone levels, while in Dagota men (pastoralists who spend much of their time away from their families), this relation was absent and testosterone levels of fathers were similar to those of non-fathers (Muller et al., 2009). As proximity to the child is necessary to show parental behavior, one would expect not (only) being or becoming a parent to be related with testosterone levels but parental behavior itself may be an important correlate of testosterone. Indeed, several studies show that fathers' testosterone is lower when they are more involved in parental

* Corresponding author.
E-mail address: m.j.bakermans@vu.nl (M.J. Bakermans - Kranenburg).

https://doi.org/10.1016/j.psyneuen.2019.104435
Received 7 February 2019; Received in revised form 4 September 2019; Accepted 4 September 2019

care (Alvergne et al., 2009; Edelstein et al., 2017; Mascaro et al., 2013; Weisman et al., 2014). Besides this association with parental involvement, paternal testosterone levels were also linked to quality of parenting, in particular sensitive and nurturing behavior (Fleming et al., 2002; Storey et al., 2011).

It has been suggested that at least two other factors besides proximity to offspring are critical for testosterone levels in males. A first crucial factor seems to be a monogamous relationship with a partner, which makes the investment in offspring part of inclusive fitness. Partnering might be a necessary condition for a committed role in child-rearing and the downregulation of testosterone levels (Gettler, 2016; Grebe et al., 2019). Second, two dimensions of paternal behavior might be distinguished, that is, warm, sensitive and playful interactions on the one hand, and paternal behavior to defend and protect offspring against intruders and other dangers on the other hand (Van Anders et al., 2012). The Offspring Defense Paradox suggests that sensitive caregiving might lower testosterone levels in fathers, whereas testosterone levels might be elevated when fathers are provoked to protect their infant (Van Anders et al., 2012). For obvious ethical reasons only few studies have been conducted on testosterone levels in the context of real dangers threatening the infant's safety (but see some studies on paternal responses to infant crying, Van Anders et al., 2012, 2014; Fleming et al., 2002). Thus, environmental characteristics (cues that the context is safe or potentially unsafe, as indicated by infant crying sounds) may affect testosterone levels in males.

Furthermore, age of the child might moderate the association between testosterone and parenting. A study on fathers of preschoolers did not find a relation between testosterone levels and observed parenting (Endendijk et al., 2016). Any decline of testosterone levels during the perinatal period and after birth of a child may depend on parental investment and may rebound with time. Some studies on testosterone reactivity to parent-child interaction support this hypothesis showing a decline in testosterone levels of fathers after interacting with their own infant (Bos et al., 2018; Kuo et al., 2016; Storey et al., 2011) but other studies did not find change over time in fathers (Delahunty, 2003; Kuo et al., 2018). One study showed a decrease in men's testosterone after nurturing care of a crying infant simulator (Van Anders et al., 2012) but these findings were not replicated in a larger trial (Van Anders et al., 2014).

The Steroid/Peptide theory of social bonds presents a model in which testosterone levels are embedded in an interactive endocrine system with other peptides such as oxytocin, vasopressin, or cortisol (Van Anders et al., 2012; Bos, 2017; see also Abraham and Feldman, 2018; Bos et al., 2018; Voorthuis et al., 2017). However, multi-peptide studies on fathers are still rare, which precludes a meta-analytic approach. Therefore, we focus on studies of testosterone in fathers as an essential component of any theory involving the hormonal basis of parenting, and more specifically paternal caregiving (Feldman and Bakermans-Kranenburg, 2017). In a series of meta-analyses we estimated the combined effect sizes in three categories of testosterone studies in males: experimental studies on reactivity to infant signals or to interaction with the child ('Reactivity'), studies on parenting quality or involvement ('Parenting quality'), and studies comparing fathers with males who were not parents ('Parental status'), taking into account similarity in partner status (i.e., whenever possible, either both groups of males were partnered; or, when fathers were not partnered, a non-partnered comparison group was used). Based on the 'challenge hypothesis' we expected that (1) reactivity tests eliciting (more) caregiving behaviors would lead to lower testosterone levels –except when parental protection is elicited, (2) lower testosterone levels would be associated with more parental involvement and higher parenting quality, and (3) fathers would have lower testosterone levels compared to non-fathers.

Furthermore, we investigated the effects of various characteristics that could moderate the meta-analytic effect sizes. The moderating role of publication year and sample size were investigated to detect possible publication bias. Study design and type of testosterone sampling were also considered as moderators. For study design, within-subject analyses were distinguished from between-subject analyses, and for reactivity studies, RCTs were distinguished from other studies. Type of testosterone sampling might affect the results as salivary testosterone contains more free testosterone compared to plasma testosterone. Whether male participants were partnered or not was used as a moderator in all three domains of inquiry. We expect that the difference between fathers and non-fathers will be most pronounced when both groups are in a (stable) partner relationship, enhancing the probability of fathers' biological relatedness to the child, which – according to the Challenge Hypothesis - predicts heightened involvement in child care. Parental status was included as moderator in the meta-analyses of studies on reactivity and parenting quality.

2. Material and methods

2.1. Literature selection

We performed a systematic literature search in four databases (PubMed, Web of Science, PsychInfo and ProQuest). Search terms were ("testosterone" OR "dehydroepiandroterone" OR "dhea") AND ("parent*" OR "father*" OR "mother*" OR "patern*" OR "matern*"). Other terms ("sex development" OR "sexual development" OR "sexual behavior" OR "polycystic ovary syndrome" OR "pcos" OR "klinefelter" OR "autis*") were used to exclude clinical papers (an asterix means the search included but was not limited to that exact word or fragment). The search was limited to humans and papers in the English language. Publications up to October 2018 were included. A PRISMA flow chart (Liberati et al., 2009) of the literature search and exclusion during selection and coding is provided in Fig. 1.

The literature search resulted in 1,435 hits. The first screening was based on title and abstract and performed by four coders. In a set of 194 randomly selected results (14%), three coders were reliable in inclusion versus exclusion with the expert coder (WMM). Due to the high exclusion rate and thus a skewed distribution of excluded and included studies, PABAKs were calculated (Prevalence Adjusted and Bias Adjusted Kappa, see Byrt et al., 1993). PABAK of the three raters compared to the expert coder were .93, .98 and .98.

The remaining 1,241 papers were subsequently coded by one of the coders. To include as many studies as possible, in this step the selection was not restricted to specific study designs or quality of the study. All dimensions of parenting were included: being a parent (comparing parents with non-parents), parenting behavior, as well as parental perceptions, cognitions, and emotions. Any study with information on parental testosterone was considered. This allowed us to include studies with testosterone administration, studies with baseline testosterone values measured in blood or saliva, or testosterone reactivity. We considered studies with testosterone as dependent or independent variable. Furthermore, we included studies on paternal testosterone as well as maternal testosterone. Studies were not restricted to specific populations (geographically, number and age of children), although we did exclude clinical samples as parenting might be influenced by the clinical condition or setting. In case of uncertainty based on title and abstract, the paper was included in this phase. Studies were excluded in this first selection step if it was not an original study (e.g. review article), if the paper was not about parenting, not about parental testosterone, if it included a clinical sample, or if the subjects in the paper were not human.

