

Reproducibility in social neuroendocrinology: Past, present, and future

Oliver C. Schultheiss & Pranjal H. Mehta

Please cite as:

Schultheiss, O. C., & Mehta, P. H. (in press). Reproducibility in social neuroendocrinology: Past, present, and future. In O. C. Schultheiss & P. H. Mehta (Eds.), *Routledge international handbook of social neuroendocrinology*. Abingdon, UK: Routledge.

Many scientific disciplines are currently embroiled in what could rightfully be called a revolution. It is a revolution spurred by a crisis in scientists' trust in the reproducibility of published research, that is, whether the finding obtained in a study can be obtained again in subsequent, closely matched or even exact replications of this study. This revolution aims at changing the way we do and report science towards increased transparency and rigour. In this chapter we discuss (i) the origins of this revolution; (ii) examples from social neuroendocrinology relevant to the revolution; and (iii) practical recommendations to strengthen future research in social neuroendocrinology.

1. A primer on reliability, validity, and power

To better understand why reproducibility has become such an issue in the empirical sciences, we think it is useful to briefly sketch out what happens when we test hypotheses, because that will put the reader in a better position to appreciate what can go wrong in the process. The main players in issues of reproducibility are reliability, validity, and statistical power.

Imagine you are a researcher who wants to test the extent to which cortisol is associated with creativity. To examine the association, besides a good measure of creativity (we will not go into that), you need a reliable measure of cortisol. Reliability refers to the precision of a measurement. If you run a sample through a cortisol assay and repeat that process again and again for this sample, will the same value show up again and again? The answer is, of course: never exactly, but hopefully sufficiently close. Thus, measured values of 1.22 ng/mL, 1.18 ng/mL, and 1.24 ng/mL of the same sample would constitute reasonably reliable measurements, because although the values differ, they are close together in terms of absolute levels, converging on an average of 1.22 ng/mL. This would be a reliable assay. In contrast, the same average could be achieved through the following series of measurements: 0.93 ng/mL, 1.39 ng/mL, and 1.42 ng/mL. The latter measurement process is less precise, causing a lot more variance around the estimated average. This would be an unreliable assay. Because cortisol measured with an unreliable (= imprecise) assay is less likely to be correlated with a measure of creativity than cortisol measured with a reliable (= precise) assay, you will opt for the latter method. If you run several studies, you will also be more likely to reproduce a specific association with the reliable assay than with the unreliable one, all else being equal.

But although measure reliability is a necessary precondition for reproducible science, it is not a sufficient one. The second key element for reproducibility is validity; that is, whether a measure actually measures what it purports to measure. For instance, just because an assay kit claims that it measures cortisol does not mean that that is the case. Perhaps the assay picks up something completely different. Perhaps it measures cortisol, but also other steroids, and therefore is not a specific measure of cortisol. Or perhaps it is an exact measure of cortisol in one medium (e.g., serum), but not in another (e.g., saliva). This is a recurring topic in social neuroendocrinology (e.g., Carter et al., 2007; Horvat-Gordon, Granger,

Schwartz, Nelson, & Kivlighan, 2007; Valstad et al., 2017) and an issue that we cover at length with regard to steroid measurements in a separate chapter (see Schultheiss, Dlugash, & Mehta, this volume). Clearly, as a researcher you can only test your hypothesis regarding the link between cortisol and creativity if your measures are valid. Reproducibility can be hampered if more valid measures are used in some studies and less valid ones in others, even if they are equally reliable. Thus, consistency in the use of highly valid measures is key. (Note that this also has another implication: you could get a specific result and be able to replicate it consistently with an invalid measure, but not with a valid measure. This example shows that high reproducibility is not necessarily the same as high validity!).

The third key element for reproducible science is statistical power. According to Cohen (1992), statistical power denotes the probability (in %) of obtaining (or replicating) a statistically significant effect, given a certain sample size (N), statistical threshold criterion (e.g., typically $p < .05$), and population effect size (e.g., expressed as r , d , or odds ratio). Returning to our illustrative example, if you had reason to expect the association between cortisol and creativity to be $r = .40$ in the general population (e.g., based on meta-analytic estimates of associations between cortisol and other psychological variables) and you were to employ the standard alpha level of $.05$ in your research, then you would have an 80% chance of obtaining a significant result if you tested $N = 47$ participants. Or, in other words, if you ran 10 studies with 47 participants each, assuming the true effect size in the population is $r = .40$, the association between cortisol and creativity could be expected to pass the $.05$ significance threshold in 8 of the 10 studies. The sample effect sizes picked up in those 10 studies are expected to scatter around the population effect size, with some coming out higher and some lower. Some of the latter will not exceed the $.05$ threshold.

Of course, if you want to make sure that you get an effect of a certain expected size with greater likelihood, then – all else being equal – you could increase N so that your power will be 99%. Now, you would need to test 106 participants to ensure that you would see the effect emerge as significant in 99 out of 100 studies. If you actually ran those 100 studies, the observed per-sample effect sizes would show much less scatter around the population effect size than effect sizes obtained in the previous set of studies with an N of 47. The reason is because the increase in sample size decreases the width of the confidence interval and thus makes the effect size estimate more precise.

Now let us revisit the population effect size we based our power calculations on. An $r = .40$, equivalent to a d of .87 or an odds ratio of 4.87, is actually a rather rare population effect size in the behavioral sciences if one uses meta-analytic findings as an approximation (and, as we will discuss below, there are good reasons to think that even those may represent overestimations). Richard, Bond, and Stokes-Zoota (2003) conducted a mega-analysis (that is, a meta-analysis of meta-analyses; see Hattie, 2008) based on a century of research in social psychology and found an average effect size of $r = .21$, equivalent to a d of .43 and an odds ratio of 2.18. If we take this as the basis of the population effect size for the hypothesized association between cortisol and creativity and want to make sure that we find a significant effect ($p < .05$) with a power of 80%, we would need to test 176 individuals. And if we want to avoid not seeing the finding in 1 out of 5 studies, but only in 1 out of 10, equivalent to a power of 90%, we would need to test 235 individuals. Samples of this size are the exception rather than the rule in social neuroendocrinology and other domains of psychology. And typically, published effect sizes are much bigger than $r = .21$, too. While this may sound like good news at first blush, small samples combined with large sample effect sizes actually hint at a problem in our field. We will later discuss why.

For now, we conclude that the likelihood of obtaining or replicating a statistically significant effect present in the overall population is a function of the reliability and validity of the measurements involved, the size of the effect, the significance threshold chosen, and the size of one's sample.

2. Some milestones and focal issues of the replication crisis

A key catalyst for the replication crisis and the revolution it triggered was a paper published by John Ioannidis (2005) whose title claimed that “most published research findings are false”. Ioannidis argued that *low statistical power* and *bias in the way research is conducted and published* are critical factors that can lead to the publication of false positive findings (that is, reporting a sizeable effect or a relationship in a study when in fact it is only minuscule or in a different direction altogether in the overall population).

There are several drawbacks to studies with low statistical power. Low power, by definition, indicates that there is a low chance of discovering an effect in a study that is genuinely true in the population. That is, a low-powered study is likely to produce a false negative result, or commit a Type II error (see Figure 1). But even when a low-powered study is “lucky” enough to discover a true effect (typically determined by a statistical significance test), the effect size observed in the low-powered study is likely to be much larger than the true effect size in the population, a phenomenon referred to as the *winner's curse* (Ioannidis, 2008). Somewhat paradoxically, in studies with low statistical power, unreliable measurement can often contribute to such inflated effect size estimates, whereas in studies with high statistical power unreliable measures almost always attenuate observed effect sizes (Loken & Gelman, 2017). And finally, even if the best measures are employed, statistically significant effects can even be obtained for an actual population effect size of 0 due to sampling fluctuations – that is, five percent of all samples drawn from such a population will

yield a significant result based on standard statistical thresholding ($p < .05$; Type I error) and thus represent false positives.

This would not be a problem in a world in which all findings, significant or not, from studies with high statistical power as well as from studies with low statistical power, were eventually published: after a couple of years one could take stock of all studies, meta-analyze them, and come up with a pretty clear-cut verdict whether the overall population effect reliably differs from zero and, if so, how much and in which direction. Unfortunately, in the real world of academia, there is publication bias - many journal editors and reviewers tend to favor the publication of statistically significant results. Thus, the right half of the box depicted in Figure 1, containing both true and false positives, is overrepresented in the literature, a phenomenon that has been termed *excess significance* (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafo, 2013). In contrast, findings falling into the left half of the box shown in Figure 1, containing both true and false negatives, are less likely to get published. This could be termed, in analogy to excess significance, *rare non-significance*.

The preference of journals for positive findings leads to the second factor which according to Ioannidis (2005) contributes to the publication of false findings: *researcher bias*. This refers to the flexibility scientists have in analyzing their data and reporting their results, which, in the context of journals' bias for publishing statistically significant findings, can increase the chances of a false positive result. Examples of such biases, often referred to as "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) or "questionable research practices" (John, Loewenstein, & Prelec, 2012), include:

- The originally targeted outcome does not show the expected relationship, but another, peripheral measure does and is then presented as the focal variable. The more

measures are included in a study, the more likely it is to find a significant effect for at least one of them due to mere chance.

