


# On Attenuated Interactions, Measurement Error, and Statistical Power: Guidelines for Social and Personality Psychologists

Personality and Social  
Psychology Bulletin  
1–10  
© 2020 by the Society for Personality  
and Social Psychology, Inc  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/0146167220913363  
journals.sagepub.com/home/pspb  


Khandis R. Blake<sup>1,2</sup>  and Steven Gangestad<sup>3</sup>

## Abstract

The replication crisis has seen increased focus on best practice techniques to improve the reliability of scientific findings. What remains elusive to many researchers and is frequently misunderstood is that predictions involving interactions dramatically affect the calculation of statistical power. Using recent papers published in *Personality and Social Psychology Bulletin* (PSPB), we illustrate the pitfalls of improper power estimations in studies where attenuated interactions are predicted. Our investigation shows why even a programmatic series of six studies employing  $2 \times 2$  designs, with samples exceeding  $N = 500$ , can be woefully underpowered to detect genuine effects. We also highlight the importance of accounting for error-prone measures when estimating effect sizes and calculating power, explaining why even positive results can mislead when power is low. We then provide five guidelines for researchers to avoid these pitfalls, including cautioning against the heuristic that a series of underpowered studies approximates the credibility of one well-powered study.

## Keywords

statistical power, effect size, fertility, ovulation, interaction effects

Received August 1, 2019; revision accepted February 22, 2020

Interaction effects, and moderation effects more generally, are common in psychology. Human behavior is notoriously sensitive to context, and testing interactions allows researchers to determine the degree to which the relationship between two variables is generalizable, or specific to certain environments or groups (Mackinnon, 2011). When conducting this research, it is typically insufficient to show that an interaction exists: usually the onus is on showing that the interaction corresponds to a pattern than supports one conclusion over another. Attenuated interactions (also referred to as spreading interactions) are one of these patterns, and characterize situations where the effect of a moderator is to reduce or eliminate, but not reverse, the main effect. Naturally, not all interaction effects are attenuated: when interactions are crossed, moving across a moderator variable results in a reversal of the direction of a main effect. Nonetheless, attenuated interactions are often patterns of interest.

Even in light of the replication crisis, researchers greatly underappreciate the fact that attenuated interactions require surprisingly large sample sizes to achieve adequate statistical power. For example, assume that a study with one between-subject factor with two levels has 80% power to detect a genuine main effect. Assume, as well, that an additional two-level between-subject factor (a moderator) generates a true

fully attenuated interaction effect: at one level of the moderator, the genuine main effect just noted exists; at the other level of the moderator, the main effect is completely absent. For a study to possess 80% power to detect the interaction effect, sample size per cell must be double that of the study examining the main effect in the “non-attenuated” condition alone. As the addition of the moderator also doubles the number of cells in the study (as the design has become a  $2 \times 2$  factorial design rather than a simple two-condition design), the total sample size must increase by a factor of four to retain its efficiency (see also Simonsohn, 2014).

One way to appreciate why power to detect the attenuated interaction is much lower than power to detect the true effect in the “effect-present” condition of the moderator is to understand each effect’s size. If a medium effect size exists in the “effect-present” condition ( $d = .50$ ) and none

<sup>1</sup>UNSW Sydney, Australia

<sup>2</sup>The University of Melbourne, Victoria, Australia

<sup>3</sup>The University of New Mexico, Albuquerque, USA

## Corresponding Author:

Khandis R. Blake, Melbourne School of Psychological Sciences,  
The University of Melbourne, Melbourne, Victoria, Australia 3010.  
Email: khandis.blake@unimelb.edu.au

exists in the other condition, then the effect size of the interaction is half that in the “effect-present” condition (thus,  $d = .25$ ). Naturally, all else equal, smaller true effects can be detected with targeted power only with larger sample sizes—and, indeed, halving of the effect size requires a quadrupling of total sample size to maintain comparable power. Even though these statistical facts have been demonstrated (Aiken et al., 1991; McClelland & Judd, 1993), here we show they are frequently misunderstood by researchers. Our primary aim is to explain these phenomena and guide researchers to avoid the common pitfall of underpowering their attenuated interactions.

The second condition affecting effect size estimates, and thus a priori power calculations, concerns measurement error as it pertains to operationalizing independent variables or predictors of interest. Some independent variables are often very easy to assess accurately, such as group membership, age or biological sex. Others are harder to assess accurately, perhaps because measures are unduly invasive, impractical, or impossible to obtain with 100% accuracy. For these measures, researchers operationalize the variables of interest with proxies, each of which contains a certain degree of measurement error. Current fertility, or ovulatory status, illustrates these latter measures. Ovulatory status can only be confirmed with transvaginal ultrasound (Porterfield, 2001), though it can be estimated with a small degree of error by measuring particular hormone concentrations (e.g., Guermendi et al., 2001), and with larger error, from the current day of a woman’s menstrual cycle. Measurement error naturally diminishes true manifest interaction effect sizes, relative to effects of true “latent” target variables of interest, and hence statistical power. The effect of measurement error on statistical power has also been demonstrated statistically (Busemeyer & Jones, 1983), but we show that effects are frequently underestimated. Our secondary aim is to illustrate the pitfalls of disregarding this type of measurement error and to show how calculation of sample sizes and resultant statistical power can correct for measurement error in predictor variables.