Google Scholar was searched for additional results using the following search terms: (parenting testosterone human –reproduction –animal* -rodent* -rat* -pig* -ape –rabbit* -gerbil* -ostrich –mice –mouse –bird* -vole* -marmoset* -romantic –testis –testes –testicular –"sexual behavior" –"polycystic ovary syndrome") and excluding patents and citations. The results were sorted on relevance and screened. Of the 32 relevant results found, three were new studies. Finally,

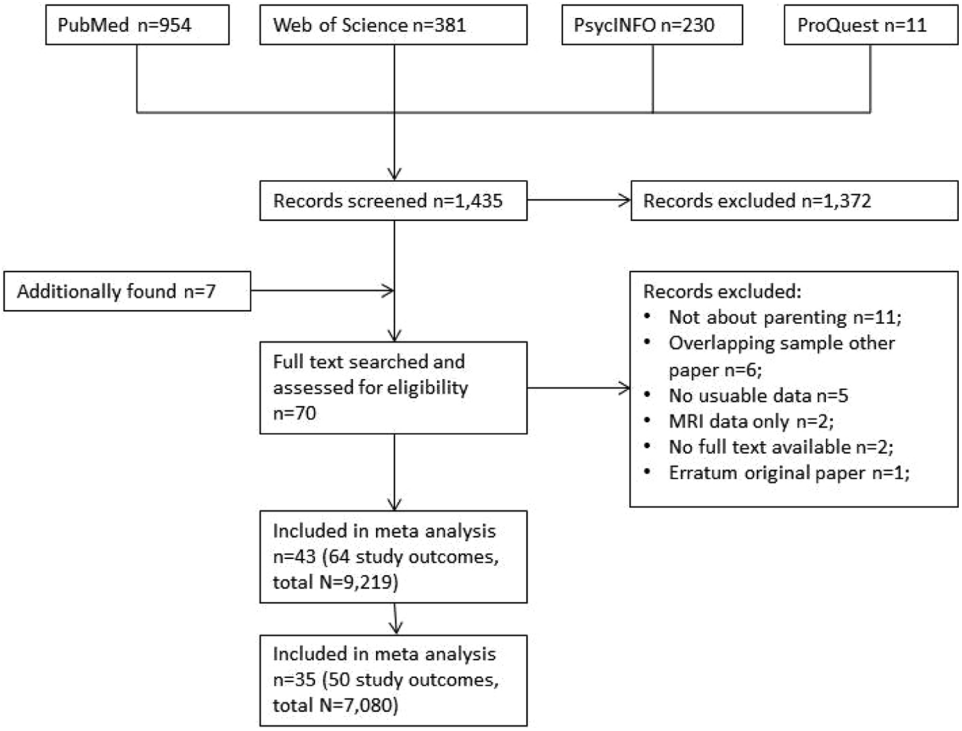


Fig. 1. Flow chart of literature search and selection.

through reference lists and personal communications another four papers were included in the eligibility assessment.

2.2. Eligibility and coding

Two coders (WMM and MJBK) independently scored the 70 included papers to assess final eligibility of each study and agreed (100%) upon including 40 studies with 58 study outcomes in the meta-analysis. One of the reviewers suggested including three additional papers and kindly provided the necessary statistics. Reasons for exclusion were, among others, overlapping samples (n = 6) and no useable data (n = 5), see Fig. 1. The set of 70 studies included only two fMRI studies (Bos et al., 2010; Kuo et al., 2012). Based on the different methodology and measures of such studies, it was decided to exclude these studies. Furthermore, one study (Weisman et al., 2014) presented results based on baseline testosterone as well as testosterone levels after an oxytocin administration. As this was the only study outcome related to hormone or neuropeptide administration, we decided to only include the baseline results in the analyses. Because the challenge hypothesis focuses on males and because the number of studies on females was low, we decided to focus the analyses on males. However, we briefly report combined effect sizes for parental status and parental quality in females without conducting moderator or related analyses. Sample sizes and effect sizes of the 35 papers including 50 study outcomes on N = 7,080 participants are listed in Table 1. Based on the 'challenge hypothesis', we expected a negative association between testosterone levels and parenting quality or involvement and coded the included study outcomes accordingly, thus associations in line with the hypothesis were assigned positive effect sizes.

The coding system is shown in Supplementary Table 1. Background information included type of publication and year of publication. Sample characteristics were mean age, ethnicity, whether or not the participants were partnered, and parental status of the participants. If a study included both male and female participants, only the study outcomes for males were included. Study characteristics included type of testosterone sampling (saliva or blood), design characteristics (RCT or non-RCT, within-subjects or between-subjects), sample size, response rate, intention to treat, blinding of subjects, blinding of researchers, and reporter of outcome measure. Some characteristics were only of interest for RCTs. Two coders (WMM and MBK) independently coded all studies. Average intercoder agreement across moderators, based on PABAKs and correlations, was .93. Discrepancies were resolved by discussion.

2.3. Data analysis

For all outcomes, the effect sizes were computed as Hedges' g. Hedges' *g* is commonly used in meta-analysis as it corrects for the bias that could occur in small samples (Borenstein et al., 2009). Using the program Comprehensive Meta-Analysis (CMA) (Borenstein et al., 2014), results of individual studies were transformed into common metrics and overall effect sizes per outcome category were computed. For each study outcome, effect sizes (*g*) are shown in Table 1. The outcomes were categorized into reactivity studies (change in T level from before to after experimental paradigm or parent-child interaction, $k = 12$), studies on parenting quality or involvement ($k = 18$), and studies that compared parents with non-parents (parental status, $k = 20$). For studies that reported more than one outcome within a specific outcome domain (e.g., various indicators of parenting quality, or morning and evening testosterone levels related to parental status), we computed a combined effect size for the study within the outcome domain using CMA. Furthermore, five studies (Alvergne et al., 2009; Bos et al., 2018; Gray et al., 2002; Kuo et al., 2018, 2016) reported on outcomes across different domains. For example, baseline T levels were associated with parenting quality, and in addition results of a reactivity test were reported. In those cases the different outcomes were included in the respective domains, for which separate meta-analyses were conducted. Doing so we ensured that within the three study domains

Table 1
Characteristics of included studies, per study outcome.

