Neuroscience and Biobehavioral Reviews 98 (2019) 221–233

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Review article

Pair-bonding, fatherhood, and the role of testosterone: A meta-analytic review

Nicholas M. Grebe[a,*,1], Ruth E. Sarafin[b,1], Chance R. Strenth[b,1], Samuele Zilioli[c,d]

[a] Department of Evolutionary Anthropology, Duke University, Durham, NC, USA
[b] Department of Psychology, University of New Mexico, Albuquerque, NM, USA
[c] Department of Psychology, Wayne State University, Detroit, MI, USA
[d] Department of Family Medicine and Public Health Sciences, Wayne State University, Detroit, MI, USA

ARTICLE INFO

Keywords:
Testosterone
Pair-bonding
Challenge hypothesis
Fatherhood
Life-history theory
Meta-analysis

ABSTRACT

Males of many species must allocate limited energy budgets between mating and parenting effort. The Challenge Hypothesis provides a framework for understanding these life-history trade-offs via the disparate roles of testosterone (T) in aggression, sexual behavior, and parenting. It predicts that males pursuing mating opportunities have higher T than males pursuing paternal strategies, and in humans, many studies indeed report that men who are fathers and/or pair-bonded have lower T than childless and/or unpaired men. However, the magnitude of these effects, and the influence of methodological variation on effect sizes, have not been quantitatively assessed. We meta-analyzed 114 effects from 66 published and unpublished studies covering four predictions inspired by the Challenge Hypothesis. We confirm that pair-bonded men have lower T than single men, and fathers have lower T than childless men. Furthermore, men more oriented toward pair-bonding or offspring investment had lower T. We discuss the practical meaningfulness of the effect sizes we estimate in relation to known factors (e.g., aging, geographic population) that influence men's T concentrations.

1. Introduction

The hormone testosterone (T) possesses a wide range of physiological and psychological functions across vertebrates. However, much scientific and widespread interest in T focuses on its role in promoting male-typical behavior across species (see Fine, 2017 and Sapolsky, 2017, for two recent popular science examples). These psychological functions of T may be situated within a broader theoretical framework regarding the evolutionary biology of the endocrine system. Within this framework, diverse functions of T are conceived of as intertwined components of an adaptive resource allocation strategy: given environmental conditions (e.g., abundance of resources, risk of extrinsic mortality), individual conditions (e.g., mutation load, susceptibility to infection), and finite resource budgets, an organism must maximize its fitness by modifying the allocation of energy and effort towards certain classes of activities (e.g., growth, somatic maintenance) at the expense of others (e.g., reproduction) (Del Giudice et al., 2015). Hormones, including T, may constitute a major biological mechanism by which these coordinated trade-offs are achieved (Ketterson and Nolan, 1992). In particular, T may underlie variation in male reproductive strategies, due to its functions as a coordinating biological messenger that facultatively adjusts behavior, morphology, and physiology to secure reproductive opportunities and, in so doing, increase reproductive fitness (Wingfield et al., 1990; Gettler et al., 2011).

1.1. Reproductive fitness trade-offs in males

Reproductive fitness—an individual's success in passing on genes to the next generation—can be increased through investment in either mating effort or parenting effort. With mating effort, individuals emphasize finding and attracting partners and competing with rivals for these mating opportunities. This strategy, across males of many species, is characterized by increased aggressive behavior, risk-taking, and investment in costly ornamentation (Archer, 2006; McGlothlin et al., 2007; Ligon et al., 1990; Parker et al., 2002; Setchell et al., 2008; Rose et al., 1971; Muller and Wrangham, 2004). In contrast, individuals allocating more energy to parenting effort invest time and resources into long-lasting, stable mating relationships and care for offspring, either through provisioning or direct involvement in child rearing (Ziegler and Snowdon, 2000; Kaplan and Lancaster, 2003; Fernandez-Duque et al.,

* Corresponding author at: Department of Evolutionary Anthropology, Campus Box 90383, Durham, NC, 27708, USA.
E-mail address: nicholas.grebe@duke.edu (N.M. Grebe).
[1] Authors share co-first authorship and are listed in alphabetical order.

https://doi.org/10.1016/j.neubiorev.2019.01.010
Received 12 June 2018; Received in revised form 8 January 2019; Accepted 9 January 2019
Available online 09 January 2019
0149-7634/

2009). Although both strategies can increase reproductive fitness, there is an inherent trade-off between the two, such that allocation of energy toward one strategy reduces the pool of resources available for involvement in the other.

Relative to females, male vertebrates typically have a smaller obligate investment in offspring, permitting a higher degree of flexibility in the optimal balance between mating and parenting investment (Trivers, 1972; for counterarguments, see Kokko and Jennions, 2008). This balance can vary both *between* males and *within* males, depending on ecological conditions, social context, and genetic variation. Males in a highly 'fit' condition (whether due to intrinsic advantages such as favorable genes, advantageous environmental conditions that result in energetic surpluses, or both; Maynard-Smith, 1989) may obtain the greatest marginal benefits from investing in mating effort in the form of mate attraction and rival fighting. However, the costs associated with pursuing new mates (physical harm, pathogen contraction, etc.) and leaving potential offspring unsupported might hinder males that exclusively pursue this strategy from achieving optimal reproductive fitness (Kokko and Jennions, 2008). Similarly, males investing solely in parenting effort may miss mating opportunities that can result in surviving offspring with little investment on their part. Therefore, as with many life-history trade-offs, strategies richly embody contingencies: certain aspects of the internal and external environment favor investment in one kind of strategy, but a shift in these conditions may lead to an adaptive shift in the balance between mating and parenting effort. These shifts are neither conscious nor instantaneous; instead, it is thought that they result from physiological and neuromodulatory effects that unfold over time as regulated by T.

1.2. The Challenge Hypothesis

A theoretical model encompassing trade-offs between mating and parenting effort within males, the ability of males to switch between varying reproductive strategies, and the role of T during these switch points was first developed in avian seasonal breeders. Wingfield et al., (1990) found that baseline T levels increased at the beginning of the breeding season, which appeared to facilitate mate acquisition and territory formation. During confrontations with other males, T levels surged from the new baseline to the physiological maximum; these surges predicted increased aggressive behaviors, which aided in defending mates and territory. At the end of the breeding season, birds' T decreased as they maintained their pair bonds and provisioned offspring. In short, these males were shifting between mating-dominant and parenting-dominant strategies, with T mediating the behavioral changes that reflected these strategies. Wingfield and colleagues dubbed this framework on T and male reproductive strategies "The Challenge Hypothesis" (CH): in their formulation, males increase their mating effort in response to mating opportunities and challenges from other males, and T surges—at multiple timescales—permit this reallocation of effort. Although the CH was originally formulated to explain within-male behavioral shifts in avian species, it has since spawned a large body of supporting evidence conducted at multiple levels of analysis, including between-male comparisons and examinations across multiple animal taxa.

Experimental research in birds comparing T-treated males to controls has corroborated the hormone's role in controlling shifts between reproductive strategies. Male sparrows injected with T competed more with other males and fed their young far less frequently than controls, whereas birds treated with flutamide, an androgen receptor antagonist, showed the opposite pattern (Hegner and Wingfield, 1987). Dark-eyed junco males treated with T were more attractive to females, but they strayed further from the nest after their offspring hatched (Ketterson and Nolan, 1999). Male house finches with experimentally increased T fed offspring less frequently but sang, an index of mating effort, more frequently than controls (Stoehr and Hill, 2000). Similar effects have been found in Lapland longspurs, jays, and western screech owls (Hunt et al., 1999; Vleck and Brown, 1999; Herting and Belthoff, 1997). Importantly, however, predicted associations between T and mating/parenting effort have not been universally found in birds. Some systematic analyses suggest that the link between T and parenting behavior in birds may be restricted to certain passerine species (Hirschenhauser et al., 2003). A recent review of pair-bonding in the zebra finch presents mixed evidence for associations between T and pair-bonding, suggesting that effects may be context-specific (Prior and Soma, 2015).

Though parental care is relatively rare in reptiles and fishes, evidence supporting the CH has also been found in a number of these species. In teleost fishes, rises in T and 11-ketotestosterone (11-KT) have been associated with the display of dominant behaviors and increases in territoriality during mating season, but these hormones decrease outside of the mating season or when a male is providing paternal care, albeit not universally across species (Cardwell et al., 1996; Oliveira et al., 1996; Cardwell and Liley, 1991; Francis and Fernald, 1993; Mayer et al., 1993; Kindler et al., 1989; Sikkel, 1993; see Oliveira et al., 2002 and Hirschenhauser and Oliveira, 2006 for reviews). A similar pattern has been found in amphibian species (Townsend and Moger, 1987; Orchinik et al., 1988). In reptiles, T is linked to social rank, dominance, male-male competition, and aggression (Greenberg and Crews, 1990; Schuett et al., 1996; Thompson and Moore, 1992). One intriguing recent paper reports that estradiol-17β, in addition to androgens, increases in response to competition in male cichlids (Scaia et al., 2018), suggesting a possible role for estrogens in the CH.

As our analysis concerns the role of T in human males, perhaps the most relevant comparative evidence hails from non-human primates. Unlike birds, many primate species do not have specific breeding seasons; of those that do, increases in baseline T as predicted by the CH are observed (Dixson and Lunn, 1987). Yet it appears that non-seasonal breeders might still shift between mating and parenting effort via the effects of T. Lemurs, mandrills, and chimpanzees exhibit increased T and aggressive behavior when in the presence of a parous female, and male tamarins exhibit increased T and arousal behaviors when presented with the scent of an ovulating female, suggesting that males switch to mating effort when mating opportunities are salient (Cavigelli and Pereira, 2000; Setchell et al., 2008; Muller and Wrangham, 2004; Sobolewski et al., 2013; Ziegler et al., 2005). Moreover, T levels increase in male tamarins coinciding with their partners' ovulation, which may function to increase reproductive success (Ziegler et al., 2004). In group-established male howling monkeys, T levels and aggression increase with the threat of an outside male (Cristóbal-Azkarate et al., 2006). As in birds, fathering behaviors in primates are correlated with a drop in T. Male marmosets and siamangs that carried their offspring and participated in more paternal care had lower T, suggesting that males switch to parenting effort in these situations (Nunes et al., 2001; Morino, 2015). Marmoset fathers exposed to the scent of their infant experienced a drop in T (Prudom et al., 2008). In sum, there is some evidence to suggest that the CH may also apply to non-human primates.

1.3. The Challenge Hypothesis and humans

Researchers have recently turned to examine the strength of evidence in favor of the CH in humans (e.g., Archer, 2006; Wingfield, 2017). Several original predictions of the CH concern links between T and aggression, and in humans, these predictions have been both reviewed (e.g. Carré & Archer, 2017; Wingfield, 2017) and meta-analyzed (Archer, 2006). However, a number of other predictions stemming from the CH more broadly concern the role of T in the balance between mating and parenting effort in men (e.g. Burnham et al., 2003; Gettler et al., 2011); these predictions have not been subjected to a formal meta-analysis. Below, we outline these predictions as adaptations of the CH for human mating systems.

Human breeding systems are characterized by a wide diversity of

mating systems (e.g., monogamy, polygyny, polygynandry) as well as configurations of offspring care (e.g., maternal, paternal, communal). Human sexual activity is not confined to a specific season or particular window of female sexual receptivity (Thornhill and Gangestad, 2008); consequently, CH-inspired predictions for humans will differ in some respects from those in other species. However, because T's coordination of physiological and psychological effort toward mate acquisition is thought to be conserved across animal taxa (Roney and Gettler, 2015), some predictions made for humans will closely resemble those advanced for other species.