- Cases or entire experimental conditions are dropped that keep results from becoming statistically significant (that is, dropping these cases or conditions turns a non-significant result into a statistically significant one), but the research report is silent about this fact.
- Sample size is not determined ahead of time but when results finally get significant and data collection is then stopped. This optional stopping strategy inflates the Type I error rate (Simmons et al., 2011; but also see Lakens, 2014).
- Stringent control variables are not included, because they would decrease effect size and significance levels. The opposite also represents researcher bias: including covariates for no apparent reason other than that results get significant only when they are included.
- The original hypothesis does not pan out, but another, unexpected finding emerges from the research and is then presented in the manuscript as the a priori hypothesis (this is called *hypothesizing after results are known*, or HARKing [Kerr, 1998], and used to be the officially recommended approach for writing papers in psychology; see Bem, 2003; Sternberg, 2003).
- A less valid or less rigorous measure is preferred over a more valid or rigorous one, simply because the former is more likely to support one's hypothesis than the latter.
- Researchers typically scrutinize all methods and data-processing steps particularly carefully when a study fails to support their hypothesis, hunting for methodological glitches that could explain this negative outcome. But they do not go to the same lengths for studies that support their hypothesis.

And bias can occur in innumerable other ways, and even in the most well-intentioned and principled researchers. After all, science is a human endeavor and thus susceptible to the many ways in which we protect, defend, and uphold our most cherished concepts and hypotheses.

Meta-analyses based on such a body of literature are bound to overestimate true effect sizes and may even indicate a reliable effect size across studies when the relationship in question is nil in the population (Bakker, van Dijk, & Wicherts, 2012). Bakker et al ran simulations showing that even for a population effect size of zero, underpowered studies that also exploit questionable research practices can lead to meta-analytical effect sizes estimates (*d*) of up to .48. Such estimates cover a large spectrum of the average effect sizes in actually published meta-analyses (e.g., Richard et al., 2003)!

Ioannidis' (2005) paper was largely aimed at biomedical and epidemiological research, and his key arguments had been made before (e.g., Cohen, 1962; Greenwald, 1975; Meehl, 1967; Smith, 1970). But his paper had an immense impact not only on the fields it focused on, but across many other scientific disciplines, too. One reason for this effect may have been its provocative title. The other reason may have to do with the diminishing returns on investment in biomedicine at the time. For instance, two independent investigations that examined the reproducibility of published landmark studies in oncology and drug development found that results for only 11% (Begley & Ellis, 2012; 53 studies) to 25% (Prinz, Schlange, & Asadullah, 2011; 67 studies) of them could be reproduced. One group of researchers even put a price tag on the squandering of resources on non-reproducible published findings in biomedical research: US\$ 28 billion per year in the USA alone (Freedman, Cockburn, & Simcoe, 2015). Thus, although funding agencies spent more and more money on

understanding and curing illnesses such as cancer, scientific progress had slowed down noticeably, thus failing to deliver to patients the effective treatments they urgently need (Harris, 2017).

But the impact of Ioannidis' paper was not limited to the biomedical sciences. Vul, Harris, Winkielman, and Pashler (2009) applied Ioannidis' argument to functional magnetic resonance imaging (fMRI) studies of emotion, personality, and social cognition, arguing that much of that research is severely underpowered and capitalizes on chance by presenting those among up to 100,000 data points representing the brain that show significant ($p < .01$) or highly significant ($p < .001$) activation (see also Button et al., 2013). Even if there is no actual systematic brain activation effect, this would still result in 1,000 or 100 significant voxels, respectively. Bennett, Miller, and Wolford (2009) provided a humorous illustration of this issue by reporting that a dead salmon completing a social-perspective-taking task in an fMRI scanner showed significant brain activation – an artifact of improper statistical thresholding.

The discussion of problematic practices in analyzing and reporting data spilled over into mainstream psychology after the publication of a paper claiming to provide evidence for precognitive abilities in humans (Bem, 2011) and the criticism it drew regarding its data-analytical strategy (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). The discussion further intensified when the Open Science Collaboration (2015) - which included one of us (PHM) - tried to replicate findings from 100 studies published in three leading psychology journals in the year 2008, using designs with high statistical power and adhering as much as possible to the original, published study protocols. Only 39% of the studies were deemed to represent successful replications, and the effect sizes of the replication studies were almost always substantially lower than those reported in the original publication (see Camerer

et al, 2016, for related observations in the field of economics). Thus, whether it is biomedical or behavioral research, Ioannidis's (2005) prediction about the low reproducibility of published research turns out to be a valid diagnosis of the state of several scientific disciplines.

3. The way we were: A personal look back

In the following sections of this chapter, we will present some case studies highlighting challenges with reproducibility in social neuroendocrinology. Our objective is not pointing fingers at others but presenting evidence, acknowledged by the originators of the research, for problematic research strategies in our own field. As researchers who have been, and still are, passionate about our own pet theories and hypotheses and who have been socialized into academia under the "old rules" that contributed to the replication crisis, we do not claim to have been immune to the problems of conducting and reporting research outlined previously.

As a case in point, the first author of this chapter (OCS) recognizes many of the processes that lead to published false positives in his own first published social neuroendocrinology paper (Schultheiss, Campbell, & McClelland, 1999). The study was conducted in 1997 to test a simple idea: Winning or losing a dominance contest leads to increases or decreases in testosterone in men with a high, but not in those with a low, implicit need for power (nPower). Thus, it was designed to test an interaction between an experimentally varied factor (winning versus losing) and a quantitatively assessed motivational disposition. Assuming a medium-sized (Cohen, 1992) correlation of .30 between nPower and testosterone change among winners and a similarly sized negative correlation of -.30 in losers (already a generous estimate, given the typical personality/social psychology effect size of $r = .21$; Richard et al., 2003), this would have required at least 44 participants per experimental condition and thus a total of 88 participants to detect the expected effect

with 80% probability at $p < .05$. But the first author only had resources to test 42 participants in total. Even if the estimated effect size had been true, this cut his effective power down to a 46% chance of observing the predicted effect, making it an underpowered study from the get-go. Then, after all data were collected, the picture stories had been coded for nPower and the last saliva sample run through the gamma counter, the pre-planned regression analyses testing the hypothesized effect revealed – nothing! There was no sign of a significant interaction between contest outcome and nPower in the data.

When OCS had recovered from this blow, he began to scrutinize all data file processing syntax for signs of data misalignment or any evidence for data miscoding. When that could be ruled out, he remembered that in past research on power motivation, researchers had often distinguished between subtypes of power motivation that could be assessed with more differentiated coding systems than the one he had used. So he took advantage of the flexibility inherent in this approach and recoded the stories using a number of different variants of nPower coding systems. This eventually turned up the desired evidence: a highly significant three-way interaction between contest outcome, an egocentric variant of power motivation termed personalized power (p Power), and a more social variant termed socialized power (s Power). The effect was due to a positive association between p Power and testosterone changes in winners in the absence of s Power that did not occur in s Power-present winners or in losers in general. Overall, the regression model accounted for 46% of the variance, corresponding to $R = .68$. Individual correlation coefficients in follow-up decompositions of the interaction reached levels of almost .90, but were based on n s as low as 5. This is the set of findings that was eventually published. No mention was made in the published paper that other coding systems had also been used for assessing power motivation, but without success. The published paper is thus a textbook case of the undisclosed flexibility

that researchers have when analyzing and reporting their findings, the spectacular effect sizes that can be associated with underpowered samples, and the tendency of reviewers and journal editors to accept papers with such spectacular results for publication.

So were the findings ever replicated? The answer is: it depends. The first replication attempt (Schultheiss & Rohde, 2002) was unsuccessful. These authors tried the p/s Power coding approach, failed, but did not report this fact in their paper. Instead they switched to an analytic approach that combined an overall measure of nPower with a word-count index of inhibition to predict testosterone responses to a dominance contest outcome (Schultheiss et al., 1999, had tried this, too, but without success and without mentioning it in the paper). Again, the study was underpowered ($N = 66$), particularly with regard to the complex three-way interaction of contest outcome, nPower, and the inhibition index. The flexibility in OCS's switch of analytical strategy from Schultheiss et al (1999) to Schultheiss and Rohde (2002) is only barely defensible on the grounds that previous research unrelated to hormones has documented functional similarities between p Power and uninhibited nPower and s Power and inhibited nPower (e.g., McClelland, Davis, Kalin, & Wanner, 1972) and the fact the he was generally right in the sense that high levels of power motivation did predict testosterone increases in winners and/or decreases in losers, once a moderating factor was taken into account.

The nPower x contest outcome interaction originally expected by the first author emerged in later studies with better statistical power, and not just for testosterone (Schultheiss, Wirth, Torges, Pang, Villacorta, & Welsh, 2005; Oxford, Tiedtke, Ossmann, Özbe, & Schultheiss, 2017), but also for cortisol (Wirth, Welsh, & Schultheiss, 2006) and, in women, for estradiol (Oxford et al., 2017; Stanton & Schultheiss, 2007). Nevertheless, these

studies also feature an inconsistent result, namely the reversed nPower x contest outcome effects for men observed by Oxford et al (2017; cf. Schultheiss et al., 2005). It remains to be resolved whether these paradoxical results may have been due to differences in the contest paradigms employed. OCS also had conducted one additional study with 56 male US students, but failed to find a significant nPower x contest outcome effect on post-contest testosterone changes. A reanalysis of the data for the purposes of this chapter suggests that the predicted effect of a positive correlation between nPower and testosterone changes among winners and a negative correlation among losers was present immediately after the contest, but did not reach accepted significance levels for the interaction, $p = .18$. The results were therefore not published (filedrawer $N = 1$), although these participants were included as part of Wirth et al.'s (2006, Study 2) report on nPower x contest outcome effects on cortisol changes.