Author/year of publication/outcome[a]	Outcome category	Sample size (N)	Effect size (Hedges' *g*)
Alvergne et al., 2009 P	Parental Status	81	0,492
Alvergne et al., 2009 PQ	Parenting Quality	49	0,364
Berg and Wynne-Edwards, 2001 P	Parental Status	37	0,991
Bos et al., 2018 males R	Reactivity	49	0,899
Bos et al., 2018 males PQ	Parenting Quality	49	0,000
Burke and Bribiescas, 2018 P	Parental Status	48	0,212
Delahunty, 2003 R	Reactivity	22	0,000
Dorius et al., 2011 males PQ	Parenting Quality	352	−0,020
Edelstein et al., 2017 males PQ	Parenting Quality	27	0,144
Endendijk et al., 2016 males PQ	Parenting Quality	217	0,032
Fleming et al., 2002 PQ	Parenting Quality	13	0,225
Gettler et al., 2011a play intervention R	Reactivity	42	0,000
Gettler et al., 2011b developmental P W	Parental Status	269	0,110
Gettler et al., 2011b neither vs father only P	Parental Status	232	−0,063
Gettler et al., 2011b partner vs partner + father P	Parental Status	123	0,134
Gettler et al., 2015 PQ	Parental Quality	104	0.378
Gettler and Oka, 2016 partnered P	Parental Status	877	0.117
Gettler and Oka, 2016 never married P	Parental Status	463	0.187
Gettler and Oka, 2016 divorced P	Parental Status	165	0.298
Gordon et al., 2017 PQ	Parenting Quality	49	0,292
Gray et al., 2002 between P	Parental Status	29	0,324
Gray et al., 2002 within PQ	Parenting Quality	15	−0,119
Gray et al., 2004 P	Parental Status	43	−0,213
Gray et al., 2006 P	Parental Status	60	0,467
Gray et al., 2017 PQ	Parenting Quality	338	0,000
Julian and McKenry, 1989 PQ	Parenting Quality	37	0,936
Kuo et al., 2016 baseline PQ	Parenting Quality	142	−0,040
Kuo et al., 2016 reactivity R	Reactivity	142	0,080
Kuo et al., 2018 R	Reactivity	289	0,000
Kuo et al., 2018 PQ	Parenting Quality	289	−0,040
Kuzawa et al., 2009 partnered P	Parental Status	176	0.454
Kuzawa et al., 2009 non-partnered P	Parental Status	714	0.000
Lawson et al., 2017 PQ	Parenting Quality	81	0,406
Mascaro et al., 2013 PQ	Parenting Quality	58	0,553
Mascaro et al., 2014 P	Parental Status	131	0,622
Mazur, 2014 P	Parental Status	522	−0,204
Muller et al., 2009 Datoga P	Parental Status	80	0,058
Muller et al., 2009 Hadza P	Parental Status	25	0,959
Perini et al., 2012 P	Parental Status	67	0,538
Simon, 2012 PQ	Parenting Quality	127	0,113
Storey et al., 2000 P W	Parental Status	8	2,632
Storey et al., 2011R	Reactivity	12	0,581
Van Anders et al., 2012 cry RCT	Reactivity	25	1,157
Van Anders et al., 2012 responsive RCT	Reactivity	26	1,145
Van Anders et al., 2012 unresponsive RCT	Reactivity	26	0,000
Van Anders et al., 2014 cry RCT	Reactivity	45	0,000
Van Anders et al., 2014 responsive RCT	Reactivity	43	0,000
Van Anders et al., 2014 unresponsive RCT	Reactivity	45	0,000
Waldvogel and Ehlert, 2018 PQ	Parenting Quality	182	0,261
Weisman et al., 2014 baseline PQ	Parenting Quality	35	0,795

[a] R = Reactivity, PQ = Parenting Quality, P = Parental Status, W = Within-subject, RCT = randomized trial.

each individual participant was only included once, by aggregating multiple data-points pertaining to the specific hypothesis. The influence of potentially outlying single study outcomes on the overall combined effect sizes was examined using the one-study-removed approach (Borenstein et al., 2009).

Heterogeneity was assessed using $Q_{homogeneity}$ statistic as well as I-square (Borenstein et al., 2009; Higgins and Thompson, 2002). The Q-test examines the presence of heterogeneity against a null hypothesis of homogeneity. As Q is depending on the number of studies in the analyses, I-square is provided as well, because it may be a better indicator in the case of varying numbers of studies in the analyses. I-square describes the percentage of variation due to heterogeneity rather than chance. An I-square lower than 50% is indicative of homogeneity (Higgins et al., 2003). Although the included studies in the three study domains showed variety in design and measures, the sets of study outcomes showed substantial homogeneity of effect sizes. Accounting for remaining heterogeneity, we used random effects models to compute the combined effect sizes and 95% confidence intervals (CIs).

Small studies with null effects or unexpected results are less likely to be published and potential publication bias should be examined. As no single meta-analytic method consistently seems to outperform all others (Carter et al., 2019) we used several approaches including the trim and fill method (Duval and Tweedie, 2000a, b), the Begg and Mazumdar (1994) rank correlation test, and the Egger's regression intercept approach (Egger et al., 1997). In the domain of parental status sufficient significant outcomes were found to conduct a *p*-curve analysis examining the influence of publication bias and potential *p*-hacking (Simonsohn et al., 2015).

The effect of categorical moderators was tested using $Q_{contrast}$. Different effect sizes between subsets of a moderator are reflected by a

Table 2
Meta-analytic results for testosterone and parenting outcomes.

	k	N	Hedges' *g*	95% CI	$Q_{homogeneity}$	I^2	$Q_{contrast}$[a]
Reactivity studies	*12*	*766*	*0.19*	*−0.03 to 0.42*	*14.56*	*24.42*	
Relationship status	5	267	0.26	−0.10 to 0.62	6.243	35.90	0.31
with partner	7	499	0.13	−0.15 to 0.41	7.53	20.35	
without partner							
Parental status	5	534	0.18	−0.11 to 0.47	7.66	47.67	0.01
children of their own	7	232	0.21	−0.14 to 0.55	6.65	09.81	
mixed with & w/o children							
Testosterone sample	10	732	0.19	−0.04 to 0.43	13.98	35.61	n.a.
saliva	2	34	0.20	−0.54 to 0.93	0.54	00.00	
blood							
Design	6	210	0.25	−0.15 to 0.66	6.45	22.46	0.16
rct	6	556	0.16	−0.10 to 0.41	7.70	35.09	
non-rct							
Design	5	414	0.23	−0.15 to 0.61	7.69	47.98	0.06
within subjects	7	352	0.17	−0.11 to 0.44	6.79	11.63	
between subjects							
Parenting Quality	*18*	2,164	*0.14**	*0.03 to 0.24*	*22.07*	*23.00*	
Relationship status	17	2,151	0.16**	0.05 to 0.27	20.28	21.09	n.a.
with partner	1	13	0.04	−0.09 to 0.18	00.00	00.00	
without partner							
Parental status	17	2,137	0.14*	0.03 to 0.25	22.03	27.38	n.a.
children of their own	1	27	0.23	−0.94 to 1.39	00.00	00.00	
mixed with & w/o children							
Testosterone sample	15	2,020	0.08	−0.01 to 0.17	13.70	00.00	n.a.
saliva	3	144	0.54*	0.20 to 0.89	1.85	00.00	
blood							
Design	3	168	0.22	−0.09 to 0.53	1.58	00.00	n.a.
within subjects	15	1,996	0.14*	0.02 to 0.25	19.90	29.64	
between subjects							
Reporter of outcome	8	1,442	0.11	−0.03 to 0.26	11.49	39.08	0.03
self-report	3	134	0.40*	0.05 to 0.75	0.74	00.00	
other-report	7	588	0.13	−0.05 to 0.32	6.87	12.64	
observational							
Parental status	*20*	4,150	*0.22***	*0.09 to 0.35*	*56.10***	*66.13*	
Partnered	10	1,562	0.29**	0.10 to 0.48	22.13**	59.32	1.08
with	10	2,588	0.16	-0.02 to 0.33	27.33**	67.06	
without							
Testosterone sample	14	1,984	0.24**	0.09 to 0.40	30.06**	56.75	0.12
saliva	6	2,166	0.19	−0.05 to 0.43	24.95**	79.96	
blood							
Design	2	277	0.98	−1.37 to 3.33	3.17	68.48	n.a.
within subjects	18	3,873	0.23**	0.09 to 0.36	52.92**	67.88	
between subjects							

[k] = number of study outcomes; N = total sample size; 95% CI = 95% confidence interval around point estimate; $Q_{homogeneity}$ = homogeneity statistic; I^2 describes the percentage of variation due to heterogeneity rather than chance; $I^2 > 50\%$ means rather large heterogeneity, $I^2 < 50\%$ means homogeneous; $Q_{contrast}$ = statistic of contrast between moderator subgroups; n.a. = not applicable; a subgroups with $k < 4$ excluded from the comparison; b As study outcomes for 6 studies were meta analytically combined, k of outcome groups do not add up to k of the total set; *$p < 0.05$; **$p < 0.01$; contrasts were only tested with sub-sets $k \geq 4$.

significant $Q_{contrast}$ value. In line with published meta-analyses (e.g., Bakermans-Kranenburg et al., 2003), only subsets with at least 4 studies were included. To test for moderating effects of the continuous variables publication year and sample size, meta-regression analyses were performed. In these analyses, slopes (b) that significantly differ from zero (at $p < .05$) indicate a significant moderator.