1) Men's baseline T levels will not differ between seasons. They will, however, covary positively with men's mating effort. Thus, single men, who are presumably actively searching for mates, will have higher T than pair-bonded men, who are less likely to be actively seeking for mates (e.g. Burnham et al., 2003; but see prediction 2).
2) Within paired men, individuals who report greater commitment or investment in their current relationship will have lower T concentrations than men who report less commitment or greater interests in extra-pair sexual opportunities (both reflections of increased mating effort; e.g. McIntyre et al., 2006).
3) Within single men, individuals who report a greater number of sexual partners, and those with less restricted sociosexuality (Simpson and Gangestad, 1991)—both indicative of greater investment in acquiring new mates—will have higher T concentrations (e.g., Puts et al., 2015).
4) Fathers, who are presumed to invest at least some degree of effort in parental care, will have lower T levels than non-fathers (e.g. Gettler et al., 2011).
5) Fathers with a greater degree of involvement in parenting their offspring will have lower T levels than fathers with minimal investment in parenting (e.g. Weisman et al., 2014).

1.4. The current analysis

Dozens of empirical studies have investigated the above predictions, and narrative reviews have, in general, concluded that these predictions are supported by scientific evidence (see, e.g., Ellison and Gray, 2009). Nevertheless, a meta-analysis of this literature is timely for several reasons. First, some scholars argue that only some of these predictions are supported in humans; for example, Mazur (2017) argues that marriage/pair-bonding, but not fatherhood, should predict a decrease in T. Our analyses will be able to adjudicate disagreements such as these through a quantitative analysis of the literature as a whole. Further, the precise effect size of CH-derived comparisons is unknown. Statistical significance need not imply practical significance, but through a meta-analysis, we gain the ability to provide an accurate estimate of T differences between groups of men, which can be compared to other factors known to relate to changes in T, such as aging, certain medical conditions, or exogenous administration. Lastly, at least two characteristics of the CH literature in humans present challenges to theoretical interpretation that can be fruitfully addressed in a meta-analysis. First, there exists no single standard method to analyze whether 'relationships' or 'fatherhood' predict decreased T—for instance, studies may include or exclude covariates, measure T from samples taken at various times of day, and may adopt different operational definitions of 'pair-bonded'. This analytic flexibility has recently been scrutinized in the psychological literature as a major obstacle to determining the 'true' support for an effect (Simmons et al., 2011). Second, CH effects, like the vast majority of empirical findings in the social sciences, are disproportionately drawn from WEIRD (Western, Educated, Rich, Industrialized, and Democratic; Henrich et al., 2010) populations. In the current review, 73% of the effects in our dataset come from Western samples. Thus, in our analyses we tested how subjective analytic decisions and the disproportionate representation of certain populations might affect effect size estimates, and in so doing, we also provide recommendations for future studies.

2. Methods

2.1. Search strategy

We located studies through multiple channels, including reference sections of published articles, online database search engines, and email and personal correspondence with researchers in this area. Our familiarity with the literature on human male behavioral endocrinology, as well as the list of studies cited in Gray and Campbell (2009), acted as a starting point for our search. We next performed searches on Google Scholar and Web of Science using "relationship status testosterone", "romantic relationships testosterone", "human parental testosterone", and "endocrinology of social relationships" as search phrases. Lastly, with the goal of locating unpublished data and manuscripts not identified through these methods, we emailed colleagues known to have conducted human-subjects research on the behavioral correlates of T (whether or not this research was specifically framed in terms of the Challenge Hypothesis) and requested data that could be included in the meta-analysis. We discontinued our literature search in October 2017.

We first restricted our search to studies that assessed relationships between two narrowly-defined predictor variables (pair-bond status and fatherhood status) and T concentrations (whether assessed through blood or saliva samples). Studies that assessed T indirectly, such as via assessments of masculinity, voice pitch, or fluctuating asymmetry were not included. However, a substantial number of effects pertinent to our predictions assessed continuous characteristics rather than the binary variables of fatherhood or pair-bond status. In these studies, T level was usually a predictor of behavioral outcomes such as time spent with children, relationship satisfaction, or interest in extra-pair copulations. We thus included these effects in two additional categories, grouped as pair-bond behaviors and fathering behaviors. Henceforth, we refer to analyses on pair-bond status and fatherhood status as our "primary analyses", and those on pair-bonding behavior and fathering behaviors as our "secondary analyses". We also chose to limit our analysis to heterosexual men. Though previous research suggests that women, but not homosexual men, experience reductions in T during pair-bonding (van Anders and Watson, 2006; van Anders and Goldey, 2010), a dearth of studies concerning these populations limit the utility of meta-analyses. Our initial search identified 127 relevant effects from 49 published manuscripts and 31 unpublished effects.

2.2. Inclusionary criteria

The 127 total effects were reduced to a working data set that balanced the desire to include as many effects as possible while limiting the dataset to only include effects that would facilitate a meaningful examination of the CH. Thus, we had a number of criteria that determined which effects would be included in the analyses:

1 Men's T concentrations decline across the lifespan (e.g., Kelsey et al., 2014); this presents a potentially important confound because fathers and men in committed relationships may be older on average than single and/or childless men. In some cases, effects were presented in papers both as raw T differences and, separately, adjusting for age (whether via including it as a covariate in the statistical model, analyzing a cohort of men across time, or matching paired men and/or fathers with age-matched controls). Whenever possible, we selected the results controlling for age, as it likely represented the more accurate estimate of the effect of interest. However, we also included effects that did not control for age when they were the only estimates available in the manuscript. When requesting unpublished effects, we asked all authors to share data or unpublished comparisons including age as a covariate. In the results, we compare the strength of age-controlled compared to non-age-controlled

effects.

2 In many cases, we included multiple effects from the same paper. We elected to do this when effects represented distinct pieces of information despite their non-independence—this is distinct from the criterion described above, because non-age-controlled samples provided no additional value when age-controlled comparisons were available. Most commonly, we included multiple effects from papers when authors reported one set of results for morning samples, and one for afternoon samples (e.g., Berg and Wynne-Edwards, 2001; Gray et al., 2006; Muller et al., 2009). Other manuscripts reported multiple operationalizations of an effect of interest (e.g., both relationship 'commitment' and 'satisfaction'; Hooper et al., (2011)). We control for the non-independence of these effects with multilevel analyses (see below).

This reduced set of effects consisted of 114 total effects: 60 for relationship status, 28 for fatherhood status, 16 for relationship behaviors, and 10 for fathering behaviors. All effects are described and categorized in a spreadsheet contained in our Supplemental Online Materials (SOM), and at https://osf.io/4r3a5/.

2.3. Obtaining effect sizes and coding moderators

Studies reported effects as *t*-statistics, *F*-statistics, or Pearson *r* correlations; unpublished effects were provided to us as *t*-statistics, Pearson *r* correlations, or raw data from which we calculated *t*-statistics. In some cases, effects of interest were not reported in the manuscript but were presented in figures or graphs. For six effects, means and standard deviations/standard errors were extracted from published figures using an online application (http://arohatgi.info/WebPlotDigitizer/), which were then used to calculate test statistics. All effects were converted into Fisher's *z*-transformation for meta-analytic estimates and transformed back into Pearson correlations for reporting results.

Pair-bond status. Pair-bond status effects (k = 60, 38 published) were included if they compared the T levels of two or more groups of men as grouped by pair-bonding status, though the operational definition of "pair-bonded" differed between studies. In our meta-analytic dataset, 15 considered only married men as pair-bonded, two only considered unmarried men in committed relationships, and 12 contained a mix of married and unmarried men in the pair-bonded group. For the remaining 31 effects, the distinction was unclear. However, due to the large global variation in mating systems, we left it to the original researchers to determine what constituted 'paired' vs. 'unpaired' and included the study as long as a distinction was made. Twelve effects included fathers in the comparison and nine did not; it is unclear in the remaining 39 effects. Of the 38 published effects, 35 collected T via saliva and three via blood (serum or plasma). Forty-four of the 60 effects (73%) of effects came from Western samples. See SOM for coding.

Pair-bond behavior. Pair-bond behavior effects (k = 16, 12 published) were diverse and included T's associations with relationship satisfaction, relationship commitment, relationship length, interest in novel partners, number of sexual partners, and sociosexual orientation. Effects were coded such that higher values represented greater mating effort (e.g., less restricted sociosexual orientation, lower relationship commitment, higher number of sexual partners). Nine of these effects examined paired men, three examined single men, and three provided insufficient information on relationship status. Of the effects, several (k = 9) assessed relationship satisfaction or commitment via an existing measure on relationship quality such as the Investment Model Scale (IMS, Rusbult et al., 1998) or the relationship satisfaction scale (Hendrick, 1988). Others considered individuals' interest in extra-pair partners (k = 1), number of sexual partners (k = 2), mating success (a composite score combining previous sexual experiences such as age of the first intercourse and number of partners; k = 1) or sociosexual orientation (Simpson and Gangestad, 1991; k = 3). Fourteen of 16 effects (88%) came from Western samples.

Fatherhood status. Fatherhood status effects (k = 28; published = 22) included 27 between-subjects effects and one within-subjects effect that assessed men's T levels before and after the birth of their first child. Of the between-subjects effects, studies differed regarding the pair-bond status of participants. If paired men do indeed have lower T than single men, and fathers are more likely to be paired than non-fathers, then this variation might confound any fatherhood status effect. Again, given the large global variation in relationship and paternity norms, we included the effect if it separated fathers from non-fathers, regardless of pair-bond status. However, to assess the impact of a "pair-bond confound", we coded the extent to which comparisons of fathers to non-fathers also compared single to paired men. Effects fell into one of four categories: only men (fathers and non-fathers) with the same relationship status were compared (no confound; k = 6); some but not all men were compared that had different relationship statuses (partial confound; k = 4); all fathers were paired, and all non-fathers were unpaired (full confound; k = 4); there was insufficient information regarding pair-bond status (unknown; k = 14). Of the published effects, 18 collected T via saliva and four via blood. Seventeen of 28 effects (61%) were drawn from Western samples.

Fathering behavior. Fathering behavior effects included ten published effects. Fathering behaviors were assessed in variety of ways, including partner reports of involvement (k = 2), a self-report composite of 'male parenting effort' (see Gray et al., 2002; k = 1), experience as a parent (k = 1), reaction to infant cries (k = 1), affectionate touch (k = 1), gaze towards infants (k = 1), use of 'motherese' (high-pitched, rhythmic speech directed toward infants; k = 1), and time spent with the offspring (k = 1). One effect assessed "caregiving behaviors," but the authors did not further operationalize this variable. Eight of these 10 effects came from Western samples.

2.4. Data analysis plan

All analyses were conducted on Fisher's *z*-transformed correlation coefficients. *F* and *t*-statistics were converted to *r* using formulas in Borenstein et al., (2011); Kendall's tau values were converted to *r* using the formula provided by Walker (2003). For the binary domains of relationship status and fatherhood, *r* represents the point-biserial correlation between pair-bond/fatherhood status and T concentrations; for the continuous domains, *r* represents the linear association between indices of either 'pair-bonding behavior' or 'fathering behavior' and T concentrations. We conducted four sets of analyses, one for each of the domains identified above: pair-bond status, pair-bond behaviors, fatherhood status, and fathering behaviors. For each set, we conducted a series of analyses to establish a plausible range of effect sizes, in recognition of the different strengths and weaknesses that individual techniques possess (Simonsohn et al., 2014a; McShane et al., 2016) and the lack of consensus regarding how best to correct for bias in meta-analyses (Carter and McCullough, 2018). The techniques we used for our analyses, and the accompanying justifications, are detailed below.