Vongas and al Hajj (2017) provide a final twist to this story. These researchers used the same experimental paradigm originally introduced by Schultheiss et al (1999), but reframed it such that the contest outcome would be an indicator of future leadership ability and also took other steps to make it methodologically more sophisticated. Vongas and al Hajj (2017) measured p Power with a slightly different coding system than Schultheiss et al (1999), by using all nPower coding categories of the Winter (1994) coding manual except the prosocial one that focuses on unsolicited help and advice (personal communication by John Vongas). Across two studies with statistical power approaching 80% ($Ns = 84$ and 72), they reported significant contest outcome x p Power interaction effects on testosterone changes. They thus provided a partial replication – *sans* the contribution of an s Power measure -- of the Schultheiss et al (1999) findings and also of the findings reported by Schultheiss et al (2005).

What is the bottom line of this personal account of one researcher's career in social neuroendocrinology? It shows that a passion for a hypothesis, coupled with a journal system that rewards underpowered research yielding over-the-top effect sizes, is likely to produce published findings that, through no ill will or deliberate intention to deceive on the part of the author, the editor, and the reviewers, are difficult to replicate, unless some undisclosed flexibility is used, and that therefore should be viewed with caution. At least there was a trajectory from underpowered (Schultheiss et al., 1999) to less underpowered (Schultheiss & Rohde, 2002) and finally adequately powered studies (e.g., Oxford et al., 2017; $N = 326$) and towards the inclusion of data sets and data analysis scripts with the publication for the sake of transparency (Oxford et al., 2017). Still, there is a degree of variability and inconsistency in the line of work that started with Schultheiss et al (1999) that remains to be resolved in direct replication studies, conducted as preregistered studies or registered reports (see below).

4. Power posing

While the aforementioned example illustrates that the field of behavioral endocrinology probably has never been immune to the pitfalls of false-positive research, this study and many similar others from various laboratories did not lead to public critiques or systematic discussions of their merits and weaknesses, although this would have been justified. By contrast, Carney, Cuddy, and Yap's (2010) study did lead to public critiques and discussions. These researchers tested the idea that a brief enactment of body postures signaling high social power, compared to postures signaling low power, would lead to corresponding changes in physiology, feeling, and behavior. Testing 42 participants randomly allocated to either the high-power or the low-power condition, Carney et al. found that participants in the former condition showed an increase in salivary testosterone, a decrease in salivary cortisol, a strong sense of subjective power, and a strong propensity towards risky

decision-making. In contrast, participants in the low-power condition showed a decrease in testosterone, an increase in cortisol, felt less powerful, and also made less risky decisions after the intervention. In the abstract, the authors drew the following conclusion from these findings: “That a person can, by assuming two simple 1-min poses, embody power and instantly become more powerful has real-world, actionable implications” (Carney et al., 2010, p. 1363).

Responding to mounting criticism that power-posing effects could not be replicated in other laboratories (e.g., Simmons & Simonsohn, 2017; Davis et al., 2017; Garrison, Tang, & Schmeichel, 2016; Raney, Dreber, Magnusson, Leiberg, Sul, & Weber, 2015), Carney, Cuddy, and Yap (2015) argued that effects of power posing can and have been replicated in many studies (see Cuddy, Schultz, & Fosse, in press, as well as Gronau, Van Erp, Heck, Cesario, Jonas, & Wagenmakers, 2017, for recent updates on the effects of power posing on feelings of power). But we note that from a social neuroendocrinology perspective, it is remarkable that none of the studies Carney et al (2015) cite in support of their hypothesis ever focused on the hormonal effects originally reported in Carney et al (2010). These effects had already come under scrutiny by Stanton (2011) who argued that treating gender as a covariate in the analyses reported by Carney et al (2010) did not do justice to the differences in overall levels and the specific mechanisms of testosterone release in men and women. With the benefit of hindsight, we would add that the sample collection and hormone assay methods used by Carney et al (2010) now appear to have had doubtful validity, too (Schultheiss, 2013; Schultheiss, Dlugash, & Mehta, this volume; Welker, Lassetter, Brandes, Prasad, Koop, & Mehta, 2016) – a caveat that also applies to many other studies, including some of those from our own laboratories.

So were there any attempts at replicating Carney et al's (2010) endocrine effects, and what did they find? We could identify four studies that had specifically aimed to replicate these effects (Davis et al., 2017; Ranehill, Dreber, Jonahhesson, Leiberg, Sul, & Weber, 2015; Ronay, Tybur, van Huijstee, & Morssinkhof, 2017; Smith & Apicella, 2017) and provide an overview of them in Table 1, next to Carney et al's (2010) original results. A careful analysis of these studies and comparison with the original power-posing study reveals the following. First, three of the four replication studies – with Davis et al (2017) being the exception -- feature substantially larger samples than the original study, determined by statistical power analyses and often increasing sample sizes beyond the results of these analyses. Second, all three four attempted to replicate the original power-posing manipulation closely (i.e., taking on the two consecutive postures in each condition described in the original paper for at least 1 min), but there were some differences as well. Davis et al (2017) manipulated power posing in participants with a diagnosis of social anxiety disorder before they entered a free-speech task. Ranehill et al. (2015) used longer times for the poses than the original study. Two studies aimed at minimizing experimenter effects by having a computer provide the instructions (Ranehill et al, 2015; Smith & Apicella, 2017). Smith and Apicella (2017) added another between-subject control condition with a neutral power pose and an additional factor, namely, whether participants had previously won or lost in a game of tug-of-war against another participant. Notably Ronay et al's (2017) study was a registered report that was accepted in principle before data collection even started (more on registered reports below). Third, the studies by Davis et al (2017), Ranehill et al (2015), and Ronay et al (2017) suffer from the same problem of treating gender as a covariate already criticized by Stanton (2011), and all four studies, like the original study, base their conclusions on hormone assays whose validity for the accurate assessment of testosterone has recently come under scrutiny (see Schultheiss

et al., this volume; Welker et al, 2016). However, because all four studies are comparable with regard to these last issues, these general criticisms cannot explain any between-study differences in the results reported.

Table 1 shows that none of the replication studies was able to replicate either the testosterone increase or the cortisol decrease associated with high-power postures, relative to low-power postures, reported by the original study. In fact, the pooled effect size as well as the individual effects reported by each replication study suggest that the effect of power posing on testosterone may be in the opposite direction of the effect reported in the original study (i.e., a sign, or Type S, error according to Gelman & Carlin, 2014). In terms of absolute magnitude the pooled effect sizes of all replication studies are in the small range (Cohen, 1992), whereas the effect sizes originally reported by Carney et al (2010) are medium- to large-sized (a magnitude, or Type M, error according to Gelman & Carlin, 2014). With regard to the effects of power posing on endocrine variables, Carney et al.'s (2010) original study is thus a good example of journals favoring the publication of studies with seemingly large, positive effects that are difficult to replicate.

There is also evidence for low statistical power and undisclosed flexibility contributing to Carney et al's (2010) results. The low statistical power of the original study is evident from the sample size, which unrealistically presupposes a large effect ($d = .78$) for simple between-group comparisons at $p < .05$ and a power of 80%. Some of the flexibility that went into data processing is disclosed in the original publication – one outlier was omitted from the analyses for testosterone, two were omitted from the analyses for cortisol. Crede and Phillips (2017) reanalyzed the Carney et al. (2010) data and demonstrated that not only the removal of outliers played a role in obtaining the results for testosterone and cortisol eventually reported in the original publication, but also other data-analytic decisions not

discussed in Carney et al. (2010; and to be fair: usually not addressed in *any* published studies). These included the use of an ANCOVA approach that treated baseline hormone measures as a covariate instead of an analysis of hormone change scores (see Van Breukelen, 2005), the question of adding the baseline of the “other” hormone (e.g., cortisol) whenever an analysis was directed at a target hormone (e.g., testosterone), and the decision to use gender as a covariate instead of analyzing the data for each gender separately. Crede and Phillips (2017) identify in each of their tables for testosterone and cortisol 54 possible ways to analyze the data resulting from the permutation of decisions related to outlier omission, covariate selection, and the consideration of gender. And for each hormone, only one out of 54 combinations of data-analytic decision yielded the desired $p < .05$ effect. These were the effects that were eventually published.

Carney et al. (2010) have to be given credit that they made their data set available to others for reanalyses, one of which led to the Crede and Phillips (2017) publication. Dana Carney, the lead author, eventually agreed with the critics of the original study, stating that in light of the many failed replication studies and the strong reservations other researchers voiced about the original study, she does “not believe that ‘power pose’ effects are real” (Carney, undated). In the same statement, Carney also conceded that the sample was too small, that gender should have been treated differently than as a covariate in the analyses of testosterone, and that the sample had been filled up until effects on the original focal measure – risk taking – became significant at $p < .05$ in one type of analysis, which was eventually reported, but not in another, equally valid analytic design. She now characterizes these decisions as instances of p-hacking (i.e., ensuring that an effect just barely makes the .05 significance level; Simmons et al., 2011) and using researcher degrees of freedom (Carney, undated). In essence, she confirms Crede and Phillips’ (2017) conclusion that key findings of

the Carney et al. (2010) study must be attributed not to an actual effect present in the overall population, but to exploiting undisclosed flexibility in data collection, processing, and analysis (note that this may be a premature conclusion with regard to feelings of power induced by power posing; see Cuddy et al., in press; Gronau et al., 2017).