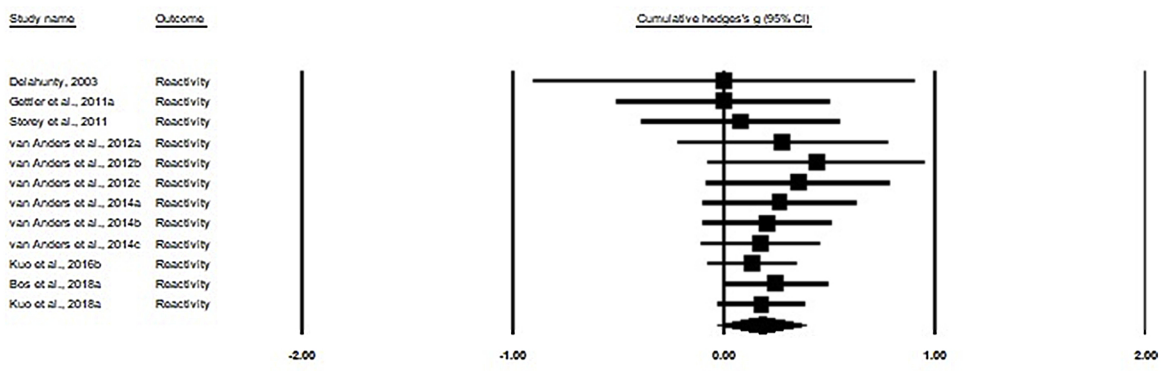
3. Results

3.1. Reactivity studies

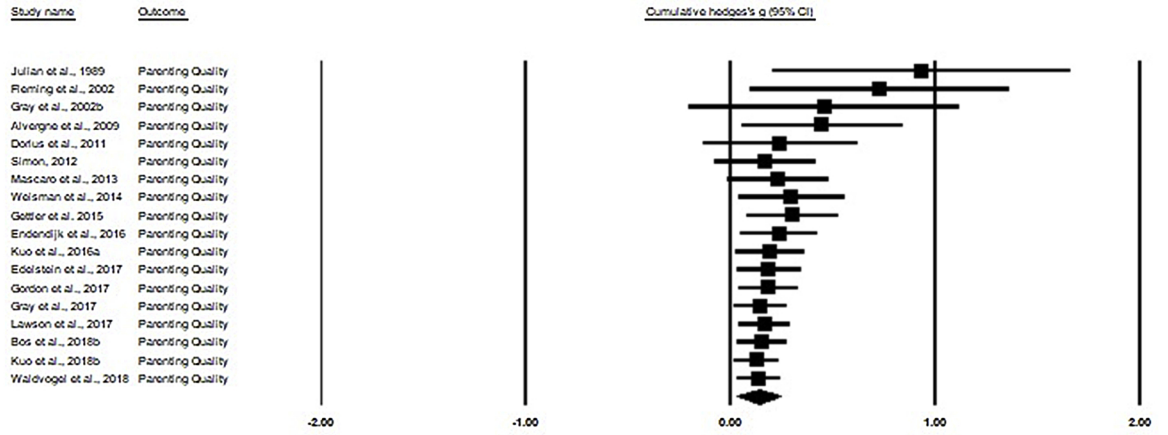
We found a non-significant combined effect size $g = 0.19$ (95% CI [−0.03, 0.42] in a homogeneous set of 12 reactivity studies (including 6 RCTs) with N = 766 participants ($Q = 14.56$, $p > 0.05$, $I^2 = 24.42$ see Table 2). Reactivity studies, mostly triggering caregiving behaviors, did not show a significant combined effect on testosterone levels. Excluding two study outcomes that might have triggered protective paternal behavior by using only cry stimuli (Van Anders et al., 2012, 2014), we did not find a stronger combined effect (Hedges' g = 0.15 (95% CI [−0.06, 0.35]). The one-study-removed approach did not result in a substantially different estimate of the combined effect size ($g = 0.18$ (95% CI [−0.03, 0.39]). Of the potential moderators, the sub-sets of studies different on partner relationship, parental status, and study design were large enough (≥4) to test the moderator contrasts. None of these variables moderated the effect size in this set of studies. The meta-regression with the continuous variables publication year and sample size did not show slopes that significantly differed from zero ($p > .05$).

Egger's regression intercept was not significant (intercept = 1.07, $p = .13$) but Begg and Mazumdar's rank correlation test appeared to be significant (Kendall's tau = 0.57, $p = .01$). The trim-and-fill analysis showed one study that needed to be trimmed and filled, resulting in a marginally lower and still non-significant combined effect size. The cumulative meta-analysis ranking studies according to publication year

a. Cumulative meta-analytic effect sizes for studies on reactivity



b. Cumulative meta-analytic effect sizes for studies on parenting quality



c. Cumulative meta-analytic effect sizes for studies on parental status

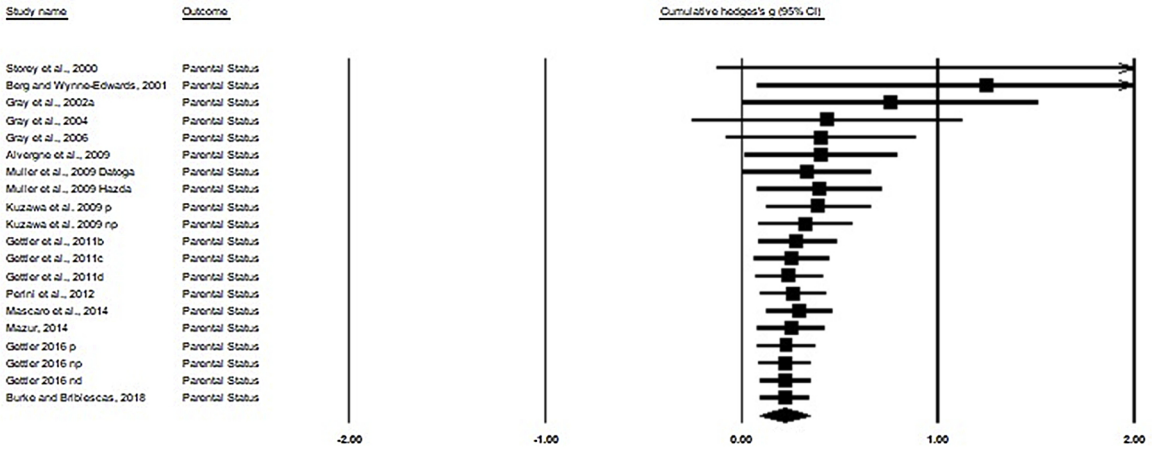


Fig. 2. Cumulative meta-analytic effect sizes for studies on (a) reactivity, (b) parenting quality, and (c) parental status, based on year of publication.
a. Cumulative meta-analytic effect sizes for studies on reactivity.
b. Cumulative meta-analytic effect sizes for studies on parenting quality.
c. Cumulative meta-analytic effect sizes for studies on parental status.

showed no trace of a winner's curse (see Fig. 2), as the first pioneering study by Delahunty (2003) showed a non-significant finding. Given that only three study outcomes were significant, *p*-curve analysis did not seem feasible.