2.4.1. Traditional meta-analyses

Our traditional meta-analyses were conducted using multilevel modeling, which specified random effects at two levels: effects nested within studies, and studies within the overall set of effects. This approach permitted us to include multiple, non-independent effects from the same study when estimating mean effect size within domains (see Raudenbush and Bryk, 2002); it also conceives of the true mean *r* within a given domain as varying over the population of studies. Finally, this approach also allows for the examination of the effect of study-level moderators in meta-regressions. Effects were weighted by the inverse of their variance, providing more weight to more precisely estimated effects in the dataset (Borenstein et al., 2011). All multilevel analyses were conducted using the 'metafor' package (Viechtbauer, 2010) in R version 3.3.1. Because these analyses utilized multiple effects per study,

included both significant and non-significant results, and also included unpublished effects, they have the advantage over other techniques of providing an estimate based on the largest overall sample size.

Potential sources of bias have been identified in traditional meta-analytic techniques. Publication bias, the increased chance for statistically significant findings to be published compared to non-significant effects, is one of the oldest and most well-known sources (Sterling, 1959). Because of strong incentives to report statistically significant effects, publication bias is likely a ubiquitous feature of scientific literature (Simonsohn, 2012), which, if left unaddressed, can lead to substantial overestimates of an underlying effect. Two types of corrections for publication bias are commonly pursued. First, meta-analysts attempt to reduce the influence of publication bias by seeking out the entirety of published *and* unpublished studies to include in analyses. We attempted to do this for our analyses (see Search Strategy). However, because meta-analysts are unlikely to identify all unpublished studies to overcome publication bias completely, meta-analyses may still overestimate effect size. Thus, a second type of correction concerns methods that statistically adjust estimates (e.g., trim-and-fill; Duval and Tweedie, 2000; PET-PESSE; Stanley and Doucouliagos, 2017). While gaining popularity in the meta-analytic literature, these types of analyses may actually lead to estimates of effect size more biased than those derived from uncorrected analyses, especially in datasets that fail to conform to idealized assumptions (e.g., homogeneous effects, little to no publication bias)—and violations of these assumptions are likely very common in real-world datasets (Ledgerwood, 2016; for simulations and criticism of trim-and-fill, see Terrin et al., (2003) and Simonsohn et al., (2014b); for simulations and criticism of PET-PEESE, see Carter et al., (2017) and Simonsohn, (2017). For this reason, we chose not to calculate estimates using these statistical corrections for publication bias. At the same time, we acknowledge that we were likely not able to recover every unpublished effect. Thus, estimates from these analyses may represent slight overestimates, or at least 'upper bounds', of a plausible effect size for a given domain. To help establish a realistic range of effect sizes, we opted to also perform several alternative meta-analytic analyses that have recently been argued to provide adjusted effect size estimates with minimal bias.

2.4.2. *P-curve and alternative selection models: alternative forms of meta-analysis*

Importantly, some features of the source data itself may lead true effect sizes to be overestimated, and even the inclusion of unpublished effects to 'open the file drawer' may not be sufficient. *P*-hacking—a term referring to the assortment of subjective, defensible decisions in data collection and analysis that researchers can exploit to artificially inflate the likelihood of obtaining statistically significant effects—distorts the literature as a whole by "replacing" null effects with larger, statistically significant effects (Simonsohn et al., 2014aa). Hence, in the presence of *p*-hacking, meta-analyses that account for publication bias *per se* may still detect a "true" effect size greater than zero when in fact none exists (e.g., Harris et al., 2014).

P-curve is a procedure developed to detect *p*-hacking (Simonsohn et al., 2014a, 2014b). The *p*-curve (Simonsohn et al., 2014a) is a distribution of *p*-values that are published, statistically significant (i.e., ranging from ˜0 to .05) and in the predicted direction for a given research domain. The shape of the *p*-curve provides diagnostic information regarding the evidential value in a set of studies (versus the influence of *p*-hacking and publication bias). Given even modest statistical power (30%), a *p*-curve examining true effects will be markedly right-skewed, with 43% of *p*-values under .01 (e.g., Hung et al., 1997).

Moreover, *p*-curve analyses can do more than simply detect whether reported effects are real or spurious; they can also generate estimates of effect size. As such, *p*-curve constitutes an alternative or supplement to traditional meta-analysis. *P*-curve furthermore has several desirable features relative to traditional meta-analysis. One need not comprehensively sample all reported effects, or search for unpublished findings; only significant published values are needed. In absence of *p*-hacking, *p*-curve returns unbiased estimates of mean true effect size, unaffected by publication bias. This remains true when the effects included in *p*-curve analyses are heterogeneous (Simonsohn et al., 2014a, 2014b; Gervais, 2015).[2]

When *p*-hacking has affected results, *p*-curve underestimates true effect size (Simonsohn et al., 2014a,b). As this bias is the opposite direction of the bias in traditional meta-analysis, *P*-curve analysis may be a valuable complement as a more 'lower-bound' estimate.

All *p*-curves were run using the latest app on *p*-curve.com, *p*-curve 4.05. This app yields *Z*-tests for right skew, left skew, and a curve flatter than one with 33% power. Negative *Z*-values indicate an effect in the expected direction. We estimated effect sizes with the *p*-curve method using the R code reported by Simonsohn et al. (2014b), which yields an estimate by minimizing the Kolmogorov-Smirnov fit statistic to observed *p*-values (for details, see Simonsohn et al., (2014b). Although Simonsohn and collogues expressed these values as Cohen's *d*, we convert them to Pearson's *r* to match the estimates from the traditional meta-analysis. Lastly, we also estimated mean effect size using the *p*-uniform method (van Assen et al., 2015). This procedure employs a model identical to *p*-curve—and thus generates nearly identical estimates of effect size—but uses an alternative estimator that has the added benefit of yielding a 95% confidence interval, which *p*-curve does not readily provide (for details, see Simonsohn et al., 2014b; van Assen et al., 2015). All such estimates were generated using the 'puniform' R package (van Aert, 2016).

McShane et al. (2016) recently offered a critical discussion of using *p*-curve as a meta-analytic tool, emphasizing an important point: *p*-curve and *p*-uniform assume that the overall set of effects in an analysis—published and unpublished, significant and non-significant alike—is homogeneous, which may not be realistic. Indeed, statistically significant effects (for which *p*-curve *does* estimate an accurate effect size in the face of heterogeneity) are not a random subset of all effects. Most often, the chance that a finding is deemed significant scales positively with the size of the true effect investigated. Especially when mean true effect size is small, and heterogeneity of true effects in the population is large, the bias resulting from assuming homogeneity in a set of effects can be substantial (McShane et al., 2016).

Alternative selection models have been developed that estimate not only a true mean effect but also the heterogeneity of effects (e.g., the standard deviation of effect size in the population, τ) based on the variability of observed effect sizes relative to that expected from sampling variability alone. Simulations show that these selection models can estimate mean true effect size without bias across a variety of conditions (Hedges and Vevea, 1996; see also McShane et al., 2016). As McShane et al. (2016) acknowledge, estimated effect sizes generated by heterogeneous selection models on significant effects only are inherently very unstable. Fortunately, instability lessens when published non-significant effects are included in the analysis, and thus the most desirable models (1) allow for the inclusion of published non-significant effects while (2) simultaneously accounting for effect size heterogeneity (Hedges and Vevea, 2005).

Under publication bias, non-significant effects are presumed to have less chance to appear in the literature; however, selection models can actually estimate or assume varying chances of non-significant effects appearing in the literature, relative to significant effects. The effects of publication bias on estimation can be examined through sensitivity analyses. For instance, one can compare estimates of mean effects

[2] We note that the p-curve authors and other groups disagree on how well p-curve handles heterogeneity. Disagreements may stem from differing interpretations on whether an effect size estimate should reflect the average of all possible conducted studies, or the average of studies included in the p-curve. See http://datacolada.org/67.

assuming that non-significant effects have 20%, 30%, or 40% the chance of being published relative to significant effects. We estimated effect sizes using alternative selection models that included non-significant findings from published studies. We estimated findings assuming non-significant effects had 20% or 40% the chance of being published compared to significant findings in the predicted direction. The former level is in a range McShane et al. (2016) claim is likely representative of most literatures. The latter level could apply to literatures in which authors can argue that failures to replicate high profile findings can meaningfully contribute to the literature. In these analyses, we also permitted true effect sizes to vary across studies. Hence, we estimated, within subsets of study populations, the standard deviation of effect sizes τ. These analyses were performed on the published effects only, as these models account for publication bias only; including unpublished effects may lead to downwardly biased estimates. All analyses were performed using the R code provided in the supplementary material from McShane et al. (2016).

2.4.3. Summary of analyses performed

For each domain of effects, we performed the same series of analyses. First, we performed a simple multilevel analysis with random effects at the effect and study level. Second, we assessed whether publication status moderated mean effect size estimates; specifically, we performed a meta-regression to estimate whether, consistent with an influence of publication bias, published effects were significantly larger than unpublished effects in a given domain. We also assessed the impact of variation in conceptualization and measurement between studies (operationalization of 'pair-bond status', the confounding of pair-bond status with fatherhood status, inclusion of an age covariate, Western versus non-Western samples) via moderator analyses. Third, we performed *p*-curve and *p*-uniform analyses on a given set of effects. For these analyses, we reduced a set of effects to only include independent, published, and statistically significant effects in the published direction, as specified by Simonsohn et al. (2014a, 2014b). In instances where there was more than one statistically significant, non-independent effect in a domain, we selected the median effect size (see SOM for the full set of effects used in our analyses). Finally, we estimated average effect size using a pair of alternative selection models, as outlined above and in McShane et al. (2016). These models permit effect size heterogeneity and the chance publication of non-significant results. The first selection model assumes non-significant results have 20% the chance of publication as significant results, whereas the second model assumes this relative probability to be 40%. Complete R code for the analyses described above is available at https://osf.io/4r3a5/.

3. Results

3.1. Primary analyses: pair-bond status

3.1.1. Multilevel meta-analysis

The first set of analyses tested the hypothesis that men down-regulate T within the context of a romantic relationship, and thus pair-bonded men should have lower T concentrations than single men. The overall analysis included 38 published effects from 25 studies (N = 9536 data points), and 22 unpublished effects (N = 1502). A multilevel meta-analysis including study as a random factor (k = 60) yielded a mean effect size estimate of r = 0.149, 95% CI: .115:.183. See Fig. 1. On average, single men exhibited higher T concentrations than men in committed relationships.

3.1.2. Moderators

One common argument is that, due to publication bias, unpublished effects are likely to be weaker than published ones. Unpublished effects in the pair-bond status domain were estimated to be smaller (unpublished r = .097; published r = .177), and this difference was significant (p = .039). As testosterone levels decline with age, and men in relationships may be systematically older than single men, we also tested whether age-controlled effects differed from effects without an age control. We found no evidence that these two categories of effects differed (age controlled: r = .156; age not controlled: r = .144; p = .759). The difference in effect sizes between Western and non-Western samples fell just short of significance (Western r = .130, non-Western r = .207; p = .056).

Finally, we tested whether differing operationalizations of 'relationship status' led to different estimates of its effect on T. When comparing across categories of how studies classified whether men were 'paired' (married only, paired but unmarried, mix of married and unmarried, undefined), effect sizes ranged from r = .091 -.164; the overall difference between categories failed to reach significance (p = .787); pairwise comparisons similarly revealed that none of the contrasts between individual categories approached significance.