In our experience, it is a rare case that a researcher who was once passionate about a study and its findings not only responds to criticism by making her or his data available to others, but eventually acknowledges that he or she may have gotten it wrong. We suspect that in their heart of hearts, all scientists know that some of the studies they have published over the years are less likely than others to actually represent a true effect that can be replicated. But very few would admit this openly and thus help pave the way to the self-correction that should be part and parcel of the scientific endeavor, but in reality hard to come by. Following the discussions about many other findings that have come under scrutiny in recent years (the Carney et al., 2010, study is hardly the only one), our impression is that defensiveness is the modal response of most researchers who have published findings that are then critiqued by others. Ultimately, this is a problematic response for scientists, even though it may be viewed as a very human response, because it leads to costly and unproductive debates instead of helping to clear the path toward better and more conclusive studies that can correct the scientific record. Carney, in our view, has done the painful, but right thing. Kudos.

5. A peek inside one laboratory's file drawer

A third illustrative example of the danger of false-positive science in social neuroendocrinology comes from the laboratory of Olivier Luminet (Lane, Luminet, Nave, & Mikolajczak, 2016). Luminet and his collaborators have conducted a series of eight studies in which oxytocin or placebo was administered via an intranasal spray to participants and the effect of this treatment on social behavior was examined. In terms of experimental design,

hormone administration, and sample size, this research is similar to a large number of recent studies from various laboratories examining the effects of oxytocin administration on affiliative behavior, social cognition, and trust (e.g. De Dreu et al., 2010; Hurlemann et al., 2010; Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). And initially, Luminet’s group was successful, too, with studies suggesting that oxytocin increases trust (Mikolajczak, Pinon, Lane, de Timary, & Luminet, 2010 [note that Lane et al., 2016, p. 11, later stated that if analyzed properly, the effect reported in this paper was not reliable]; Mikolajczak, Gross, Lane, Corneille, de Timary, & Luminet, 2010) and a willingness to share emotions with others (Lane, Luminet, Rime, Gross, de Timary, & Mikolajczak, 2013), with one additional paper suggesting that oxytocin effects on social behavior may depend on alexithymia, a self-report measure of individual differences in accessing one’s emotions (Luminet, Grynberg, Ruzette, & Mikolajczak, 2011).

However, the Luminet laboratory also conducted several studies that failed to document hypothesized effects of oxytocin on dependent measures. One was a failed replication of the oxytocin effect on trust. This paper, although initially rejected by one journal (see Lane et al., 2016), was eventually published by another journal, and its findings are therefore on record (Lane et al., 2015). Three other papers, documenting null effects of oxytocin administration on social conformity, social mimicry, and compassion, were submitted to journals, but repeatedly rejected and remain unpublished (see Lane et al., 2016), illustrating how the previously discussed “rare non-significance” in the published literature comes about. Lane et al (2016, p. 2) drew the following conclusion from this state of affairs:

“After realising that our publication portfolio has become less and less representative of our actual findings, and because the nonpublication of our null results might contribute to generating a publication bias in [intranasal oxytocin administration]

research, we decided to retrieve these studies out of our drawer, hoping that other laboratories will do the same.”

Lane et al (2016) conducted a meta-analysis of all 8 studies conducted in the Luminet laboratory, comprising a total of 453 research participants, 25 research paradigms, and 13 dependent variables. It should be noted that in 6 of these studies, 60 participants or more were tested, thus providing these studies the statistical power to reliably (> 80%) detect mean-difference effects sizes of $d = 0.65$, which corresponds to a medium-to-large effect size (Cohen, 1992). It is our impression that such sample sizes are already rather high compared to many other hormone administration studies and thus provide a somewhat better basis for testing targeted hypotheses.

The results of the in-lab meta-analysis yielded sobering results. An overall analysis of all emotion, cognitive, and behavioral dependent variables revealed an average effect size of $d = .003$, which was not significantly different from a null effect. This finding was not moderated by the type of dependent measure used or the specific theory of oxytocin’s effect tested.

Lane et al (2016) state that the difficulty in replicating effects of oxytocin on theoretically relevant outcomes has turned them from believers into skeptics. Addressing the question how the large literature on oxytocin effects in humans could have accumulated in the first place, given this difficulty to replicate basic effects, they offer two explanations. One is that the publication record in its totality represents false positives, resulting from selective publication of the few studies out of many conducted that happened to cross the threshold of statistical significance. The implication is that in the overall population, the effect of intranasally administered oxytocin on social cognition and behavior of the type tested in

published studies is close to zero and that the hypothesis of oxytocin effects on social cognition and behavior is plain wrong.

The other explanation Lane et al (2016) offer is that effects of intranasal oxytocin in studies of human participants is due to methodological and statistical artifacts. These may include pre-existing differences between verum and placebo groups, particularly in small-*N* studies, single-blind studies in which the experimenter knows about a participant's treatment-group assignment, although the participant is not informed (see Rosenthal & Rosnow, 1969), unsubstantiated assumptions about the dosage and timeline of intranasal delivery of oxytocin and its effects on the brain (e.g., Leng & Ludwig, 2016; Walum, Waldman, & Young, 2016), and unrealistically high targeted effect sizes, given the sample sizes tested.

Lane et al. (2016) do not rule out the possibility that effects of oxytocin on social cognition and behavior are moderated by situational or dispositional factors, leading to a strong effect under one set of boundary conditions and a null effect under another. This would be consistent with the difficulty of replicating any direct main effects of oxytocin administration. But, as the authors emphasize repeatedly in their paper, the most problematic aspect of intranasal oxytocin research is the lack of published *direct* replications, be it of the main effect of hormone administration or any interaction effects with other variables.

They draw the following conclusion from their own data and their knowledge of the field of oxytocin research:

“We believe that a systematic shift in the [intranasal oxytocin administration] publication process is essential for revealing the true state of the world. Pre-registration of ex-ante hypotheses, replication attempts of the findings before their submission, and the submission of null results and failed replications for publication, especially when the studies are well-powered to detect the original findings, should be encouraged. Review processes

should insist on fully reporting all of the candidate moderators that were measured and tested and encourage publication of well-conducted studies, regardless of their results.” (Lane et al., 2016, p. 13)

We fully agree with Lane et al.’s (2016) conclusions, which are also echoed by Walum et al. (2016), and applaud them for the scientific integrity they demonstrate by laying open their lab’s research record.

6. Gathering around the campfire of highly reproducible research

Our previous examples document that social endocrinology is not immune to the factors that gave rise to the replication crisis sweeping psychology and other sciences. This should not be surprising, given that our field has been subject to the same incentives, pressures, academic socialization practices, and publication biases as other fields. It seems that the constant hunt for sexy and sensational new findings may lure our field into the dark wilderness of irreproducible research. What is a better alternative?

It is our impression that the behavioral sciences have made the greatest progress and achieved the most profound insights when they have focused on well-documented, highly reproducible findings. Pavlovian conditioning, that is, the learned association between unconditioned rewards and punishers and the stimuli that reliably predict them, is a classic example. Because Pavlovian conditioning is a very robust phenomenon resulting from a well-defined general testing paradigm, this type of learning has become a prime vehicle for research on such different topics as animal cognition, emotional processing, or molecular changes involved in learning (e.g., LeDoux, 1996, 2002). It became the backbone of research in biopsychology and neuroscience, helping to make both Noble-prize-winning discoveries (e.g, Eric Kandel’s research on synaptic changes in the *aplysia californica*) and to further the progress of our understanding of brain functions in general. Indeed, it could be argued that

biopsychology and neuroscience in their present form are inconceivable without the firm fundament of Pavlovian conditioning and its close cousin, instrumental conditioning. Over time the limits of Pavlovian learning processes also became apparent (Seligman, 1970) and some specific findings resulting from Pavlovian conditioning paradigms have turned out to be hard to replicate (e.g., Maes et al., 2016). But the phenomenon itself was so robust and pervasive that its limits were not apparent for a long time and actually required clever experimental setups to document them (Garcia & Koelling, 1966).

Other examples of strong testing paradigms producing valuable phenomena for entire scientific disciplines include the Stroop test for the cognitive sciences (MacLeod, 1991), binocular rivalry in consciousness studies (Dehaene, 2014), the ultimatum game for the study of decision-making (Güth, Schmittberger, & Schwarze, 1982), and the strange situation test (Ainsworth & Bell, 1970) and the marshmallow test in developmental psychology (Mischel, Shoda, & Rodriguez, 1989). Does social neuroendocrinology have similarly robust testing paradigms that can be used as a secure platform from which substantial questions can be addressed?

Without a doubt, our field features such robust campfires of replicable phenomena around which researchers can gather and use them to shine a light on new questions. They appear to come in two varieties: testing paradigms that are either strong for a-priori, conceptual reasons or that result from a serendipitous observation with a well-replicated empirical track record. An excellent example of a strong conceptual testing paradigm in social neuroendocrinology work with animal models is gonadectomizing individuals and then reinstating hormones through external administration. This allows to bring hormone concentrations under experimental control and study their causal effects on, for instance, sexual preference, mating behavior, parenting, or aggression. This approach is typically

combined with robust testing paradigms imported from other subdisciplines (e.g., learning psychology, ethology, neuroscience; see, for instance, Nelson, 2011).