3.2. Parenting quality

For 18 study outcomes (including N = 2,164 participants) on parenting quality, a significant combined effect size of Hedges' $g = 0.14$ (95% CI [0.03, 0.24]) was found in a homogeneous set of outcomes

(Q = 22.07, p > 0.05; I^2 = 23.00, see Table 2). The one-study-removed approach did not result in a different estimate of the combined effect size or 95% CI. Only one potential moderator, reporter of outcome, had large enough (≥ 4) sub-sets, but the moderator contrast between self-report and observational measures of parenting quality was not significant. Only one study did not include partnered males, and excluding this study resulted in a somewhat larger effect size (g = 0.16 (95% CI [0.05, 0.27]). The meta-regression with publication year showed a slope that was just not significant (p = .06) but the meta-regression with sample size resulted in a slope that was significantly different from zero (p < .001), with larger studies yielding smaller effect sizes.

Testing for publication bias, the Begg and Mazumdar's rank correlation test was significant (Kendall's tau = 0.37, p = .03). This was confirmed by Egger's regression intercept (1.56) that showed a significant asymmetry of the funnel plot (p = .007), indicating potential publication bias against small studies with small effects. Applying trim-and-fill showed that six studies needed to be trimmed and filled resulting in a non-significant adjusted effect size of 0.02 (95% CI [−0.04, 0.09). The cumulative meta-analysis, ranking studies according to publication year, showed a potential winner's curse (see Fig. 2) as the first pioneering study (Julian and McKenry, 1989) found a large effect size of Hedges' g = 0.94 in a sample of N = 37. This was the effect size with the highest and significant standardized z = 2.32 (p < .05) in the distribution of effect sizes (Fisher Z transformed) in this set of study outcomes. With only three study outcomes in this set being significant, p-curve analysis did not seem feasible. It should be noted that for females (N = 826 participants, 10 study outcomes), we found a non-significant combined effect size of Hedges' g = 0.04 (95% CI [−0.09, 0.18]) for parenting quality in a homogeneous set of outcomes (Q = 6.64, p > 0.05; I^2 = 00.00).

3.3. Parental status

For 20 study outcomes (with N = 4,150 participants) comparing testosterone levels of parents and non-parents, a significant combined effect size of Hedges' g = 0.22 (95% CI [0.09, 0.35]) was found in a heterogeneous set of outcomes (Q = 56.10, p < 0.01, I^2 = 66.13, see Table 2). Excluding the two studies comparing testosterone levels before and after birth within the same participants (within-group design), leaving 18 between-subject study outcomes, provided similar results (g = 0.23, 95% CI [0.09, 0.36]). The one-study-removed approach did not result in a different estimate of the combined effect size or 95% CI (g = 0.22 (95% CI [0.09, 0.35]). Two potential moderators had large enough (≥ 4) sub-sets to conduct a moderator test, partnership status and type of testosterone sampling (saliva or blood), but neither of them was significant. The meta-regression with publication year did not show a significant slope (p = .20), but the meta-regression with sample size resulted in a slope that was significantly different from zero (p = .02), with – again- larger studies showing smaller effect sizes.

Testing for publication bias, the Begg and Mazumdar's rank correlation test was significant (Kendall's tau = 0.41, p = .01). This was confirmed by Egger's regression intercept (2.40) that showed a significant asymmetry of the funnel plot (p = .002), indicating potential publication bias against small studies with small effects. Applying the trim-and-fill method showed that seven studies needed to be trimmed and filled, resulting in a non-significant adjusted effect size of Fisher Z = 0.04 (95% CI [−0.02, 0.11). The cumulative meta-analysis ranking studies according to publication year showed a potential winner's curse (see Fig. 2); the first pioneering study (Storey et al., 2000) found a very large effect size of Hedges' g = 2.63 in a sample of N = 8. This was the effect size with the highest and significant standardized z = 3.51 (p < .001) in the distribution of effect sizes (Fisher Z transformed). With eight significant study outcomes p-curve analysis did seem feasible (one significant negative outcome was not included in the p-curve analysis). The aggregate with the Stouffer method yielded a full p-curve Z = −1.88, p = .03, whereas the half p-curve was Z = −1.11, p = .13, hence the p-curve does not indicate evidential value (Simonsohn et al., 2015). For females (N = 1,234 participants, 6 study outcomes) a significant combined effect size of Hedges' g = 0.26 (95% CI [0.14, 0.38]) was found in a homogeneous set of outcomes (Q = 3.79, p > 0.05; I^2 = 00.00).

4. Discussion

The current series of meta-analyses provided equivocal support for the 'challenge hypothesis' as applied to fathers in their role of caregivers. The reactivity studies, triggering paternal behavior and responses through interaction with the child or exposure to infant cues, did not show a significant combined effect size. Having a child (parental status) and displaying more active paternal involvement or higher parenting quality were related to lower testosterone levels compared to (partnered) males without children or fathers with lower involvement or parenting quality. However, according to conventional criteria (Cohen, 1988) the combined effect sizes were small and publication bias might have inflated the meta-analytic results. In both research domains, the first pioneering but small and underpowered studies illustrated a 'winner's curse' (Button et al., 2013) as the strength of their findings have not yet been replicated. The most promising evidence for the challenge hypothesis is found in studies on the association of lower testosterone levels with the transition to parenthood, but it must be noted that most of these studies have low statistical power as well, with elevated risk of unpublished reports of studies with negative or null effects. Only few studies on testosterone as related to parental status in females have been conducted, but our preliminary finding of a significant albeit small combined effect size might be taken as an encouragement to further explore this association.

The challenge hypothesis suggests that testosterone levels of adult men are basically low but rise under conditions of competition and reproductivity, leading to a flexible adaptive neuroendocrine system (Archer, 2006; Gettler, 2016). For competitive mating, higher levels of testosterone are favorable, but lower testosterone levels are supposed to be more conducive for maintaining partner relationships and caring for offspring. As human babies are extremely helpless during the first few years of their life, intensive (bi-parental) care is essential for a child to survive (Hrdy, 2011). According to the challenge hypothesis, the downregulation of testosterone levels during the transition to parenthood would facilitate the shift from competition to care. We did indeed find lower levels of testosterone in males transiting to parenthood but the meta-analytic effect size should be considered small (Cohen, 1988). With a Hedges' g of 0.22, 59% of the parents would show a testosterone level below the mean of the non-parental individuals (Cohen's U3 = 59), and 91% of the two groups would overlap. In other words, if 100 men transit to parenthood, only about six men will have lower testosterone compared to if they had not made the transition to fatherhood (Magnusson, 2014).

Neuroscientific studies have been characterized as underpowered, and have been suggested to suffer from power failure (Button et al., 2013) with high risk of false positive results due to too small sample sizes. In the area of testosterone research, a similar power failure might exist. In general, larger studies lead to more precise estimates of the true population effect size. In the set of parental status studies, six studies had sample sizes larger than N = 200, and all of these had null effects or even negative effect sizes (see Table 1). In the domain of neuroendocrine studies of parenting, samples larger than 200 are not easy to conduct because data collection and data coding methods are

labor intensive, which creates a ceiling to the number of participants included. In the set of studies in our meta-analysis the median sample size was n = 58, with a range from 8 (Storey et al., 2000) to 877 (Gettler and Oka, 2016, partnered males). Using the combined effect size of $g = 0.22$ as the best indicator for the expected effect size, a study on the difference in testosterone levels between fathers and non-fathers would need a sample of more than 500 subjects to reach a power of .80 with an alpha level of 0.05. This indicates that the majority of the studies included in the current meta-analysis are underpowered and that for replications rather large samples are required. In this context it should also be noted that, compared to between-subject designs, studies with within-subject designs have superior power as a result of the reduced error term in within-subject studies (see for a numerical example Van IJzendoorn and Bakermans-Kranenburg, 2016). The current meta-analysis included some studies with a within-subject design, and their example may be followed more often in future studies.