3.1.3. p-curve and p-uniform

Only independent, statistically significant published effects were used in the *p*-curve and *p*-uniform analyses (k = 18), per the recommendations in Simonsohn et al. (2014a, 2014b). The *p*-curve for these effects demonstrated significant right skew, consistent with results containing evidential value (i.e., not being due entirely to *p*-hacking and publication bias—full p-curve: Z = -2.61, p = 0.005; half *p*-curve: Z = -1.99, p = 0.023). *P*-curve's estimate of the true mean effect was r = 0.083. The average effect size estimate from *p*-uniform was, as expected, nearly identical: r = 0.078. The confidence interval for these estimates did not reject the null hypothesis of no effect (95% CI: -.042:.133).

3.1.4. Alternative selection models

Finally, we present estimates derived from alternative selection models. In the overall set of published effects, when heterogeneity was permitted to be non-zero, and non-significant results had 20% the chance of being published relative to significant results, the estimated effect size was r = .124. Increasing this probability to 40% raised the estimate to r = .154.

3.2. Secondary analyses: pair-bond behaviors

3.2.1. Multilevel analysis

This analysis included 12 effects from 8 studies (N = 1388 data points), and 4 unpublished effects (N = 186). An analysis of this body of effects (k = 16) yielded an average effect size estimate of r = .215, 95% CI: .138:.291. Unpublished and published effects were approximately equal in size (effect of publication status: r = .-005, p = .962). Overall, men with higher T concentrations exhibited more behaviors indicative of an interest in finding and acquiring new mates (i.e., "mating effort").

3.2.2. Other moderators

Pair-bond behaviors measured varied across studies. We coded whether the effect we extracted from a given study pertained to a) paired men's attitudes or behaviors in their current relationship (k = 10), b) single men's sexual behavior or attitudes (k = 3); or c) a mix of the two (k = 3). Effect sizes ranged from r = .118 -.315. The largest effect sizes pertained to current relationship attitudes/behaviors (r = .315), whereas single men's sexual attitudes/behaviors was smaller (r = .195); however, this difference was non-significant (p = .131). Effects from non-Western samples were larger than those from Western samples (r = .363 and r = .203, respectively), but with only two non-Western effects, this difference was non-significant (p = .317).

3.2.3. p-curve and p-uniform

The estimated effect size with *p*-curve (k = 5) was r = .123. Estimates from *p*-uniform revealed that the small number of significant published effects led to an imprecise estimate with a confidence interval

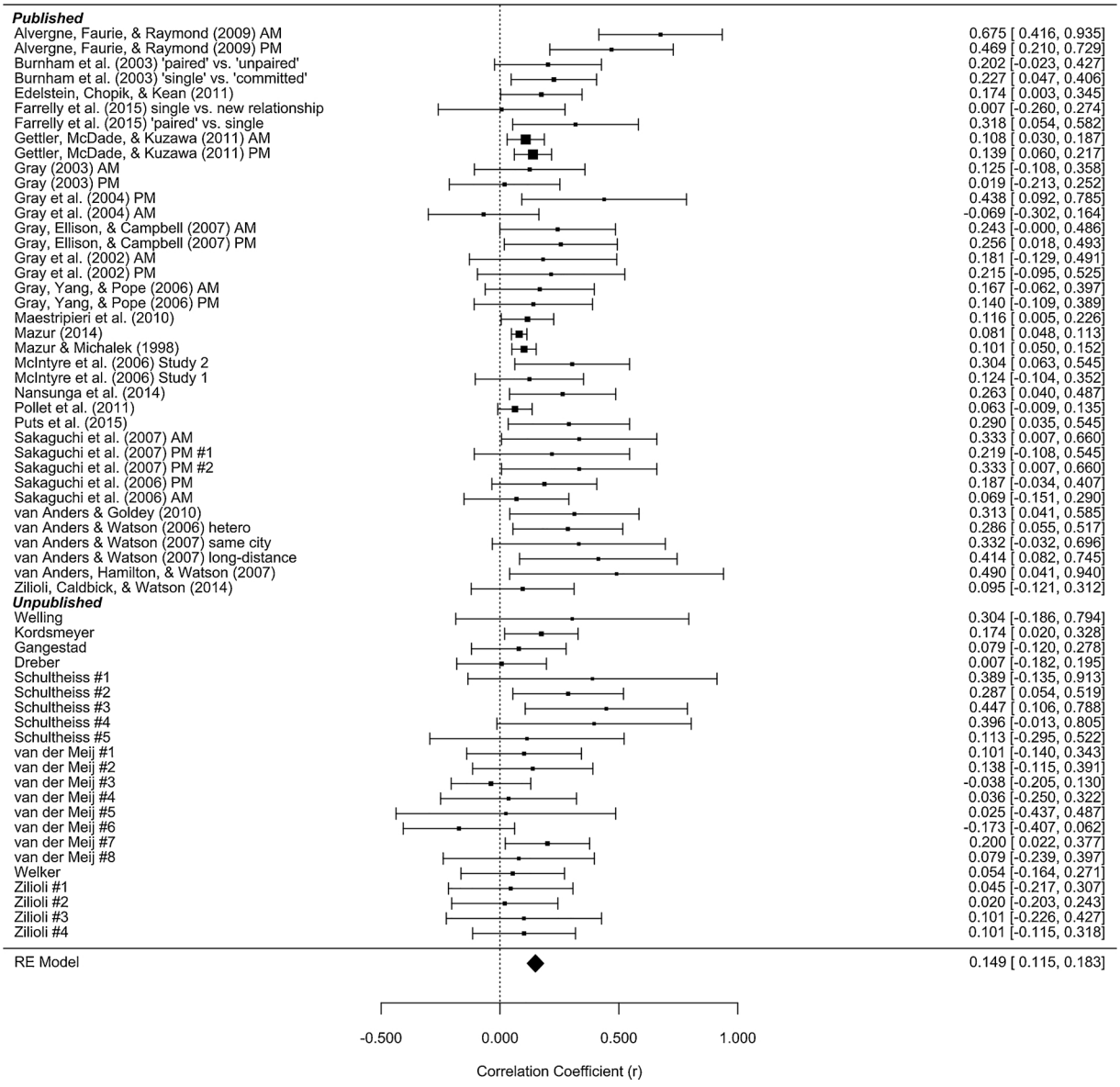

Fig. 1. Forest plot of pair-bond status effects. Brackets represent 95% confidence intervals for individual effects. Width of diamond represents the 95% confidence interval for the overall effect size estimate.

overlapping zero: r = .131, 95% CI: -.146:.406. Unlike the traditional estimates in this domain, these estimates do not find evidence for a robust effect. However, the *p*-curves indicated evidential value: full *p*-curve: Z = -2.14, p = 0.016; half *p*-curve: Z = -1.46, p = 0.072.

3.2.4. Alternative selection models

A heterogeneous selection model, in which non-significant results have 20% the chance of being published relative to significant results, estimated the effect size to be r = .141. Increasing this probability to 40% raised the estimate to r = .169.

3.3. Primary analyses: fatherhood status

3.3.1. Multilevel meta-analysis

The first set of analyses in the fatherhood domain tested whether fathers, who are hypothesized to upregulate parenting effort and downregulate mating effort, have lower T than non-fathers. The overall analysis included 22 published effects from 16 studies (N = 5223 data points), and 6 unpublished effects (N = 1091). A multilevel meta-analysis of effects nested within studies (k = 28) yielded a mean effect size estimate of r = 0.189, 95% CI: .111:.267, suggesting that fathers overall have lower concentrations of T than non-fathers. See Fig. 2.

3.3.2. Moderators

In this domain, unpublished effects were smaller than published effects (published r = .233; unpublished r = .077), consistent with a set of studies affected by publication bias. This difference fell short of significance (p = .067).

As with relationship status, age might introduce a confound into comparisons based on fatherhood status, as fathers tend to be older than non-fathers. However, we found no evidence that age-controlled effects differed from non age-controlled effects (p = .458). Lastly, we examined how the presence of a “pair-bond confound” moderated effect size estimates (i.e., comparing fathers to non-fathers may also be comparing paired to single men; see Methods). Estimates varied appreciably across categories, ranging from r = .155 (when there was insufficient information regarding pair-bond status) to r = .248 (when there was no confound with pair-bonding), to r = .313 (when fatherhood was fully confounded with pair-bonding). The existence of strongest effects in the ‘full confound’ category—in which paired

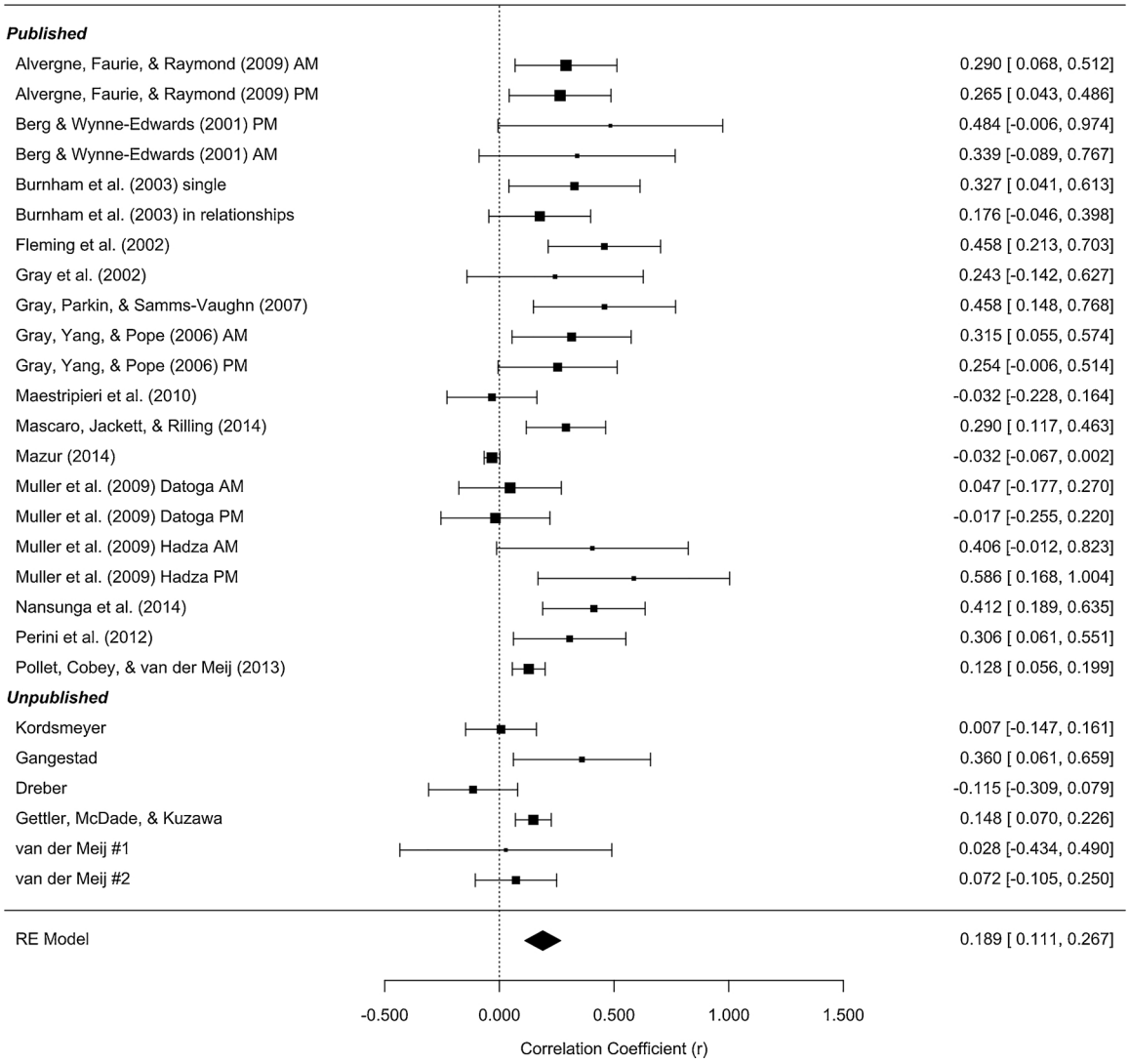

Fig. 2. Forest plot of fatherhood status effects. Brackets represent 95% confidence intervals for individual effects. Width of diamond represents the 95% confidence interval for the overall effect size estimate. Effect for Storey et al. (2000) not depicted.

fathers were compared with unpaired non-fathers—is consistent with expectations that bias inflates the estimates of these effects. However, all possible pairwise comparisons of confound categories failed to reach statistical significance (all $p > .05$).