Because gonadectomy, combined with hormonal reinstatement, is not a research approach feasible with human participants, options are more limited in human social neuroendocrinology. Naturally occurring variations in hormone levels, such as circadian rhythms, during the menstrual cycle, or in the transition from fertility to menopause, can serve as alternatives, again for well-established a-priori reasons. But they do not allow the same strong causal inferences that can be drawn from purely experimental animal models. Moreover, for studies trying to exploit menstrual-cycle effects, there is the difficulty of determining cycle phase (Blake, Dixson, O’Dean, & Denson, 2016). And this line of research is not immune to the dangers of false-positive science, as reflected by the discussions surrounding menstrual-cycle “effects” on political preferences (Durante, Rae, & Griskevicius, 2013; Gelman & Carlin, 2014), mate preferences (Gildersleeve, Haselton, & Fales, 2014; Harris, Pashler, & Mickes, 2014), or clothing style (Blake, Dixson, O’Dean, & Denson, 2017; Eisenbruch, Simmons, & Roney, 2015). Still, as several chapters in this volume document, if used prudently, the natural-variation approach can yield systematic insights into neuroendocrinological phenomena such as interhemispheric coupling (Hausmann & Burt, this volume), memory and decision-making (Hampson, this volume) or emotional and motivational processes (Gingnell, Hornung, & Derntl, this volume; Diekhof, Reimers, & Holtfrerich, this volume).

One strong testing paradigm available not only to researchers working with animals, but also to those working with humans, is the administration of exogenous hormones, sometimes combined with the temporary pharmacological suppression of endogenous hormones. Although the validity of such methods is still an issue of debate in the case of

peptides like oxytocin (see above; Leng & Ludwig, 2016), this method is viable for steroid hormones, whose free, unbound fraction readily passes the blood-brain barrier. Thus, the mechanism by which steroid administration affects the brain are well understood, and there is evidence of its efficacy using physiological indicators. For instance, testosterone can be administered orally and its effects on genital function have been documented (e.g., Corona, Rastrelli, Morgentaler, Sforza, Mannucci, & Maggi, in press; Tuiten, van Honk, Koppeschar, Bernaards, Thijssen, & Verbaten, 2000). Moreover, testosterone administration can be combined with the administration of a gonadotropin-releasing-hormone antagonist, which results in a downregulation of the hypothalamic-pituitary-gonadal axis (Goetz, Tang, Thomason, Diamond, Hariri, & Carre, 2014). This treatment transiently reduces gonadal (endogenous) testosterone release to hypogonadal levels, making the experimental paradigm akin to animal studies using gonadectomy with subsequent exogenous hormone reinstatement.

Other well-documented hormone-manipulation interventions include the downregulation of the hypothalamic-pituitary-adrenal axis through dexamethasone suppression (also termed “chemical adrenalectomy”; Lupien & McEwen, 1997, p. 21) or, conversely, the simulation of a strong stress response through the administration of hydrocortisone (e.g., Schwabe, Tegenthoff, Hoffken, & Wolf, 2010). Of course, both approaches can be combined to mimic animal studies in which the hormone-producing gland is removed and hormone levels are then restored through exogenous hormone administration (Lupien & McEwen, 1997). Overall, (steroid) hormone administration studies, particularly when coupled with transient glandular suppression, provide excellent testing paradigms with well-described underlying mechanisms of hormone function and proven effects on relevant outcome measures. Note, however, that this does not necessarily imply that all targeted outcomes will be affected in a hypothesized manner or that this research is immune to low

statistical power, undisclosed flexibility, or publication bias. Our argument is that when these obstacles to greater scientific rigor are removed, the hormone administration/suppression approach represents a well-described, mechanistic tool for elucidating causal mechanisms of hormones in social neuroendocrinology.

Sometimes strong testing paradigms are also the result of a serendipitous observation, backed up by highly consistent empirical replications, like in the case of the Trier Social Stress Test (TSST, Kirschbaum, Pirke, & Hellhammer, 1993). In the TSST, the experimenter asks research participants to first prepare (anticipation period, 10 min) and then actually give an impromptu job application presentation in front of an unresponsive jury of two confederates of the experimenter (5 min). Subsequently, participants also perform a math task in front of the jury by counting down from 1,022 in steps of 13. If they make a mistake, a jury member asks them to start over again from 1,022 (5 min). Saliva or blood samples taken before (baseline, anticipation), during (speech, math task), and after the TSST show a robust and strong cortisol increase starting during the anticipation period, peaking about 40 min after TSST onset, and returning to baseline levels about 90 min after TSST onset (Goodman, Janson, & Wolf, 2017; Kirschbaum et al., 1993). (The TSST also elicits strong responses for other hormonal and psychophysiological parameters such as prolactin, growth hormone, heart rate, or blood pressure.)

In general, salivary cortisol levels rise two- to threefold in the majority of participants (Kudielka, Hellhammer, & Kirschbaum, 2007), thus making the stress-axis effect elicited by the TSST a large-sized one. This suggests that the social-evaluative stress that characterizes the testing situation in the TSST is a near universal elicitor of strong endocrine responses in humans. Goodman et al's (2017) meta-analysis of the effect of the TSST on cortisol responses yields a large effect size of $d = 0.925$, based on the within-subject comparison between the

pre-TSST baseline and post-TSST saliva samples (see also Dickerson & Kemeny, 2004). This suggests that in studies assessing TSST-induced cortisol changes within subjects, a sample size of 9 participants would be sufficient to detect the stress-induced cortisol effect with a probability of 80% at $p < .05$. For a 90% probability, a total sample size of 12 would be sufficient. The size of the TSST effect is all the more remarkable as it goes in a direction opposite to the circadian cortisol drop during waking hours usually observed over assessments covering similar time spans as the TSST procedure. Adding a non-stressful control group therefore will typically yield similar or even higher effect sizes when comparing cortisol concentrations in samples collected after the end of the TSST and control procedures (e.g., Wiemers, Schultheiss, & Wolf, 2015). Thus, the TSST represents a robust paradigm for stimulating a strong, highly replicable endocrine stress response by psychological means.

Due to its robustness, the TSST is now a frequently used platform for exploring stress responses and their relationship with other factors, such as gender differences, age, mental health, social support, or immunological changes (see Kudielka, Hellhammer, & Wüst, 2009; Kudielka & Zänkert, this volume; Rohleder, this volume). We think it is notable that so far the TSST represents the only testing paradigm in which a standardized psychological situation produces such a robust hormonal response. Similarly strong psychological-stimulation paradigms for other hormones, such as testosterone, estradiol, progesterone, but also peptides like oxytocin or vasopressin, are sorely missing so far, and their development remains an important task for future research in social neuroendocrinology.

7. Conclusion and recommendations

So far in this chapter, we have chronicled the replication crisis in science, with a particular focus on the behavioral sciences and the core reasons for why the crisis came about. We have pointed out that a combination of factors on the side of researchers (underpowered

studies, combined with undisclosed flexibility in analyzing and reporting results) and journals (valuing novel, significant, strong-effect findings over replications or null results, even if based on methodologically rigorous studies) has fueled this crisis by producing findings that cannot be replicated. We have shown by example that social neuroendocrinology is not immune to findings that have been difficult to replicate, but also pointed out that there are clear-cut cases of solid, replicable research built on conceptually or empirically strong testing paradigms. All is not lost, so to speak. But what can researchers in our field do to improve the quality of future studies and insure that false-positive findings are minimized?

We propose that the royal road to a better science of social neuroendocrinology is based on Chambers's (2017) model of registered studies. This approach requires authors, in a first stage, to undergo peer review of a proposed study, justifying its importance and necessity, listing and justifying the hypotheses, and detailing its methods. Requests by editors and reviewers for changes regarding the theory or the methods can be addressed in subsequent revision(s), until in-principle acceptance is given. Only then can the study author start collecting data. In a second review stage, the originally involved editor and reviewers then evaluate the final paper (now including results and discussion sections), verifying that the study has been conducted as approved and that results are presented and discussed in line with the originally proposed hypotheses. The paper can also include additional, post-hoc exploratory analyses clearly labeled as such. At this second stage, editors and reviewers can also request revisions. The final revised paper is then published. The most important aspect of registered reports is, however, that in-principle acceptance is not based on the results of a study (i.e., whether they are significant, have large effect sizes, etc), *but on the merit of the question it tries to answer and the rigor of the methods it employs*. Registered reports thus represent a complete departure from the de facto model of scientific review and publishing,

which was always overfocused on outcomes. To the extent that this model, which we deem to be the best one presently available, is gradually adopted by journals and researchers in our field, one important goal will be to monitor whether it generates unintended side effects. After all, the traditional model of publishing, which gave rise to the present reproducibility crisis, was not implemented to generate false positives. Yet it did. Incentives and their boundary conditions can sometimes generate truly weird side effects, as Skinner's (1948) famous case of superstitious behavior in pigeons illustrates.