After finalizing a previous draft of the current paper a meta-analytic review of the literature on "Pair-bonding, fatherhood, and the role of testosterone" was published (Grebe et al., 2019). That meta-analysis was broader than ours in examining also associations between pair-bonding and testosterone from the perspective of the challenge hypothesis. However, they also examined the association of parental status with testosterone levels, as well as the relation between parenting behavior (involvement) with testosterone. It should be noted that their literature search did not lead to a completely overlapping set of studies, due to different search strategies and somewhat different inclusion and exclusion criteria. Furthermore, different meta-analytic techniques were used, which could lead to different findings (Carter et al., 2019). It is therefore notable that the findings of both meta-analyses pertaining to fatherhood are converging, although their interpretation of the robustness of the outcomes turns out to be somewhat divergent from ours.

For the association between paternal behavior and testosterone Grebe et al. (2019) could not show sufficient evidential value, which converges with our conclusion that using stringent criteria the verdict should be that evidence for a relation is still absent. For parental status, however, Grebe et al. (2019) concluded that the effect was "robust and non-zero", within the range of r = .15 - .19, which is comparable to Hedges' g ranging from .30 to .40. Indeed, our estimate for the studies that focused on partnered fathers and comparisons (Hedges' g = .29) is close to the range of the estimates reported by Grebe and colleagues. For several reasons, however, we are somewhat less convinced of the robustness and reproducibility of this effect size. First, various indicators for publication bias pointed at bias against small studies with small effect sizes, and the trim and fill procedure yielded a non-significant combined effect. Second, with some important exceptions (e.g., Gettler and Oka, 2016) the statistical power of studies in this field is low, with an estimated sample size of at least 700 participants required to find the meta-analytic effect size in the next empirical studies. Results should therefore be interpreted keeping these limitations in mind, and the final verdict on the challenge hypothesis postponed until more evidence has been collected.

Not only a power failure but also a failure to take context into account might explain the inability to squarely support the challenge hypothesis on fatherhood based on the available studies. Downregulation of testosterone levels might only be expected if parents are actually and effectively involved in caring for their offspring and if renewed reproductive efforts or other competitive demands remain absent (Gettler, 2016). Context canalizes the regulation of testosterone production. In a study by Muller et al. (2009) testosterone levels of fathers who spent most of their time away from their families were comparable to those of non-fathers, indicating that biological fatherhood is much less important than social fatherhood. Furthermore, Van Anders et al. (2012) showed that testosterone levels of fathers who were exposed to infant crying sounds without the possibility of offering nurturing care did not decrease. The experience of active fathering seems a necessary condition for the downregulation of testosterone levels. Lastly, some studies suggested that testosterone levels rebound as the child gets older (Barrett et al., 2013; Kuzawa et al., 2010), maybe because competitive mating efforts increase (Gettler, 2016) or social competition intensifies if fathers have to provide for more offspring (Jasienska et al., 2012). The current set of testosterone studies did not allow for detailed examination of contextual factors such as the intensity and effectiveness of fathering or the presence of competing demands modulating testosterone levels into an upward direction.

Testosterone is not a single acting hormone but part of a much more complex neuroendocrine system of hormones and peptides. The interplay between testosterone and cortisol was predictive of parenting quality in fathers before and after the birth of their child (Bos et al., 2018). Besides testosterone, a number of other hormones and neuropeptides are probably involved in fathering (including, e.g., prolactin and progesterone; Bos, 2017), but besides testosterone three hormones may be of particular relevance in the context of parenting: oxytocin, estradiol, and vasopressin. Oxytocin administration was found to affect testosterone levels in fathers which subsequently influenced quality of the parent-child interaction (Weisman et al., 2014). Furthermore, testosterone is metabolized to estradiol, which in turn is critical for the synthesis of oxytocin (Choleris et al., 2008; Cornil et al., 2006). In our own experimental studies, we demonstrated increased parental sensitivity and decreased hostility in fathers' interactive play with their toddlers after intranasal oxytocin administration (Naber et al., 2010, 2013). High doses of oxytocin may lead to binding of oxytocin to vasopressin receptors, shifting the balance between oxytocin and vasopressin in the brain (Gimpl and Fahrenholz, 2001). Fathers with high vasopressin levels showed lower activation in circuits related to cognitive processing (inferior frontal gyrus, insula) when observing their own infant (Atzil et al., 2012). These complex interactions make the interpretation of isolated testosterone findings difficult without considering the wider neuroendocrine context. Unfortunately, the small number of multi-peptide studies, and the fact that these studies focus on different combinations of hormones and peptides in fathers, currently preclude a meta-analysis that takes the interaction between testosterone and other hormones into account.

To conclude, we did find a positive relation between lower testosterone levels and (elevated) involvement in parenting among fathers but due to possible bias and mostly underpowered studies our findings do not provide robust evidence for the challenge hypothesis. Large studies are needed to overcome the power failure, in particular to facilitate testing the role of moderators such as age of the child, number of children in the family, or more refined indicators of parenting quality and involvement. Downregulation of testosterone levels might only be expected if fathers are actually and effectively involved in caring for their offspring and if other competitive demands are absent. Given that testosterone is part of a complex neuroendocrine feedback system, we need studies that take the social and neuroendocrine context of testosterone regulation into account.

Funding

This study was supported by a European Research Council grant (ERC AdG 669249) awarded to M.J.B-K.

Declaration of Competing Interest

None.

Appendix A

Table A1

Table A1
Coding system.

Variable	Coding description/ categories
Background characteristics	
Type of publication	Journal article Dissertation Book chapter other
Year of publication	
Sample characteristics	
Gender of subjects	Males Females
Mean age	Age (years)
Ethnicity	% Caucasian.
Parental status subjects	Children of their own No children of their own Mixed: own children and no children Unknown
Relationship status	Partnered Not partnered
Study characteristics	
Type of intervention	Hormonal: Testosterone Homonal: other hormone Behavioral No intervention
Type of Testosterone sample	Saliva Blood, plasma Blood, serum Blood, not otherwise specified Other
Design characteristics	
Type of study	RCT Non-RCT Cohort design Cross sectional design Case control design Quasi-experimental design
Type of design	Within subjects Between subjects
Sample size	Number of participants in the included study outcome
Intention to treat	yes no NA (no RCT)
Blinding of subjects	yes no NA (no RCT)
Blinding of researchers	yes no NA (no RCT)
Reporter of outcome (physiological = observational)	Self report Other report Observational

References

Abraham, E., Feldman, R., 2018. The neurobiology of human allomaternal care; implications for fathering, coparenting, and children's social development. Physiol. Behav. 193, 25–34.

Alvergne, A., Faurie, C., Raymond, M., 2009. Variation in testosterone levels and male reproductive effort: insight from a polygynous human population. Horm. Behav. 56, 491–497.

Archer, J., 2006. Testosterone and human aggression: an evaluation of the challenge hypothesis. Neurosci. Biobehav. Rev. 30, 319–345.

Atzil, S., Hendler, T., Zagoory-Sharon, O., Winetraub, Y., Feldman, R., 2012. Synchrony and specificity in the maternal and the paternal brain: relations to oxytocin and vasopressin. J. Am. Acad. Child Adolesc. Psychiatry 51, 798–811.

Bakermans-Kranenburg, M.J., van IJzendoorn, M.H., Juffer, F., 2003. Less is more: meta-analyses of sensitivity and attachment interventions in early childhood. Psychol. Bull. 129, 195–215.