3.3.3. *p-curve and p-uniform*

The estimated effect size with *p*-curve using only significant, independent, published effects in the predicted direction ($k = 12$) is $r = .186$. The right skew for the half *p*-curve was highly significant, $Z = -3.30$, $p < 0.001$, indicating 'evidential value' of the body of studies in this domain. The estimate from *p*-uniform was very similar and indicated the presence of a robust effect: $r = .185$, 95% CI: .037:.339.

3.3.4. *Alternative selection models*

A heterogeneous selection model, in which non-significant results had 20% the chance of being published relative to significant results, estimated the effect size to be $r = .145$. Increasing this probability to 40% raised the estimate to $r = .190$.

3.4. *Secondary analyses: fathering behaviors*

3.4.1. *Multilevel analysis*

This analysis included 11 published effects from six studies ($N = 504$ data points). A multilevel analysis yielded an average effect size estimate of $r = .334$, 95% CI: .244:.424.

3.4.2. *p-curve and p-uniform*

The estimated effect size with *p*-Curve ($k = 5$) was $r = .173$. Estimates from *p*-uniform differed noticeably, once again likely due to the small number of significant published effects: $r = .266$, 95% CI: -.364:..573. The *p*-curves did not indicate evidential value: full *p*-curve: $Z = -0.87$, $p = 0.191$; half *p*-curve: $Z = 0.13$, $p = 0.551$.

3.4.3. *Heterogeneous selection models*

A heterogeneous selection model, in which non-significant results had 20% the chance of being published relative to significant results, estimated the effect size to be $r = .258$. Increasing this probability to 40% raised the estimate to $r = .308$.

3.5. *Summary of estimates*

For a summary of effect size estimates by domain, see Fig. 3.

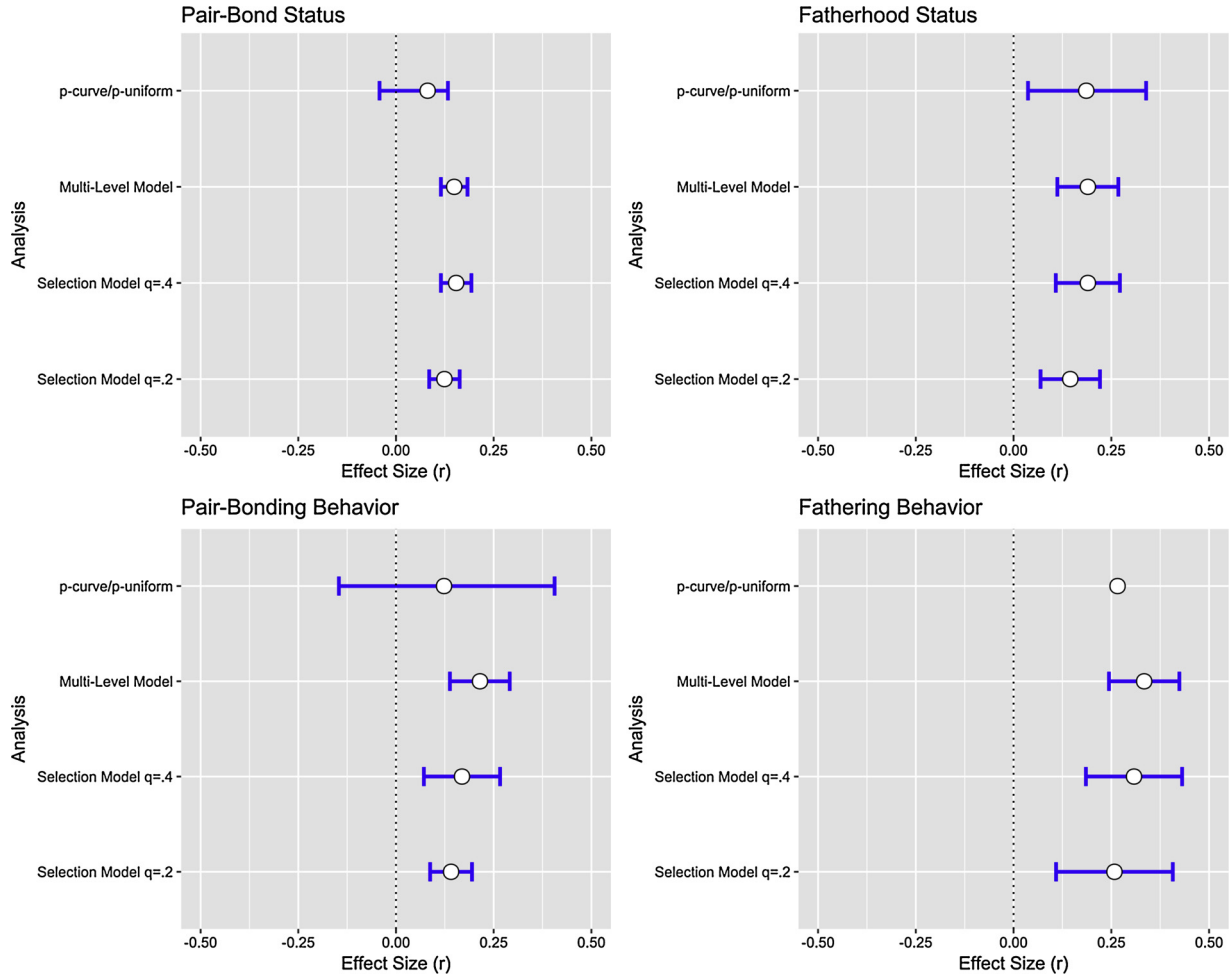

Fig. 3. Dot plot summarizing the estimates of effect size estimates by domain. Brackets represent 95% confidence intervals for each analysis type. Confidence interval omitted for Fathering Behavior *p*-curve/*p*-uniform estimate (-.364:.573).

4. Discussion

4.1. Summary of findings

In our meta-analysis, we evaluated evidence for four different predictions derived from the Challenge Hypothesis: 1) pair-bonded men will have lower T levels than single men; 2) men more committed and/or invested in their current pair-bonded relationship will have lower T levels than those less involved; 3) fathers will have lower T levels than non-fathers; and 4) fathers more involved in parenting activities will have lower T levels than fathers who are less involved. To establish a plausible range of effect sizes, we used a variety of meta-analytic techniques (traditional estimates, *p*-curve/*p*-uniform, and alternative selection models) and found that each of our predictions was supported, albeit to varying degrees. In our primary analyses, the aggregate of evidence suggests that the effect of pair-bond status and fatherhood status are both robust and non-zero. The effect of pair-bond status (r = .08–.15) was smaller than the effect of fatherhood (r = .15–.19). Within the pair-bond status domain, published effects were appreciably larger than unpublished effects, and non-Western effects were marginally larger than Western effects. Within the fatherhood domain, published effects, and those confounded with pair-bond status were appreciably larger than unpublished and non-confounded effects, respectively; however, both differences fell short of statistical significance.

We also conducted secondary analyses on sets of effects involving correlations of T with specific pair-bond or fathering behaviors thought to indicate men's balance between mating and parenting effort. Here, we found larger effects than in our primary analyses: r = .12 - .22 for pair-bond behaviors and r = .13 - .33 for fathering behaviors. However, given the smaller number of effects used in these analyses, the uncertainty in these estimates was also higher. Once again, the aggregate of evidence suggested that these effects were significantly different from zero, though analyses were less unanimous on this point for fathering behaviors.

4.2. Contextualization of effect sizes

In interpreting effect sizes, researchers often turn to Cohen, (1988; 1992) rules of thumb: r coefficients of .1, .3, and .5 are translated as small, medium, and large, respectively. By these conventions, the effect sizes we estimate generally fall between small and medium. Yet, this begs a fundamental question (as estimations of effect size often do): How should one interpret the practical significance of a "small" or "medium" effect? To address this point, it may be helpful to turn to factors—investigated in other literatures—that are known to affect T concentrations. As Cohen himself noted, "a basis for positing [effect size] which comes from theory or experience should automatically take precedence" over arbitrary benchmarks (Cohen, 1988; p. 147). With this in mind, we contextualize our effect size estimates by comparing them to known differences in T as a function of aging, ethnic or geographic population, and exogenous administration.

Following a peak shortly after puberty, men's T declines throughout the life course; however, the steepness of the decline, and whether the decline stops at a certain age, remain a matter of debate (see Kelsey

et al., (2014) and Handelsman et al., (2015) for two treatments of this issue). Furthermore, as the results of our meta-analysis show, circumstantial and/or social factors have an appreciable influence on men's T. Thus, estimates of the size of men's age-related declines differ substantially. Kelsey et al. (2014) aggregated data from 13 samples of men (total N = 10,098) to generate a predictive model for men's T across the lifespan. We used the data from Kelsey et al. to determine median T, along with standard deviations, at various adult ages. As illustrative examples, this model predicts an $r = .17$ decrease when comparing men age 20 and age 30, but only an $r = .06$ decrease between age 30 to 40. Comparing our meta-analytic estimates to the normative data from Kelsey et al., the average T difference between a single and a paired man that we report ($r \approx .10$) approximates the average T difference between a 20 year old man and a 25 year old man. The T difference between fathers and non-fathers ($r \approx .18$) is approximately equal to the difference between a 20 and 32 year old man.

As T secretion appears to increase along a gradient of socioeconomic status, urbanization, and affluence (Alvarado, 2010; Gray et al., 2006), between-population differences in men's T provide another means of contextualizing our estimates. Comparisons of populations of men both within and between societies tend to show moderate differences that get smaller with increasing ages. Taking the within-society comparison of black and white American men as an example, one study with a mean age of 31 found a difference of $r = .12$ (black men had higher average T; Ettinger et al., 1997), but a second study with a mean age of 38 only reported a difference of $r = .03$ (Ellis and Nyborg, 1992). Our pair-bond status estimate falls near the higher end of this ethnic difference, but our other estimates all exceed the magnitude of this difference. Moving to between-society comparisons, Ellison and Gray, (2009) showed that some geographic populations of men vary considerably in their average T concentrations. Here, our pair-bond status effect closely resembles the average difference between American and Nepalese men (higher in American men; $r = .10$). Our fatherhood status effect is nearly equal to the difference between Congolese and Nepalese men (higher in Congolese men; $r = .19$).

Finally, studies of exogenous T administration provide yet another benchmark. Such effects generally greatly exceed the estimates we provide that are based on social and/or life history factors. For instance, 25 mg of a weekly T injection, considered to be a very small dose, has an effect on endogenous T equal to $r = .39$ (Bhasin et al., 2001). Larger doses of exogenous T often show effects even larger (e.g., $r = .91$ for the effect of increasing a weekly injection by 100 mg; see Bhasin et al., 2001).