Realistically, however, it may take a while until the journals that social neuroendocrinologists typically publish their work in will adopt the full-blown registered-study approach championed by Chambers (2017). And even if they do, at least initially this is unlikely to be a requirement for all submitted studies but rather an optional feature. In the meantime, what else can be done to enhance the quality of science in our field? We suggest the following measures might be helpful:

a) Preregister studies

Even if the journals you usually submit your work to do not feature full preregistration in the sense of peer-review of the proposed study and its rationale and methods, you can preregister your research plan, your hypotheses, and your planned analyses in a time-stamped manner on sites such as aspredicted.org or osf.io. However, it is crucial that you be as specific as possible with regard to the hypotheses, targeted sample size and its power-analysis justification, independent and dependent variables (in the case of experimental paradigms) or measured variables (in the case of correlational research designs), analyses, and dealing with outliers and exclusions (see <http://datacolada.org/64>). This way, you can ensure that when the research is done and you are about to submit your work to a journal of your choice, editors

and reviewers can evaluate the merits of your actual findings vis-à-vis your original research plan.

b) Run power analyses

Before you run a study, do a power analysis based on earlier research in your own laboratory or published studies with a similar focus, but keeping in mind that these may be biased towards larger effects than what would be realistic (see our discussion above about journal publishing biases). If there are no published or unpublished effect sizes to go by, assume a mean effect size of $r = .21$ (which corresponds to about $d = .43$) as typically found in published social-psychology studies (Richard et al., 2003). If your laboratory is unable to test samples of the sizes suggested by your power analyses, consider engaging in a multi-site collaboration with other labs (for an example, see Knight et al, submitted). Report your a-priori power analysis in your method section, including whatever reasons made you deviate from it.

c) Practice open science

Make your data set, including raw data, and analysis scripts available in commonly accessible formats such as text files (.txt) or generic delimited file formats (e.g., .csv) so that others can reanalyze your data or test their own hypotheses on your data. If you publish open access (OA), which we would recommend, you can submit these files as a supplement with your manuscript. If you submit your work to a subscription-only, paywalled journal, make sure your data and analysis scripts are publicly available through, for instance, the open science framework (osf.io). Be sure to reference these files through a permanent link in your published manuscript. Also, if your research materials and paradigms were programmed on and presented via a computer, make those scripts available, too, so that others can replicate

your work or make use of whatever ingenious idea for testing and assessment you may have developed in your own work.

d) Use the best available methods

With best available methods we refer to the best-validated research designs and measures (see our discussion of strong paradigms above and Schultheiss et al, this volume). Methods and measures of questionable validity can be a means to the end of hidden flexibility, “enabling” findings that would not emerge, or emerge differently, if more valid methods and measures had been used.

e) Run and value replication studies

If you think that your own or someone else’s work in a particular area is so interesting or important that a lot could hinge upon whether the findings can be replicated or not, do a preregistered, properly powered, and methodologically rigorous replication study. If the original study was pretty rigorous to start with, do an exact replication. If it left methodological wriggle room (see our previous point), improve upon it and aim at making the replication as rigorous as possible. Regardless of the study’s results, try to get the replication published and do not be deterred if some journals state that they are not interested in replication studies. Submit and submit again, and you will eventually find a journal with a sufficiently enlightened editorship. If it is not a replication of your own work but someone else’s, confer with the authors of the original study before running the replication study. This can help ensure that you know everything there is to know about the original study and its methods and hence your ability to do a direct replication.

Of course, this also applies in reverse: if someone tries to replicate your own studies and contacts you with a request for more information, your very first response should be to

thank those researchers for deeming your work important enough to warrant the effort, regardless of the outcome, and provide whatever information and materials are requested.

Carefully conducted, methodologically rigorous replication studies may or may not replicate the original finding. Such outcomes should not be misconstrued as verdicts of the scientific integrity of the authors of the original finding or the replicating team. Even when the best methods are used, samples are adequately powered, and all data and materials have been shared openly, neither false positives nor false negatives can ever be conclusively ruled out due to the fundamentally probabilistic nature of statistical prediction in the behavioral sciences. And sometimes a critical moderator may have eluded all contributing author teams so far. Inconsistent results should therefore be the starting point for further, ideally collaborative, replication work and aim at resolving the issue through more studies. They should not be a cause for doubting others' skills and integrity as scientists. Indeed, we think the latter way of thinking is due to our currently prevailing academic publishing practices that place a premium on the outcome of research. We believe that shifting to registered reports will also alter the standards by which scientists will be judged in the future, with a focus then on the kind of questions that are asked and the methods that are employed.

f) Change the standards of evaluation

The recommendations we make here are neither entirely new nor unique. Others have presented similar arguments and recommendations before us, and often more comprehensively and thoughtfully than we could do in the brevity of this chapter (see, for instance, Chambers, 2017; Ioannidis, Fanelli, Dunne, & Goodman, 2015; Munafò et al., 2017; Simmons et al., 2011). But they require, as a consequence, a sea change in how we evaluate science and scientists in the future. We need to stop valuing the churning-out of many publications built on spectacular, but ultimately non-replicable findings from studies with low

statistical power or other methodological drawbacks. Instead, we need to start valuing carefully crafted research that is methodologically rigorous, sufficiently powered, up-front about its goals and methods via preregistration, and transparent with regard to the data collected and analyses performed. We also need to value research programs that focus a substantial part of their effort on internal and external replications and that thus contribute to building a more solid fundament of empirical knowledge. The “new” approach to doing science we espouse here will be more labor-intensive and perhaps yield publications at a slower pace. But it comes with the great advantage that those fewer publications will eventually advance science more than publications generated under the “old” approach. This should and will have important consequences for socializing undergraduate students, graduate students and postdocs into academia, for the standards of good science we communicate to them, and of course for hiring and tenure decisions. Under the “old” standards, committees would screen candidates for number of publications, citation frequency, and acquired external funding. With the new approach we endorse here committees should still look for productivity and relevance, but focus more on the following indicators for their evaluations:

- Are there published registered reports on the candidate’s portfolio? These would signal that other scientists deemed the research important enough to endorse its execution, regardless of the outcome.
- Does the candidate employ rigorous methods, as reflected in preregistration, explicitly reported a-priori power analyses, full disclosure of all experimental conditions, methods, exclusions, and data-analytic decisions?
- Does the candidate make her or his data sets, analysis scripts, and method materials freely available by submitting them as supplements to the publication or depositing

them on internet repositories dedicated to open science practices, to the extent allowed by law?

The more a candidate's portfolio provides affirmative answers to these questions – less so by the time someone is hired for a tenure-track position, more so when the person is up for tenure – the more likely the candidate's research program is built on a solid foundation and will therefore make lasting contributions. We need to change the academic culture in this direction, in science in general and in social neuroendocrinology in particular.

References

- Ainsworth, M. D. S., & Bell, S. M. (1970). Attachment, Exploration, and Separation: Illustrated by the Behavior of One-Year-Olds in a Strange Situation. *Child Development, 41*(1), 49-67. doi:10.2307/1127388
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060
- Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533. doi:10.1038/483531a
- Bem, D. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger (Eds.), *The Compleat Academic: A Career Guide*. Washington, DC: American Psychological Association.
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425. doi:10.1037/a0021524

- Bennett, C. M., Miller, M., & Wolford, G. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Neuroimage*, *47*(Suppl 1), S125.
- Blake, K. R., Dixson, B. J. W., O'Dean, S. M., & Denson, T. F. (2016). Standardized protocols for characterizing women's fertility: A data-driven approach. *Hormones and Behavior*, *81*(Supplement C), 74-83. doi:<https://doi.org/10.1016/j.yhbeh.2016.03.004>
- Blake, K. R., Dixson, B. J. W., O'Dean, S. M., & Denson, T. F. (2017). No compelling positive association between ovarian hormones and wearing red clothing when using multinomial analyses. *Hormones and Behavior*, *90*(Supplement C), 129-135. doi:<https://doi.org/10.1016/j.yhbeh.2017.03.005>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience*, *14*(5), 365-376. doi:10.1038/nrn3475
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436. doi:10.1126/science.aaf0918
- Carney, D. R. (undated). My position on "power poses". Retrieved from http://faculty.haas.berkeley.edu/dana_carney/pdf_Dana%20Carney%20CV%201-23-17.pdf
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*, *21*(10), 1363-1368. doi:10.1177/0956797610383437

- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and Summary of Research on the Embodied Effects of Expansive (vs. Contractive) Nonverbal Displays. *Psychological Science*, 26(5), 657-663. doi:10.1177/0956797614566855
- Carter, C. S., Pournajafi-Nazarloo, H., Kramer, K. M., Ziegler, T. E., White-Traut, R., Bello, D., & Schwertz, D. (2007). Oxytocin: behavioral associations and potential as a salivary biomarker. *Annals of the New York Academy of Sciences*, 1098, 312-322. doi:1098/1/312 [pii] 10.1196/annals.1384.006
- Chambers, C. (2017). *The 7 deadly sins of psychology. A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *J Abnorm Soc Psychol*, 65, 145-153.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Corona, G., Rastrelli, G., Morgentaler, A., Sforza, A., Mannucci, E., & Maggi, M. (2017). Meta-analysis of Results of Testosterone Therapy on Sexual Function Based on International Index of Erectile Function Scores. *European Urology*. doi:<https://doi.org/10.1016/j.eururo.2017.03.032>
- Credé, M., & Phillips, L. A. (2017). Revisiting the Power Pose Effect: How Robust Are the Results Reported by Carney, Cuddy, and Yap (2010) to Data Analytic Decisions? *Social Psychological and Personality Science*, 8(5), 493-499. doi:10.1177/1948550617714584
- Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (in press). P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value For Power-Posing Effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, 0(0), 0956797617746749. doi:10.1177/0956797617746749