Barrett, E.S., Tran, V., Thurston, S., Jasienska, G., Furberg, A.-S., Ellison, P.T., Thune, I., 2013. Marriage and motherhood are associated with lower testosterone concentrations in women. Horm. Behav. 63, 72–79.

Begg, C.B., Mazumdar, M., 1994. Operating characteristics of a rank correlation test for publication Bias. Biometrics 50, 1088–1101.

Berg, S.J., Wynne-Edwards, K.E., 2001. Changes in testosterone, cortisol, and estradiol levels in men becoming fathers. Mayo Clin. Proc. 76, 582–592.

Borenstein, M., Hedges, L.V., Higgins, J., Rothstein, H.R., 2014. Comprehensive Meta-Analysis: a computer program from research synthesis (version 3). Biostat Inc., Englewood, NJ.
Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-analysis. Wiley, Ltd., New York, NY. https://doi.org/10.1002/9780470743386.
Bos, P.A., Hechler, C., Beijers, R., Shinohara, K., Esposito, G., de Weerth, C., 2018. Prenatal and postnatal cortisol and testosterone are related to parental caregiving quality in fathers, but not in mothers. Psychoneuroendocrinology 97, 94–103.
Bos, P.A., 2017. The endocrinology of human caregiving and its intergenerational transmission. Dev. Psychopathol. 29, 971–999.
Bos, P.A., Hermans, E.J., Montoya, E.R., Ramsey, N.F., van Honk, J., 2010. Testosterone administration modulates neural responses to crying infants in young females. Psychoneuroendocrinology 35, 114–121.
Brown, R.E., Murdoch, T., Murphy, P.R., Moger, W.H., 1995. Hormonal responses of male gerbils to stimuli from their mate and pups. Horm. Behav. 29, 474–491.
Burke, E.E., Bribiescas, R.G., 2018. A comparison of testosterone and cortisol levels between gay fathers and non-fathers: a preliminary investigation. Physiol. Behav. 193, 69–81.
Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafo, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.
Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46, 423–429.
Carter, E.C., Schönbrodt, F.D., Gervais, W.M., Hilgard, J., 2019. Correcting for Bias in psychology: a comparison of meta-analytic methods. Adv. Methods Pract. Psychol. Sci. 2 (2), 115–144.
Choleris, E., Devidze, N., Kavaliers, M., Pfaff, D.W., 2008. Steroidal/neuropeptide interactions in hypothalamus and amygdala related to social anxiety. Prog. Brain Res. 170, 291–303.
Clark, M.M., Galef Jr., B.G., 1999. A testosterone-mediated trade-off between parental and sexual effort in male mongolian gerbils (Meriones unguiculatus). J Comp Psychol 113, 388–395.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
Cornil, C.A., Ball, G.F., Balthazart, J., 2006. Functional significance of the rapid regulation of brain estrogen action: where do the estrogens come from? Brain Res. 1126, 2–26.
Delahunty, K.M., 2003. Hormonal Indicators of Paternal Care in Humans. A Longitudinal Study of First-time Parents. Memorial University of Newfoundland, Ottawa, Canada.
Dorius, C., Booth, A., Hibel, J., Granger, D.A., Johnson, D., 2011. Parents' testosterone and children's perception of parent-child relationship quality. Horm. Behav. 60, 512–519.
Duval, S., Tweedie, R., 2000a. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J. Am. Stat. Assoc. 95 (89), 98.
Duval, S., Tweedie, R., 2000b. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56, 455–463.
Edelstein, R.S., Chopik, W.J., Saxbe, D.E., Wardecker, B.M., Moors, A.C., LaBelle, O.P., 2017. Prospective and dyadic associations between expectant parents' prenatal hormone changes and postpartum parenting outcomes. Dev. Psychobiol. 59, 77–90.
Edelstein, R.S., Wardecker, B.M., Chopik, W.J., Moors, A.C., Shipman, E.L., Lin, N.J., 2015. Prenatal hormones in first-time expectant parents: longitudinal changes and within-couple correlations. Am. J. Hum. Biol. 27, 317–325.
Egger, M., Davey Smith, G., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. BMJ 315, 629–634.
Endendijk, J.J., Hallers-Haalboom, E.T., Groeneveld, M.G., van Berkel, S.R., van der Pol, L.D., Bakermans-Kranenburg, M.J., Mesman, J., 2016. Diurnal testosterone variability is differentially associated with parenting quality in mothers and fathers. Horm. Behav. 80, 68–75.
Feldman, R., Bakermans-Kranenburg, M.J., 2017. Oxytocin: a parenting hormone. Curr. Opin. Psychol. 15, 13–18.
Fleming, A.S., Corter, C., Stallings, J., Steiner, M., 2002. Testosterone and prolactin are associated with emotional responses to infant cries in new fathers. Horm. Behav. 42, 399–413.
Gettler, L.T., 2016. Becoming DADS: considering the role of cultural context and developmental plasticity for paternal socioendocrinology. Curr. Anthropol. 57 (S13), S38–S51.
Gettler, L.T., McDade, T.W., Agustin, S.S., Feranil, A.B., Kuzawa, C.W., 2015. Longitudinal perspectives on fathers' residence status, time allocation, and testosterone in the Phillippines. Adapt. Human Behav. Physiol. 1, 124–149.
Gettler, L.T., McDade, T.W., Agustin, S.S., Kuzawa, C.W., 2011a. Short-term changes in fathers' hormones during father–child play: impacts of paternal attitudes and experience. Horm. Behav. 60, 599–606.
Gettler, L.T., McDade, T.W., Feranil, A.B., Kuzawa, C.W., 2011b. Longitudinal evidence that fatherhood decreases testosterone in human males. Proc. Natl. Acad. Sci. U. S. A. 108, 16194–16199.
Gettler, L.T., Oka, R.C., 2016. Are testosterone levels and depression risk linked based on partnering and parenting? Evidence from a large population-representative study of US men and women. Soc. Sci. Med. 163, 157–167.
Gimpl, G., Fahrenholz, F., 2001. The oxytocin receptor system: structure, function, and regulation. Physiol. Rev. 81, 629–683.
Gordon, I., Pratt, M., Bergunde, K., Zagoory-Sharon, O., Feldman, R., 2017. Testosterone, oxytocin, and the development of human parental care. Horm. Behav. 93, 184–192.
Gray, P.B., Campbell, B.C., Marlowe, F.W., Lipson, S.F., Ellison, P.T., 2004. Social variables predict between-subject but not day-to-day variation in the testosterone of US men. Psychoneuroendocrinology 29, 1153–1162.
Gray, P.B., Kahlenberg, S.M., Barrett, E.S., Lipson, S.F., Ellison, P.T., 2002. Marriage and fatherhood are associated with lower testosterone in males. Evol. Hum. Behav. 23, 193–201.
Gray, P.B., Reece, J., Coore-Desai, C., Dinall, T., Pellington, S., Samms-Vaughan, M., 2017. Testosterone and jamaican fathers : exploring links to relationship dynamics and paternal care. Hum. Nat. 28, 201–218.
Gray, P.B., Yang, C.F., Pope Jr., H.G., 2006. Fathers have lower salivary testosterone levels than unmarried men and married non-fathers in Beijing. China. Proc Biol Sci 273, 333–339.
Grebe, N.M., Sarafin, R.E., Strenth, C.R., Zilioli, S., 2019. Pair-bonding, fatherhood, and the role of testosterone: a meta-analytic review. Neurosci. Biobehav. Rev. 98, 221–233.
Higgins, J.P., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. 21, 1539–1558.
Higgins, J.P., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. BMJ 327, 557–560.
Hrdy, S.B., 2011. Mothers and Others: the Evolutionary Origins of Mutual Understanding. Belknap Press of Harvard University Press, Cambridge, Massachusetts.
Jasienska, G., Jasienski, M., Ellison, P.T., 2012. Testosterone levels correlate with the number of children in human males, but the direction of the relationship depends on paternal education. Evol. Hum. Behav. 33 (665), 671.
Julian, T., McKenry, P.C., 1989. Relationship of testosterone to men's family functioning at mid-life: a research note. Aggress. Behav. 15, 281–289.
Kuo, P.X., Braungart-Rieker, J.M., Burke Lefever, J.E., Sarma, M.S., O'Neill, M., Gettler, L.T., 2018. Fathers' cortisol and testosterone in the days around infants' births predict later paternal involvement. Horm. Behav. 106, 28–34.
Kuo, P.X., Carp, J., Light, K.C., Grewen, K.M., 2012. Neural responses to infants linked with behavioral interactions and testosterone in fathers. Biol. Psychol. 91, 302–306.
Kuo, P.X., Saini, E.K., Thomason, E., Schultheiss, O.C., Gonzalez, R., Volling, B.L., 2016. Individual variation in fathers' testosterone reactivity to infant distress predicts parenting behaviors with their 1-year-old infants. Dev. Psychobiol. 58, 303–314.
Kuzawa, C.W., Gettler, L.T., Huang, Y.Y., McDade, T.W., 2010. Mothers have lower testosterone than non-mothers: evidence from the Philippines. Horm. Behav. 57, 441–447.
Kuzawa, C.W., Gettler, L.T., Muller, M.N., McDade, T.W., Feranil, A.B., 2009. Fatherhood, pairbonding, and testosterone in the Philippines. Horm. Behav. 56, 429–435.
Lawson, D.W., Nunez-de la Mora, A., Cooper, G.D., Prentice, A.M., Moore, S.E., Sear, R., 2017. Marital status and sleeping arrangements predict salivary testosterone levels in rural Gambian men. Adapt. Hum. Behav. Psychol. 3 (221), 240.
Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 6, e1000100.
Mascaro, J.S., Hackett, P.D., Rilling, J.K., 2013. Testicular volume is inversely correlated with nurturing-related brain activity in human fathers. Proc. Natl. Acad. Sci. U. S. A. 110, 15746–15751.
Mascaro, J.S., Hackett, P.D., Rilling, J.K., 2014. Differential neural responses to child and sexual stimuli in human fathers and non-fathers and their hormonal correlates. Psychoneuroendocrinology 46, 153–163.
Magnusson, K., 2014. Interpreting Cohen's D Effect Size. An Interactive Visualization. (Accessed February 1st, 2019. https://rpsychologist.com/d3/cohend.
Mazur, A., 2014. Testosterone of young husbands rises with children in the home. Andrology 2, 125–129.
Muller, M.N., Marlowe, F.W., Bugumba, R., Ellison, P.T., 2009. Testosterone and paternal care in East African foragers and pastoralists. Proc. Biol. Sci. 276, 347–354.
Naber, F., van IJzendoorn, M.H., Deschamps, P., van Engeland, H., Bakermans-Kranenburg, M.J., 2010. Intranasal oxytocin increases fathers' observed responsiveness during play with their children: a double-blind within-subject experiment. Psychoneuroendocrinology 35, 1583–1586.
Naber, F.B., Poslawsky, I.E., van IJzendoorn, M.H., van Engeland, H., Bakermans-Kranenburg, M.J., 2013. Brief report: oxytocin enhances paternal sensitivity to a child with autism: a double-blind within-subject experiment with intranasally administered oxytocin. J. Autism Dev. Disord. 43, 224–229.
Nunes, S., Fite, J.E., Patera, K.J., French, J.A., 2001. Interactions among paternal behavior, steroid hormones, and parental experience in male marmosets (Callithrix kuhlii). Horm. Behav. 39, 70–82.
Okabe, S., Kitano, K., Nagasawa, M., Mogi, K., Kikusui, T., 2013. Testosterone inhibits facilitating effects of parenting experience on parental behavior and the oxytocin neural system in mice. Physiol. Behav. 118, 159–164.
Perini, T., Ditzen, B., Hengartner, M., Ehlert, U., 2012. Sensation seeking in fathers: the impact on testosterone and paternal investment. Horm. Behav. 61, 191–195.
Rosenbaum, S., Gettler, L.T., McDade, T.W., Bechayda, S.S., Kuzawa, C.W., 2018. Does a man's testosterone "rebound" as dependent children grow up, or when pairbonds end? A test in Cebu, Philippines. Am. J. Hum. Biol. 30 (6), e23180.
Simon, C.D., 2012. Social Physiology in the Postpartum Period. Northwestern University, Evanston, Il. USA.
Simonsohn, U., Simmons, J.P., Nelson, L.D., 2015. Better P-curves: making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller. J. Exp. Psychol. Gen. 144 (6), 1146–1152.
Storey, A.E., Noseworthy, D.E., Delahunty, K.M., Halfyard, S.J., McKay, D.W., 2011. The effects of social context on the hormonal and behavioral responsiveness of human fathers. Horm. Behav. 60, 353–361.
Storey, A.E., Walsh, C.J., Quinton, R.L., Wynne-Edwards, K.E., 2000. Hormonal correlates of paternal responsiveness in new and expectant fathers. Evol. Hum. Behav. 21, 79–95.
Trainor, B.C., Bird, I.M., Alday, N.A., Schlinger, B.A., Marler, C.A., 2003. Variation in aromatase activity in the medial preoptic area and plasma progesterone is associated