In sum, most natural between-population comparisons of men's T—along with our estimates that closely resemble these population differences—would be classified conventionally as "small" to "medium" effect sizes. Yet, given the extensive interest in the scholarly literature on the consequences of T differences of this size (see, e.g., Kelsey et al., 2014; Alvarado, 2010), we believe our effect size estimates suggest an appreciable *practical* (not just statistical) significance.

4.3. Strengths and limitations

In this meta-analysis, we have provided the first comprehensive quantitative review to date of several predictions derived from the Challenge Hypothesis in humans. Previous reviews have either concerned separate predictions of the Challenge Hypothesis (e.g., the link between T and aggression; Archer, 2006), or been narrative in nature and have relied on published effects only (e.g., Gray and Campbell, 2009). Drawing from the development of new theory and techniques from meta-analytic science, we have provided a plausible range of effect size estimates for predictions that vary in the methodologies used to test them, and in the power available within the literature to detect true effects. For our two main predictions, concerning pair-bond and fatherhood status, we aggregated a large enough set of effects to estimate a tight range of effect sizes, providing strong evidence of real differences in these domains. For our two secondary analyses, although the overall pattern of results was consistent with our predictions, the meta-analytic estimates for the pair-bonding behavior and fathering behavior domain did not contain sufficient statistical power to draw equally firm conclusions regarding the robustness of these factors in predicting men's T.

We sent out a widespread call for unpublished effects and obtained dozens of these effects for our analyses, which we see as a major strength of our meta-analysis. However, several caveats apply to this point. First, the majority of the unpublished effects (22 out of 30) were in the pair-bond status domain, with only a handful of unpublished effects for fatherhood status or pair-bonding behaviors (and none for fathering behaviors). Thus, the benefit of reduced bias may be largely concentrated in the pair-bond status estimates. Second, we have no doubt that we were unable to recover every unpublished effect. At the very least, we know of several datasets containing the appropriate variables that we were never able to access through authors. Third, while analyses comparing published to unpublished effects showed no statistically significant differences, the overall pattern—with unpublished effects being smaller—goes in the direction one would expect with publication bias. Perhaps a sample with greater power—i.e., more unpublished effects—would reveal significant differences.

We have attempted to address these limitations transparently. First, we fully acknowledge the challenge of generating estimates in a heterogeneous and biased literature. We believe publication bias has affected the literature on human behavioral endocrinology and the Challenge Hypothesis, as it has virtually all other fields in psychology (Simonsohn, 2012). For this reason, we made a concerted decision not to rely on any single procedure in generating our estimates. Meta-analytic techniques may treat the non-independence of effect optimally (by using multilevel models), or may treat publication bias optimally (by modeling selection processes). To our knowledge, there is no technique that does both. Traditional meta-analytic estimates are able to incorporate multiple non-independent effects in a multilevel analysis, and possess greater precision due to incorporating a larger number of effects; however, the estimates may still suffer from some degree of upward bias. A number of techniques have been proposed to account for the bias introduced by heterogeneity, publication bias, and *p*-hacking, and we used several of them to help fill out a plausible range of effect size estimates. Discussion is ongoing regarding the strengths and weaknesses of these techniques (e.g., McShane et al., 2016; Carter et al., 2017; Nelson, 2018), particularly those that are newly developed, but through our reading of these discussions, we attempted to select techniques supported by simulations and quantitative demonstrations. More generally, given widespread disagreement regarding how to best correct for bias in meta-analyses (Carter and McCullough, 2018), we feel that an approach focused on sensitivity analyses that relies on multiple techniques is superior to selecting a particular technique and cherry-picking evidence in its favor.

Importantly, the preponderance of evidence from our results suggests that bias is *not* the sole source of positive effects in any of the domains we examine. In our mind, this point is also worth emphasizing: based on our analyses, we believe effects based on the Challenge Hypothesis in men are both real *and* affected by selective reporting.

4.4. Suggestions for future research

Our meta-analysis has provided what we hope will serve as a useful reference for the average effect of social relationships on men's T in several domains. However, numerous avenues still exist to more precisely tease out the nature of relationships between life-history shifts and men's T (see e.g., Zilioli and Bird, 2017). Thanks to the work of biological anthropologists, our analyses contained a non-trivial number of effects from non-Western populations (e.g., Alvergne et al., 2009; Gettler et al., 2011; Muller et al., 2009; Nansunga et al., 2014); however, much more work remains to be done to determine the impacts of

cross-cultural socioecological variation on hormone-mediated life-history shifts.

Second, future research in humans may benefit from an increased effort to tease apart the proximate, biological connections between T and parenting effort. Our analysis, despite finding an overall negative association between T and the expression of parenting behaviors in men, is unable to speak to specific mechanisms at play that mediate the theorized trade-off. Non-human animal research, in its effort to explain complex and heterogeneous findings regarding the link between paternal behavior and T (Bales and Saltzman, 2016; Hirschenhauser et al., 2003; Wynne-Edwards and Timonin, 2007), has turned to identifying specific neural mechanisms that regulate paternal behavior. For example, studies from rodents suggest that the neuromodulatory effects of T on reward circuits, which include brain regions such as the medial preoptic area of the hypothalamus (MPOA) and bed nucleus of the stria terminalis (BST), are central to the expression of paternal behavior (reviewed in Bales and Saltzman, 2016). Similar research in humans is just beginning. A preliminary imaging study of ten human fathers identified numerous prefrontal and subcortical brain regions that associate with T responses to infants (Kuo et al., 2012). Mascaro et al., (2014), in a study of 88 fathers, found that neural responses in a brain region associated with face emotion processing (the caudal middle frontal gyrus [MFG]) were correlated with concurrent T concentrations. Another study of 70 fathers found that ventral tegmental area (VTA) activity, while associated with viewing pictures of their infants, was not correlated with T (Mascaro et al., 2013). Obviously, much more remains to be done to elucidate the role of T in the paternal brain (Swain et al., 2014; Feldman, 2015), and we strongly encourage future research in this vein.

Relatedly, the role of other neuroendocrine mechanisms in men's life-histories also remains ripe for exploration and review. Oxytocin (OT), in particular, is another hormone intimately involved in the processes of sexual pair-bonding and parenting (Gangestad and Grebe, 2017). Although recent meta-analyses have failed to establish robust effects of OT on more general prosocial behavior (e.g., trust; Nave et al., 2015), researchers have found robust support for OT's role in social processes specific to romantic relationships and parent-child bonds (Feldman, 2017; Gangestad and Grebe, 2017). Theoretical frameworks have been developed to jointly account for the effects of OT and T in social relationships (e.g. the 'steroid/peptide theory of social bonds' from van Anders et al., (2011); the 'neuroscience of social decision-making' from Rilling and Sanfey, (2011). These frameworks expand upon the CH and are necessary to advance our understanding of the behavioral endocrinology of humans' social relationships. An emblematic example in this regard concerns paternal protection of infants, which is conventionally considered an example of parenting effort, yet it has been associated with increased T (van Anders et al., 2011). It may be that interactions with concomitant OT surges contribute to positive correlations between T and certain aspects of parenting effort. Similar interactive effects may also drive associations between hormones and men's mating effort (Mascaro et al., 2014).

Prolactin, too, has been associated with paternal care across diverse animal taxa (Brown et al., 1995; Gubernick and Nelson, 1989; Lynn, 2016; Reburn and Wynne-Edwards, 1999; Saltzman and Ziegler;, 2014; Schradin and Anzenberger, 1999). Pioneering correlational and experimental work in birds has established that prolactin directly contributes to fathering behaviors such as incubation and offspring provisioning (reviewed in Lynn, 2016). Though correlational studies from monogamous rodents and New World monkeys are consistent with findings in birds, experimental work interestingly has failed to find evidence for a direct activational effect of prolactin on paternal behavior (see Saltzman and Ziegler, 2014). Finally, early evidence supports a potential role for prolactin in human fatherhood as well, though stronger conclusions await more definitive evidence (see Gangestad and Grebe, 2017). We agree with van Anders et al. (2011) that an increased emphasis on studies of prolactin in humans may reveal important roles in conjunction with T and OT.

For future research in the realms we mention above, we provide two suggestions that may aid in yielding more robust effects.

First, our analyses focus almost exclusively on cross-sectional, between-subjects comparisons of T. This was a purposeful choice, motivated by a desire to obtain effects broadly represented across the literature. At the same time, we acknowledge that a reliance on between-subjects comparisons as a test of the CH entails paying a cost in construct validity; indeed, theory specifically pertains to *within-individual* changes in T in response to life circumstances. Although we attempted to control for a number of confounds that could introduce error into between-group comparisons, extraneous factors may still have affected the observed 'baseline' concentrations of T. Thus, a more desirable method for conducting future research entails the measurement of within-person hormonal changes in response to evolutionarily-salient life events. The few longitudinal studies in humans bolster the conclusion of a T-mediated shift toward parenting effort when men become fathers (e.g., Storey et al., 2000; Gettler et al., 2011). Future work could complement these findings by shedding light on the role played by individual differences in the degree to which fathers shift from mating to parenting effort. Short-term changes in OT, too, have been shown to predict dynamics of romantic relationships (Grebe et al., 2017) and parental care (e.g., Feldman et al., 2010) in men. A promising avenue for future longitudinal research would be to clarify how these changes function as a component of men's life-history strategies.

Second, the published studies we review are generally greatly underpowered to detect the average effect sizes that we estimate. For instance, the median sample size in our pair-bond status domain is 73; this provides just a 17% chance of detecting a true effect of $r = .12$. Substituting our various effect size estimates generates a range of power from 10%–26%. The power to detect our estimated fatherhood status effect ($r = .18$) with the median sample size in our dataset ($n = 71$) is higher but still far from ideal, at 32% (ranging from 23% to 38%, again depending on effect size estimate used).[3] Statistical power at these levels presents problems of both an inflated false negative rate, and an inflated estimate of effect size ("the winner's curse"; see Ioannidis, 2008). We recommend that researchers look to power analysis, rather than previously published papers, when designing future studies on T and the Challenge Hypothesis. As an example, the sample size necessary to achieve 80% power (a figure often used as a target when designing new studies) to detect effects in the range of our fatherhood status estimate is anywhere from 200 to 360 participants, greater than the vast majority of studies in our dataset. We note that this recommendation also applies to research in behavioral endocrinology more generally, a field that often suffers from a reliance on small, underpowered samples to address hypotheses of interest (see e.g. Walum et al., 2016).

4.5. Conclusion

Across male members of many species, T has been shown to mediate trade-offs between mating and parenting effort. The same mediating role has been argued to exist in humans, manifesting in T levels of single men being generally higher than T levels of pair-bonded men and fathers, relationship investment and fathering behavior negatively associated with circulating T, and mate-seeking behaviors positively associated with circulating T. Although these relationships have been reported in narrative reviews, to date no meta-analytic evidence exists that speaks to the precise effect size, statistical significance, and practical relevance of these relationships. In the current meta-analysis, we aggregated findings from 114 effects from 66 published and unpublished studies (for a total of 19,397 data points from 17,341 individuals) and found that being single was associated with greater T levels, non-fathers had higher T levels than fathers, mate-seeking

[3] All estimates obtained from G*Power (version 3.1.9.2).

behavior was associated with higher T levels, and fathering behavior was associated with lower T levels. Our effect sizes, which would be interpreted as somewhere between "small" and "medium" based on convention, can be fruitfully interpreted in light of the effects of other factors known to relate to T, such as aging, population differences, and T administration. With these overall associations in mind, we encourage future research, based on large samples and within-subjects comparisons whenever possible, that links specific neuroendocrine mechanisms and behaviors to the dynamic nature of men's life-histories.