- Davis, M. L., Papini, S., Rosenfield, D., Roelofs, K., Kolb, S., Powers, M. B., & Smits, J. A. J. (2017). A randomized controlled study of power posing before public speaking exposure for social anxiety disorder: No evidence for augmentative effects. *Journal of Anxiety Disorders*, 52(Supplement C), 1-7.
doi:<https://doi.org/10.1016/j.janxdis.2017.09.004>
- De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., . . . Feith, S. W. W. (2010). The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans. *Science*, 328(5984), 1408-1411.
doi:10.1126/science.1189047
- Dehaene, S. (2014). *Consciousness and the brain. Discovering how the brain codes our thoughts*. New York, NY: Penguin Books.
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), 355-391.
- Durante, K. M., Rae, A., & Griskevicius, V. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. *Psychological Science*, 24(6), 1007-1016.
doi:10.1177/0956797612466416
- Eisenbruch, A. B., Simmons, Z. L., & Roney, J. R. (2015). Lady in Red: Hormonal Predictors of Women's Clothing Choices. *Psychological Science*, 26(8), 1332-1338.
doi:10.1177/0956797615586403
- Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, 13(6), e1002165.
doi:10.1371/journal.pbio.1002165

- Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4(1), 123-124. doi:10.3758/bf03342209
- Garrison, K. E., Tang, D., & Schmeichel, B. J. (2016). Embodying Power: A Preregistered Replication and Extension of the Power Pose Effect. *Social Psychological and Personality Science*, 7(7), 623-630. doi:10.1177/1948550616652209
- Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641-651. doi:10.1177/1745691614551642
- Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*, 140(5), 1205-1259. doi:10.1037/a0035438
- Goetz, S. M. M., Tang, L., Thomason, M. E., Diamond, M. P., Hariri, A. R., & Carré, J. M. (2014). Testosterone Rapidly Increases Neural Reactivity to Threat in Healthy Men: A Novel Two-Step Pharmacological Challenge Paradigm. *Biological Psychiatry*, 76(4), 324-331. doi:<https://doi.org/10.1016/j.biopsych.2014.01.016>
- Goodman, W. K., Janson, J., & Wolf, J. M. (2017). Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test. *Psychoneuroendocrinology*, 80(Supplement C), 26-35. doi:<https://doi.org/10.1016/j.psyneuen.2017.02.030>
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1), 1-20.
- Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with

- informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology*, 2(1), 123-138. doi:10.1080/23743603.2017.1326760
- Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367-388.
doi:[https://doi.org/10.1016/0167-2681\(82\)90011-7](https://doi.org/10.1016/0167-2681(82)90011-7)
- 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). *Psychological Bulletin*, 140(5), 1260-1264.
doi:10.1037/a0036478
- Harris, R. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope, and wastes billions*. New York, NY: Basic Books.
- Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. New York, NY: Routledge.
- Horvat-Gordon, M., Granger, D. A., Schwartz, E. B., Nelson, V. J., & Kivlighan, K. T. (2005). Oxytocin is not a valid biomarker when measured in saliva by immunoassay. *Physiology and Behavior*, 84(3), 445-448.
- Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., . . . Kendrick, K. M. (2010). Oxytocin Enhances Amygdala-Dependent, Socially Reinforced Learning and Emotional Empathy in Humans. *The Journal of Neuroscience*, 30(14), 4999-5007. doi:10.1523/jneurosci.5538-09.2010
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8), e124. doi:04-PLME-E-0321R2 [pii] 10.1371/journal.pmed.0020124
- Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5), 640-648. doi:10.1097/EDE.0b013e31818131e7

- Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology*, *13*(10), e1002264. doi:10.1371/journal.pbio.1002264
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524-532. doi:10.1177/0956797611430953
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196-217. doi:10.1207/s15327957pspr0203_4
- Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'--a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76-81.
- Knight, E. L., Kutlikova, H. H., Morales, P. J., Christian, C. B., Harbaugh, W. T., Mayr, U., . . . Carré, J. M. (submitted for publication). No robust effect of exogenous testosterone on cognitive reflection in three experiments.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673-676.
- Kudielka, B. M., Hellhammer, D., & Kirschbaum, C. (2007). Ten years of research with the Trier Social Stress Test—revisited. In E. Harmon-Jones & P. Winkielman (Eds.), *Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior* (pp. 56-83). New York, NY: Guilford.
- Kudielka, B. M., Hellhammer, D. H., & Wüst, S. (2009). Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoneuroendocrinology*, *34*(1), 2-18. doi:S0306-4530(08)00264-3 [pii] 10.1016/j.psyneuen.2008.10.004

- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701-710. doi:10.1002/ejsp.2023
- Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a Publication Bias in Behavioural Intranasal Oxytocin Research on Humans? Opening the File Drawer of One Laboratory. *Journal of Neuroendocrinology, 28*(4), n/a-n/a. doi:10.1111/jne.12384
- Lane, A., Luminet, O., Rimé, B., Gross, J. J., de Timary, P., & Mikolajczak, M. (2013). Oxytocin increases willingness to socially share one's emotions. *International Journal of Psychology, 48*(4), 676-681. doi:10.1080/00207594.2012.677540
- Lane, A., Mikolajczak, M., Treinen, E., Samson, D., Corneille, O., de Timary, P., & Luminet, O. (2015). Failed Replication of Oxytocin Effects on Trust: The Envelope Task Case. *PLoS One, 10*(9), e0137000. doi:10.1371/journal.pone.0137000
- LeDoux, J. E. (1996). *The emotional brain*. New York: Simon & Schuster.
- LeDoux, J. E. (2002). *The synaptic self*. New York, NY: Viking.
- Leng, G., & Ludwig, M. (2016). Intranasal Oxytocin: Myths and Delusions. *Biological Psychiatry, 79*(3), 243-250. doi:<https://doi.org/10.1016/j.biopsych.2015.05.003>
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585. doi:10.1126/science.aal3618
- Luminet, O., Grynberg, D., Ruzette, N., & Mikolajczak, M. (2011). Personality-dependent effects of oxytocin: Greater social benefits for high alexithymia scorers. *Biological Psychology, 87*(3), 401-406. doi:<https://doi.org/10.1016/j.biopsycho.2011.05.005>
- Lupien, S. J., & McEwen, B. S. (1997). The acute effects of corticosteroids on cognition: integration of animal and human model studies. *Brain Research Reviews, 24*(1), 1-27. doi:[https://doi.org/10.1016/S0165-0173\(97\)00004-0](https://doi.org/10.1016/S0165-0173(97)00004-0)

- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163-203.
- Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A.-M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General*, *145*(9), e49-e71. doi:10.1037/xge0000200
- McClelland, D. C., Davis, W. N., Kalin, R., & Wanner, E. (1972). *The drinking man*. New York: Free Press.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science*, 103-115.
- Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., Timary, P. d., & Luminet, O. (2010). Oxytocin Makes People Trusting, Not Gullible. *Psychological Science*, *21*(8), 1072-1074. doi:10.1177/0956797610377343
- Mikolajczak, M., Pinon, N., Lane, A., de Timary, P., & Luminet, O. (2010). Oxytocin not only increases trust when money is at stake, but also when confidential information is in the balance. *Biological Psychology*, *85*(1), 182-184.
doi:<https://doi.org/10.1016/j.biopsycho.2010.05.010>
- Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, *244*, 933-938.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. doi:10.1038/s41562-016-0021
- Nelson, R. J. (2011). *An introduction to behavioral endocrinology* (4 ed.). Sunderland MA: Sinauer Associates Inc.

- Open Science, C. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi:10.1126/science.aac4716
- Oxford, J. K., Tiedtke, J. M., Ossmann, A., Özbe, D., & Schultheiss, O. C. (2017). Endocrine and aggressive responses to competition are moderated by contest outcome, gender, individual versus team competition, and implicit motives. *PLoS One*, 12(7), e0181610. doi:10.1371/journal.pone.0181610
- Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*, 10(9), 712. doi:10.1038/nrd3439-c1
- Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, 26(5), 653-656. doi:10.1177/0956797614553946
- Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, 74(4), 331-363. doi:10.1037/1089-2680.7.4.331
- Ronay, R., Tybur, J. M., van Huijstee, D., & Morssinkhof, M. (2017). Embodied power, testosterone, and overconfidence as a causal pathway to risk-taking. *Comprehensive Results in Social Psychology*, 2(1), 28-43. doi:10.1080/23743603.2016.1248081
- Schultheiss, O. C. (2013). Effects of sugarless chewing gum as a stimulant on progesterone, cortisol, and testosterone concentrations assessed in saliva. *International Journal of Psychophysiology*, 87, 111-114. doi:10.1016/j.ijpsycho.2012.11.012

- Schultheiss, O. C., Campbell, K. L., & McClelland, D. C. (1999). Implicit power motivation moderates men's testosterone responses to imagined and real dominance success. *Hormones and Behavior, 36*(3), 234-241.
- Schultheiss, O. C., & Rohde, W. (2002). Implicit power motivation predicts men's testosterone changes and implicit learning in a contest situation. *Hormones and Behavior, 41*, 195-202.
- Schultheiss, O. C., Wirth, M. M., Torges, C. M., Pang, J. S., Villacorta, M. A., & Welsh, K. M. (2005). Effects of implicit power motivation on men's and women's implicit learning and testosterone changes after social victory or defeat. *Journal of Personality and Social Psychology, 88*(1), 174–188.
- Schwabe, L., Tegenthoff, M., Hoffken, O., & Wolf, O. T. (2010). Concurrent glucocorticoid and noradrenergic activity shifts instrumental behavior from goal-directed to habitual control. *Journal of Neuroscience, 30*(24), 8190-8196. doi:10.1523/JNEUROSCI.0734-10.2010
- Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77*, 406-428.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi:0956797611417632 [pii] 10.1177/0956797611417632
- Simmons, J. P., & Simonsohn, U. (2017). Power Posing: P-Curving the Evidence. *Psychological Science, 28*(5), 687-693. doi:10.1177/0956797616658563
- Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology, 38*, 168-172.