with the onset of paternal behavior. Neuroendocrinology 78, 36–44.
Van Anders, S.M., Tolman, R.M., Jainagaraj, G., 2014. Examining how infant interactions influence men's hormones, affect, and aggression using the Michigan Infant Nurturance Simulation Paradigm. Father. A J. Theory Res. Pract. Men Father. 12, 143–160.
Van Anders, S.M., Tolman, R.M., Volling, B.L., 2012. Baby cries and nurturance affect testosterone in men. Horm. Behav. 61, 31–36.
Van IJzendoorn, M.H., Bakermans-Kranenburg, M.J., 2016. The role of oxytocin in parenting and as augmentative pharmacotherapy: critical issues and bold conjectures. J. Neuroendocrinol. 28 (8).
Voorthuis, A., Bakermans-Kranenburg, M.J., Van IJzendoorn, M.H., 2017. Testosterone reactivity to infant crying and caregiving in women: the role of oral contraceptives and basal cortisol. Infant Behav. Dev.
Waldvogel, P., Ehlert, U., 2018. Testosterone is associated with perceived constraint in early fatherhood. Adapt. Hum. Behav. Psychol. 4, 69–90.
Weisman, O., Zagoory-Sharon, O., Feldman, R., 2014. Oxytocin administration, salivary testosterone, and father-infant social behavior. Prog. Neuropsychopharmacol. Biol. Psychiatry 49, 47–52.
Wingfield, J.C., Hegner, R.E., Dufty, A.M., Ball, G.F., 1990. The "Challenge hypothesis": theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. Am. Nat. 136, 829–846.
Ziegler, T.E., Prudom, S.L., Zahed, S.R., Parlow, A.F., Wegner, F., 2009. Prolactin's mediative role in male parenting in parentally experienced marmosets (Callithrix jacchus). Horm. Behav. 56, 436–443.
Ziegler, T.E., Washabaugh, K.F., Snowdon, C.T., 2004. Responsiveness of expectant male cotton-top tamarins, Saguinus oedipus, to mate's pregnancy. Horm. Behav. 45, 84–92.