References

Alvarado, L.C., 2010. Population differences in the testosterone levels of young men are associated with prostate cancer disparities in older men. Am. J. Hum. Biol. 22 (4), 449–455.

Alvergne, A., Faurie, C., Raymond, M., 2009. Variation in testosterone levels and male reproductive effort: insight from a polygynous human population. Horm. Behav. 56 (5), 491–497.

Archer, J., 2006. Testosterone and human aggression: an evaluation of the challenge hypothesis. Neurosci. Biobehav. Rev. 30 (3), 319–345.

Bales, K.L., Saltzman, W., 2016. Fathering in rodents: neurobiological substrates and consequences for offspring. Horm. Behav. 77, 249–259.

Berg, S.J., Wynne-Edwards, K.E., 2001. Changes in testosterone, cortisol, and estradiol levels in men becoming fathers. June. Mayo Clin. Proc. 76 (6), 582–592 Elsevier.

Bhasin, S., Woodhouse, L., Casaburi, R., Singh, A.B., Bhasin, D., Berman, N., Dzekov, J., 2001. Testosterone dose-response relationships in healthy young men. Am. J. Physiol. Endocrinol. Metabol. 281 (6), E1172–E1181.

Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R., 2011. Introduction to Meta-analysis. John Wiley & Sons.

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 (4), 474–491.

Burnham, T.C., Chapman, J.F., Gray, P.B., McIntyre, M.H., Lipson, S.F., Ellison, P.T., 2003. Men in committed, romantic relationships have lower testosterone. Horm. Behav. 44 (2), 119–122.

Cardwell, J.R., Liley, N.R., 1991. Hormonal control of sex and colour change in stoplight parrotfish, Sparisoma viride. Gen. Comp. Endocrin 81, 7–20.

Cardwell, J.R., Sorensen, P.W., Van Der Kraak, G.J., Liley, N.R., 1996. Effect of dominance status on sex hormone levels in laboratory and wild-spawning male trout. Gen. Comp. Endocrinol. 101 (3), 333–341.

Carter, E.C., McCullough, M.E., 2018. A simple, principled approach to combining evidence from meta-analysis and high-quality replications. Adv. Methods Pract. Psychol. Sci.

Carter, E., Schönbrodt, F., Gervais, W.M., Hilgard, J., 2017. Correcting for Bias in Psychology: a Comparison of Meta-analytic Methods. Preprint online at:. https://osf.io/preprints/psyarxiv/9h3nu.

Cavigelli, S.A., Pereira, M.E., 2000. Mating season aggression and fecal testosterone levels in male ring-tailed lemurs (*Lemur catta*). Horm. Behav. 37 (3), 246–255.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd edition. .

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155.

Cristóbal-Azkarate, J., Chavira, R., Boeck, L., Rodríguez-Luna, E., Veàl, J.J., 2006. Testosterone levels of free-ranging resident mantled howler monkey males in relation to the number and density of solitary males: a test of the challenge hypothesis. Horm. Behav. 49 (2), 261–267.

Del Giudice, M., Gangestad, S.W., Kaplan, H.S., 2015. Life history theory and evolutionary psychology. In: Buss, D.M. (Ed.), The Handbook of Evolutionary Psychology. John Wiley & Sons, Hoboken, NJ.

Dixson, A.F., Lunn, S.F., 1987. Post-partum changes in hormones and sexual behaviour in captive groups of marmosets (*Callithrix jacchus*). Physiol. Behav. 41 (6), 577–583.

Duval, S., Tweedie, R., 2000. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56 (2), 455–463.

Ellis, L., Nyborg, H., 1992. Racial/ethnic variations in male testosterone levels: a probable contributor to group differences in health. Steroids 57 (2), 72–75.

Ellison, P.T., Gray, P.B. (Eds.), 2009. Endocrinology of Social Relationships. Harvard University Press.

Ettinger, B., Sidney, S., Cummings, S.R., Libanati, C., Bikle, D.D., Tekawa, I.S., et al., 1997. Racial differences in bone density between young adult black and white subjects persist after adjustment for anthropometric, lifestyle, and biochemical differences. J. Clin. Endocrinol. Metab. 82 (2), 429–434.

Feldman, R., 2015. The adaptive human parental brain: implications for children's social development. Trends Neurosci. 38 (6), 387–399.

Feldman, R., 2017. The neurobiology of human attachments. Trends Cogn. Sci. (Regul. Ed.) 21 (2), 80–99.

Feldman, R., Gordon, I., Schneiderman, I., Weisman, O., Zagoory-Sharon, O., 2010. Natural variations in maternal and paternal care are associated with systematic changes in oxytocin following parent–infant contact. Psychoneuroendocrinology 35 (8), 1133–1141.

Fernandez-Duque, E., Valeggia, C.R., Mendoza, S.P., 2009. The biology of paternal care in human and nonhuman primates. Annu. Rev. Anthropol. 38, 115–130.

Fine, C., 2017. Testosterone Rex: Myths of Sex, Science, and Society. WW Norton & Company, New York, NY.

Francis, R.C., Fernald, R.D., 1993. Neuroendocrine effects of dominance status in an African cichlid. *XXIII International Ethological Conference* (Torremolinos, Spain) Abstracts. pp. 340.

Gangestad, S.W., Grebe, N.M., 2017. Hormonal systems, human social bonding, and affiliation. Horm. Behav. 91, 122–135.

Gervais, W., 2015. In: Lexington, K.Y., Gervais, William (Eds.), Putting PET-PEESE To The Test [Internet], June 25 [cited 2016 March 30]. Available from: http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1.

Gettler, L.T., McDade, T.W., Feranil, A.B., Kuzawa, C.W., 2011. Longitudinal evidence that fatherhood decreases testosterone in human males. Proc. Natl. Acad. Sci. 108 (39), 16194–16199.

Gray, P.B., Campbell, B.C., 2009. Human male testosterone, pair bonding and fatherhood. In: Ellison, P.T., Gray, P.B. (Eds.), Endocrinology of Social Relationships. Harvard University Press, Cambridge, MA.

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 (3), 193–201.

Gray, P.B., Yang, C.F.J., Pope, H.G., 2006. Fathers have lower salivary testosterone levels than unmarried men and married non-fathers in Beijing, China. Proc. R. Soc. Lond. B: Biol. Sci. 273 (1584), 333–339.

Grebe, N.M., Kristoffersen, A.A., Grøntvedt, T.V., Thompson, M.E., Kennair, L.E.O., Gangestad, S.W., 2017. Oxytocin and vulnerable romantic relationships. Horm. Behav. 90, 64–74.

Greenberg, N., Crews, D., 1990. Endocrine and behavioral responses to aggression and social dominance in the green anole lizard, Anolis carolinensis. Gen. Comp. Endocrinol. 77 (2), 246–255.

Gubernick, D.J., Nelson, R.J., 1989. Prolactin and paternal behavior in the biparental California mouse, Peromyscus californicus. Horm. Behav. 23 (2), 203–210.

Handelsman, D.J., Yeap, B.B., Flicker, L., Martin, S., Wittert, G.A., Ly, L.P., 2015. Age-specific population centiles for androgen status in men. Eur. J. Endocrinol. 173 (6), 809–817.

Harris, C.R., Pashler, H., Mickes, L., 2014. Elastic analysis procedures: an incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). Psychol. Bull. 140 (5), 1260.

Hedges, L.V., Vevea, J.L., 1996. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. J. Educ. Behav. Stat. 21 (4), 299–332.

Hedges, L., Vevea, J., 2005. Selection method approaches. In: Rothstein, H., Sutton, A., Borenstein, M. (Eds.), Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments. John Wiley & Sons, Chichester, UK, pp. 145–174.

Hegner, R.E., Wingfield, J.C., 1987. Effects of experimental manipulation of testosterone levels on parental investment and breeding success in male house sparrows. Auk 462–469.

Hendrick, S.S., 1988. A generic measure of relationship satisfaction. J. Marriage Fam. 93–98.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behav. Brain Sci. 33 (2-3), 61–83.

Herting, B.L., Belthoff, J.R., 1997. Testosterone, Aggression, and Territoriality in Male Western Screech-owls (Otus Kennicottue): Results From Preliminary Experiments. United States Department of Agriculture Forest Service General Technical Report NC, pp. 213–217.

Hirschenhauser, K., Oliveira, R.F., 2006. Social modulation of androgens in male vertebrates: meta-analyses of the challenge hypothesis. Anim. Behav. 71 (2), 265–277.

Hirschenhauser, K., Winkler, H., Oliveira, R.F., 2003. Comparative analysis of male androgen responsiveness to social environment in birds: the effects of mating system and paternal incubation. Horm. Behav. 43 (4), 508–519.

Hooper, A.E.C., Gangestad, S.W., Thompson, M.E., Bryan, A.D., 2011. Testosterone and romance: the association of testosterone with relationship commitment and satisfaction in heterosexual men and women. Am. J. Hum. Biol. 23 (4), 553–555.

Hung, H.J., O'Neill, R.T., Bauer, P., Kohne, K., 1997. The behavior of the p-value when the alternative hypothesis is true. Biometrics 11–22.

Hunt, K.E., Hahn, T.P., Wingfield, J.C., 1999. Endocrine influences on parental care during a short breeding season: testosterone and male parental care in Lapland longspurs (Calcarius lapponicus). Behav. Ecol. Sociobiol. (Print) 45 (5), 360–369.

Ioannidis, J.P., 2008. Why most discovered true associations are inflated. Epidemiology 19 (5), 640–648.

Kaplan, H.S., Lancaster, J.B., 2003. An evolutionary and ecological analysis of human fertility, mating patterns, and parental investment. Offspring: Human Fertility Behavior in Biodemographic Perspective. pp. 170–223.

Kelsey, T.W., Li, L.Q., Mitchell, R.T., Whelan, A., Anderson, R.A., Wallace, W.H.B., 2014. A validated age-related normative model for male total testosterone shows increasing variance but no decline after age 40 years. PLoS One 9 (10), e109346.

Ketterson, E.D., Nolan Jr, V., 1992. Hormones and life histories: an integrative approach. Am. Nat. 140, S33–S62.

Ketterson, E.D., Nolan Jr, V., 1999. Adaptation, exaptation, and constraint: a hormonal perspective. Am. Nat. 154 (S1) S4-S25.

Kindler, P.M., Philipp, D.P., Gross, M.R., Bahr, J.M., 1989. Serum 11-ketotestosterone and testosterone concentrations associated with reproduction in male bluegill (Lepomis macrochirus: centrarchidae). Gen. Comp. Endocrinol. 75, 446–453.

Kokko, H., Jennions, M.D., 2008. Parental investment, sexual selection and sex ratios. J. Evol. Biol. 21 (4), 919–948.

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 (2), 302–306.

Ledgerwood, A., 2016. Introduction to the special section on improving research practices: thinking deeply across the research cycle. Perspect. Psychol. Sci. 11 (5), 661–663.