- Smith, K. M., & Apicella, C. L. (2017). Winners, losers, and posers: The effect of power poses on testosterone and risk-taking following competition. *Hormones and Behavior*, 92(Supplement C), 172-181. doi:<https://doi.org/10.1016/j.yhbeh.2016.11.003>
- Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, 25(10), 970-975. doi:10.1037/h0029774
- Stanton, S. J. (2011). The essential implications of gender in human behavioral endocrinology studies. *Front Behav Neurosci*, 5, 9. doi:10.3389/fnbeh.2011.00009
- Stanton, S. J., & Schultheiss, O. C. (2007). Basal and dynamic relationships between implicit power motivation and estradiol in women. *Hormones and Behavior*, 52(5), 571-580. doi:S0018-506X(07)00163-8 [pii] 10.1016/j.yhbeh.2007.07.002
- Sternberg, R. J. (2003). *The psychologist's companion: A guide to scientific writing for students and researchers* (4 ed.). Cambridge, MA: Cambridge University Press.
- Tuiten, A., Van Honk, J., Koppeschaar, H., Bernaards, C., Thijssen, J., & Verbaten, R. (2000). Time course of effects of testosterone administration on sexual arousal in women. *Archives of General Psychiatry*, 57(2), 149-153; discussion 155-146.
- Valstad, M., Alvares, G. A., Egknud, M., Matziorinis, A. M., Andreassen, O. A., Westlye, L. T., & Quintana, D. S. (2017). The correlation between central and peripheral oxytocin concentrations: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 78(Supplement C), 117-124. doi:<https://doi.org/10.1016/j.neubiorev.2017.04.017>
- Vongas, J. G., & Al Hajj, R. (2017). The effects of competition and implicit power motive on men's testosterone, emotion recognition, and aggression. *Hormones and Behavior*, 92, 57-71. doi:10.1016/j.yhbeh.2017.04.005

- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4, 274-290.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426-432.
doi:10.1037/a0022790
- Walum, H., Waldman, I. D., & Young, L. J. (2016). Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies. *Biological Psychiatry*, 79(3), 251-257. doi:<https://doi.org/10.1016/j.biopsych.2015.06.016>
- Welker, K. M., Lassetter, B., Brandes, C. M., Prasad, S., Koop, D. R., & Mehta, P. H. (2016). A comparison of salivary testosterone measurement using immunoassays and tandem mass spectrometry. *Psychoneuroendocrinology*, 71, 180-188.
doi:<http://dx.doi.org/10.1016/j.psyneuen.2016.05.022>
- Wiemers, U. S., Schultheiss, O. C., & Wolf, O. T. (2015). Public speaking in front of an unreceptive audience increases implicit power motivation and its endocrine arousal signature. *Hormones and Behavior*, 71, 69-74. doi:10.1016/j.yhbeh.2015.04.007
- Winter, D. G. (1994). *Manual for scoring motive imagery in running text* (4 ed.). Department of Psychology, University of Michigan, Ann Arbor: Unpublished manuscript.
- Wirth, M. M., Welsh, K. M., & Schultheiss, O. C. (2006). Salivary cortisol changes in humans after winning or losing a dominance contest depend on implicit power motivation. *Hormones and Behavior*, 49(3), 346-352.

Table 1

Overview of main methodological features and endocrine outcomes for power-pose studies, with effect sizes based on direct comparisons between high-power and low-power poses

Study	N	Power pose	Additional factors	Effect on testosterone	Effect on cortisol	Also relevant
<i>Original study</i>						
Carney et al (2010)	42 (26♀, 16♂)	2 x 60 s, either high or low	--	$d = 0.66, p < .05$	$d = -0.89, p < .02$	--
<i>Replication studies</i>						
Ranehill et al (2015)	200 (98♀, 102♂)	2 x 180 s, either high or low	--	$d = -0.20, ns$	$d = -0.16, ns$	--
Ronay et al (2017)	108 (64♀, 44♂)	direct replication	--	$d = -0.12, ns$	$d = -0.03, ns$	--
Smith & Apicella (2017)	258 (♂ only), 160 after exclusion of	direct replication, plus neutral posing control	winning/losing in one-on-one tug-of-war (before power posing)	$d = -0.08, ns$	$d = 0.09, ns$	For testosterone, win/lose x power pose interaction: testosterone rise in

	neutral-posing controls				high-power-pose winners and low- power-pose losers
Davis et al (2017)	73 (52♀, 21♂), 53 (36♀, 17♂) after exclusion of no-posing controls	direct replication, plus no-posing control	$d = -0.31, ns$	$d = 0.25, ns$	Participants had diagnosis of social anxiety disorder
<i>Totals and weighted averages</i>	521 (198♀, 323♂)		$d = -0.16$	$d = -0.02$	

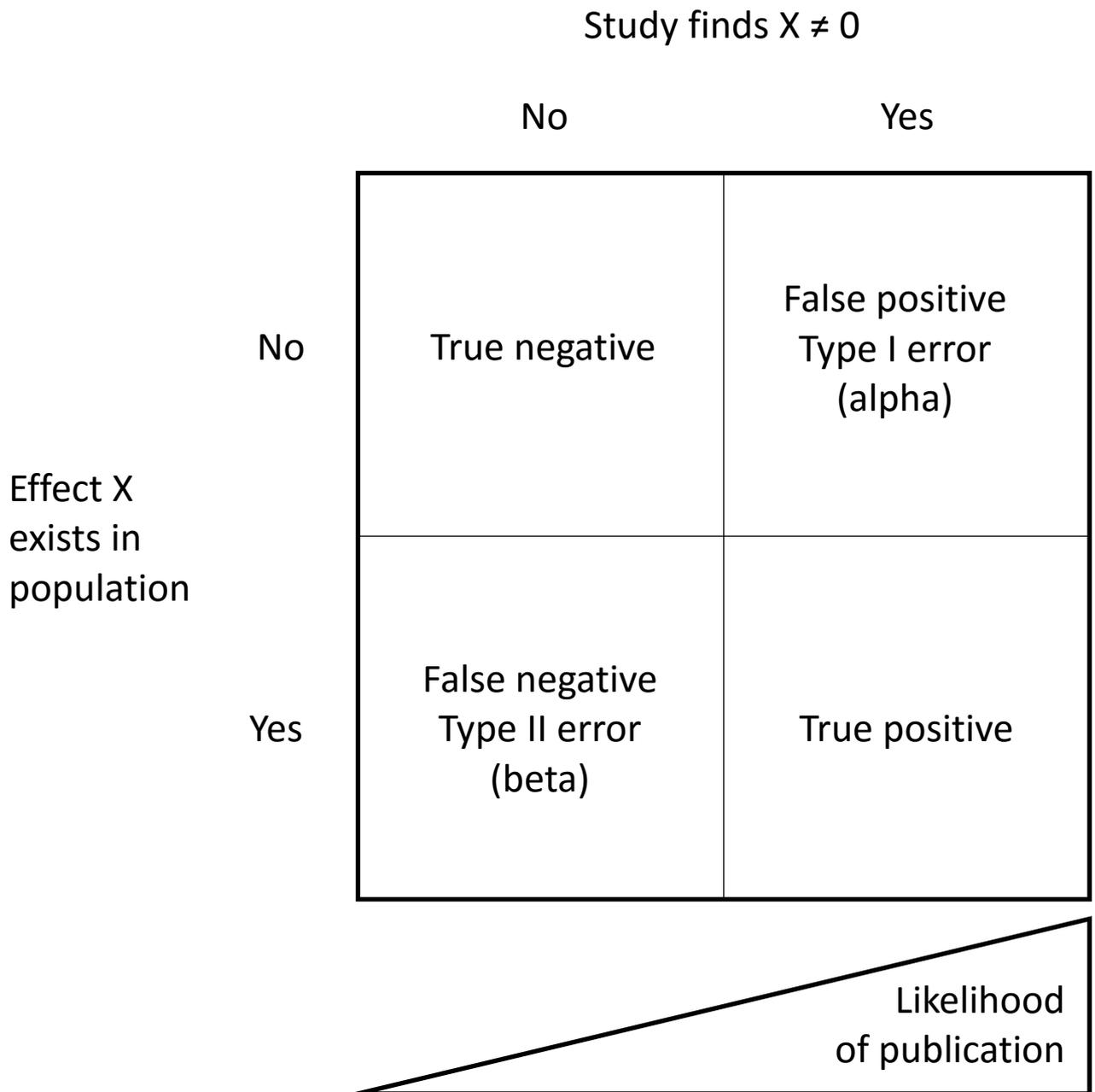


Figure 1. Probabilistic relationship between true effect and effect observed in a given study. Findings from a study may correctly indicate the absence (true negative) or presence (true positive) of an effect exceeding a certain size in a randomly drawn sample. But they may also falsely indicate the absence of an effect if one is present in the population (false negative) or the presence of an effect if one is absent in the population (false positive). Historically, journals in the behavioral sciences have preferred to publish positive (= significant) findings.