Ligon, J.D., Thornhill, R., Zuk, M., Johnson, K., 1990. Male-male competition, ornamentation and the role of testosterone in sexual selection in red jungle fowl. Anim. Behav. 40 (2), 367–373.
Lynn, S.E., 2016. Endocrine and neuroendocrine regulation of fathering behavior in birds. Horm. Behav. 77, 237–248.
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., 201305579.
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.
Mayer, I., Rosenqvist, G., Borg, B., Ahnesjo, I., Berglund, A., Schulz, R., 1993. Plasma levels of sex steroids in three species of pipefish (*Syngnathidae*). Can. J. Zool. 71, 1903–1907.
Mazur, A., 2017. Testosterone in biosociology: a memoir. Horm. Behav. 92, 3–8.
McGlothlin, J.W., Jawor, J.M., Ketterson, E.D., 2007. Natural variation in a testosterone-mediated trade-off between mating effort and parental effort. Am. Nat. 170 (6), 864–875.
McIntyre, M., Gangestad, S.W., Gray, P.B., Chapman, J.F., Burnham, T.C., O'rourke, M.T., Thornhill, R., 2006. Romantic involvement often reduces men's testosterone levels–but not always: the moderating role of extrapair sexual interest. J. Pers. Soc. Psychol. 91 (4), 642.
McShane, B.B., Böckenholt, U., Hansen, K.T., 2016. Adjusting for publication bias in meta-analysis: an evaluation of selection methods and some cautionary notes. Perspect. Psychol. Sci. 11 (5), 730–749.
Muller, M.N., Wrangham, R.W., 2004. Dominance, aggression and testosterone in wild chimpanzees: a test of the 'challenge hypothesis'. Anim. Behav. 67 (1), 113–123.
Muller, M.N., Marlowe, F.W., Bugumba, R., Ellison, P.T., 2009. Testosterone and paternal care in East African foragers and pastoralists. Proc. R. Soc. Lond. B: Biol. Sci. 276 (1655), 347–354.
Nansunga, M., Manabe, Y.C., Alele, P.E., Kasolo, J., 2014. Association of testosterone levels with socio-demographic characteristics in a sample of Ugandan men. Afr. Health Sci. 14 (2), 348–355.
Nave, G., Camerer, C., McCullough, M., 2015. Does oxytocin increase trust in humans? A critical review of research. Perspect. Psychol. Sci. 10 (6), 772–789.
Nelson, L., 2018. 71] The (Surprising?) Shape of the File Drawer. [Internet]. Leif Nelson April 30 [cited 2018 May 1]. Available from: http://datacolada.org/71.
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 (1), 70–82.
Oliveira, R.F., Almada, V.C., Canario, A.V., 1996. Social modulation of sex steroid concentrations in the urine of male cichlid fish Oreochromis mossambicus. Horm. Behav. 30 (1), 2–12.
Oliveira, R.F., Hirschenhauser, K., Carneiro, L.A., Canario, A.V., 2002. Social modulation of androgen levels in male teleost fish. Comp. Biochem. Physiol. B: Biochem. Mol. Biol. 132 (1), 203–215.
Orchinik, M., Licht, P., Crews, D., 1988. Plasma steroid concentrations change in response to sexual behavior in *Bufo marinus*. Horm. Behav. 22 338e350. package version 0.0.2.
Parker, T.H., Knapp, R., Rosenfield, J.A., 2002. Social mediation of sexually selected ornamentation and steroid hormone levels in male junglefowl. Anim. Behav. 64 (2), 291–298.
Prior, N.H., Soma, K.K., 2015. Neuroendocrine regulation of long-term pair maintenance in the monogamous zebra finch. Horm. Behav. 76, 11–22.
Prudom, S.L., Broz, C.A., Schultz-Darken, N., Ferris, C.T., Snowdon, C., Ziegler, T.E., 2008. Exposure to infant scent lowers serum testosterone in father common marmosets (Callithrix jacchus). Biol. Lett. 4 (6), 603–605.
Puts, D.A., Pope, L.E., Hill, A.K., Cárdenas, R.A., Welling, L.L., Wheatley, J.R., Breedlove, S.M., 2015. Fulfilling desire: evidence for negative feedback between men's testosterone, sociosexual psychology, and sexual partner number. Horm. Behav. 70, 14–21.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods Vol. 1 Sage Publishing., Thousand Oaks, CA.
Reburn, C.J., Wynne-Edwards, K.E., 1999. Hormonal changes in males of a naturally biparental and a uniparental mammal. Horm. Behav. 35 (2), 163–176.
Rilling, J.K., Sanfey, A.G., 2011. The neuroscience of social decision-making. Annu. Rev. Psychol. 62, 23–48.
Roney, J.R., Gettler, L.T., 2015. The role of testosterone in human romantic relationships. Curr. Opin. Psychol. 1, 81–86.
Rose, R.M., Holaday, J.W., Bernstein, I.S., 1971. Plasma testosterone, dominance rank and aggressive behaviour in male rhesus monkeys. Nature 231, 366–368.
Rusbult, C.E., Martz, J.M., Agnew, C.R., 1998. The investment model scale: measuring commitment level, satisfaction level, quality of alternatives, and investment size. Pers. 5 (4), 357–387.
Saltzman, W., Ziegler, T.E., 2014. Functional significance of hormonal changes in mammalian fathers. J. Neuroendocrinol. 26 (10), 685–696.
Sapolsky, R.M., 2017. Behave: the Biology of Humans at Our Best and Worst. Penguin Books, London, UK.
Scaia, M.F., Morandini, L., Noguera, C., Trudeau, V.L., Somoza, G.M., Pandolfi, M., 2018. Can estrogens be considered as key elements of the challenge hypothesis? The case of intrasexual aggression in a cichlid fish. Physiol. Behav.
Schradin, C., Anzenberger, G., 1999. Prolactin, the hormone of paternity. Physiology 14 (6), 223–231.
Schuett, G.W., Harlow, H.J., Rose, J.D., Van Kirk, E.A., Murdoch, W.J., 1996. Levels of plasma corticosterone and testosterone in male copperheads (*Agkistrodon contortrix*) following staged fights. Horm. Behav. 30 (1), 60–68.
Setchell, J.M., Smith, T., Wickings, E.J., Knapp, L.A., 2008. Social correlates of testosterone and ornamentation in male mandrills. Horm. Behav. 54 (3), 365–372.
Sikkel, P.C., 1993. Changes in plasma androgen levels associated with changes in male reproductive behavior in a brood cycling marine fish. Gen. Comp. Endocrinol. 89, 229–237.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.
Simonsohn, U., 2012. It does not follow: evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). Perspect. Psychol. Sci. 7 (6), 597–599.
Simonsohn, U., 2017. 59] PET-PEESE Is Not Like Homeopathy [Internet]. Uri Simonsohn. April 12 [cited 2017 October 30]. Available from:. http://datacolada.org/59.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014a. P-curve and effect size: correcting for publication bias using only significant results. Perspect. Psychol. Sci. 9 (6), 666–681.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014b. P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. 143 (2), 534.
Simpson, J.A., Gangestad, S.W., 1991. Individual differences in sociosexuality: evidence for convergent and discriminant validity. J. Pers. Soc. Psychol. 60 (6), 870.
Smith, J.M., 1989. Evolutionary Genetics. Oxford University Press, Oxford, UK.
Sobolewski, M.E., Brown, J.L., Mitani, J.C., 2013. Female parity, male aggression, and the challenge hypothesis in wild chimpanzees. Primates 54 (1), 81–88.
Stanley, T.D., Doucouliagos, H., 2017. Neither fixed nor random: Weighted least squares meta-regression. Res. Synth. Methods 8 (1), 19–42.
Sterling, T.D., 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. J. Am. Stat. Assoc. 54 (285), 30–34.
Stoehr, A.M., Hill, G.E., 2000. Testosterone and the allocation of reproductive effort in male house finches (Carpodacus mexicanus). Behav. Ecol. Sociobiol. (Print) 48 (5), 407–411.
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 (2), 79–95.
Swain, J.E., Dayton, C.J., Kim, P., Tolman, R.M., Volling, B.L., 2014. Progress on the paternal brain: theory, animal models, human brain research, and mental health implications. Infant Ment. Health J. 35 (5), 394–408.
Terrin, N., Schmid, C.H., Lau, J., Olkin, I., 2003. Adjusting for publication bias in the presence of heterogeneity. Stat. Med. 22 (13), 2113–2126.
Thompson, C.W., Moore, M.C., 1992. Behavioral and hormonal correlates of alternative reproductive strategies in a polygynous lizard: tests of the relative plasticity and challenge hypotheses. Horm. Behav. 26 (4), 568–585.
Thornhill, R., Gangestad, S.W., 2008. The Evolutionary Biology of Human Female Sexuality. Oxford University Press, Oxford, UK.
Townsend, D.S., Moger, W.H., 1987. Plasma androgen levels during male parental care in a tropical frog (*Eleutherodactylus*). Horm. Behav. 21 (1), 93–99.
Trivers, R., 1972. Parental Investment and Sexual Selection Vol. 136. Biological Laboratories, Harvard University, Cambridge, MA, pp. 179.
van Aert, R., 2016. Puniform: Meta-analysis Methods Correcting for Publication Bias. R.
van Anders, S.M., Goldey, K.L., 2010. Testosterone and partnering are linked via relationship status for women and 'relationship orientation' for men. Horm. Behav. 58 (5), 820–826.
van Anders, S.M., Watson, N.V., 2006. Relationship status and testosterone in North American heterosexual and non-heterosexual men and women: cross-sectional and longitudinal data. Psychoneuroendocrinology 31 (6), 715–723.
van Anders, S.M., Goldey, K.L., Kuo, P.X., 2011. The steroid/peptide theory of social bonds: integrating testosterone and peptide responses for classifying social behavioral contexts. Psychoneuroendocrinology 36 (9), 1265–1275.
van Assen, M.A., van Aert, R., Wicherts, J.M., 2015. Meta-analysis using effect size distributions of only statistically significant studies. Psychol. Methods 20 (3), 293.
Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36 (3), 1–48.
Vleck, C.M., Brown, J.L., 1999. Testosterone and social and reproductive behaviour in Aphelocoma jays. Anim. Behav. 58 (5), 943–951.
Walker, D.A., 2003. JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. J. Mod. Appl. Stat. Methods 2 (2), 26.
Walum, H., Waldman, I.D., Young, L.J., 2016. Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. Biol. Psychiatry 79 (3), 251–257.
Wingfield, J.C., 2017. The challenge hypothesis: where it began and relevance to humans. Horm. Behav. 92, 9–12.
Wingfield, J.C., Hegner, R.E., Dufty Jr, A.M., Ball, G.F., 1990. The" challenge hypothesis": theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. Am. Nat. 136 (6), 829–846.
Wynne-Edwards, K.E., Timonin, M.E., 2007. Paternal care in rodents: weakening support for hormonal regulation of the transition to behavioral fatherhood in rodent animal models of biparental care. Horm. Behav. 52 (1), 114–121.
Ziegler, T.E., Snowdon, C.T., 2000. Preparental hormone levels and parenting experience in male cotton-top tamarins, *Saguinus oedipus*. Horm. Behav. 38 (3), 159–167.
Ziegler, T.E., Jacoris, S., Snowdon, C.T., 2004. Sexual communication between breeding male and female cotton-top tamarins (*Saguinus oedipus*), and its relationship to infant care. Am. J. Primatol. 64 (1), 57–69.
Ziegler, T.E., Schultz-Darken, N.J., Scott, J.J., Snowdon, C.T., Ferris, C.F., 2005. Neuroendocrine response to female ovulatory odors depends upon social condition in male common marmosets, *Callithrix jacchus*. Horm. Behav. 47 (1), 56–64.
Zilioli, S., Bird, B.M., 2017. Functional significance of men's testosterone reactivity to social stimuli. Front. Neuroendocrinol. 47, 1–18.