We treat a paper by Netchaeva and Kouchaki (2018), recently published in *Personality and Social Psychology Bulletin* (PSPB), as our first case study to demonstrate these pitfalls and how they may be overcome, later extending our analysis of the first pitfall to a range of papers in PSPB. Netchaeva and Kouchaki (2018) presented an ambitious series of six studies, requiring a high degree of investment from the researchers. There is much to like about their paper, including multiple studies and attempted internal replication of effects, the utilization of multiple methods to assess proxy variables (i.e., fertility), varying sample pools to increase generalizability, and a final large study that was preregistered. In short, the project incorporated many strong research practices. Nonetheless, critical problems limit what can be concluded from the series of studies, the most notable of which concern an

underappreciation of true manifest effect sizes that could have reasonably been detected, and hence overestimation of statistical power.

As we document here, none of the six studies reported—including one with 537 participants—had anywhere close to sufficient statistical power to detect an attenuated interaction in which, within a predicted “effect-present” condition, there exists a true medium effect. The authors concluded that no attenuated interaction effect likely exists yet, in fact, their data do not support this conclusion. This article, then, offers a good case study illustrating how a well-intentioned and ambitious research program incorporating a host of strong research practices can nevertheless be hampered by misunderstandings of attenuated interactions and measurement error. By clarifying these misunderstandings, we hope that our discussion of this study can help prevent similar issues arising in future research.

### **Attenuated Interactions Affect Effect Size Estimates: A Case Study**

Netchaeva and Kouchaki (2018) sought to test whether naturally ovulating women, when conceptive in their cycles, would especially distrust, dislike, derogate, and be more attuned to the dominance of a woman dressed in red, relative to a woman dressed in another color (in their studies, blue and gray). Six studies tested their hypotheses, exposing purportedly conceptive or non-conceptive women to a woman dressed in red or blue/gray, then measuring interpersonal judgments. The prediction tested was an attenuated interaction: when women evaluated a target woman dressed in red, conceptive status was predicted to have main effects on interpersonal judgments, but when women evaluated a woman dressed in blue or gray, conceptive status was expected to have little to no effect on interpersonal judgments. Findings were mixed, with about one of six effects tested achieving statistical significance in the predicted direction ( $p < .05$ ), and a little over one in three achieving “marginal” significance ( $p < .10$ ). Ultimately, the authors concluded support for the null hypothesis.

Netchaeva and Kouchaki’s (2018) hypothesis was a risky one, as it stems from the conjunction of two separable ideas: that women are more attuned to compete intrasexually with female rivals in a conceptive state, and that the color red evokes both sexual motivation in men and competitive motivation in women (for critical treatments of the latter idea, see Francis, 2013; Lehmann & Calin-Jageman, 2017). The authors furthermore posited that conceptive women’s competitive motivations evoked by redness would manifest in a variety of forms: fundamentally, they focused on distrust, but they proposed that distrust could manifest in dislike, derogation tactics, and attunement to dominance. Our aim is not to weigh in on the a priori strength of these hypotheses. Rather, we show that even if their hypotheses are true, a misunderstanding of statistical power in attenuated interactions, and

of measurement error, meant that their six studies had little chance of detecting the effects in question.

### Calculating the Expected True Effect Size of an Attenuated Interaction

First, we examine how attenuated interactions affect the expected true size of an effect. A medium effect size (Cohen's  $d$ ) is generally considered to be  $d = .50$  (Cohen, 1988). Let us suppose that women in the fertile window (generally thought to be approximately 5–6 days per cycle; Fehring & Schneider, 2008) trust women in red less than women outside the fertile window, with  $d = .50$ . We furthermore assume that they equally trust women in blue or gray, such that  $d = 0$ . This pattern constitutes a fully attenuated interaction: an effect present in one condition completely evaporates in the other condition. The effect size of the interaction—the standardized difference in the dependent variable along diagonals of a  $2 \times 2$  matrix—is thus half that of the “effect-present” condition effect size (thus  $d = .25$ ). So long as sample size in the two groupings compared are equal, as should (approximately) be the case in this instance, we can express this effect size in a metric equivalent to  $r$  in Equation 1, with conversion of  $d$  to  $r$  such that when  $d = .25$ :

$$r = \frac{d}{\sqrt{d^2 + 4}} \quad (1)$$

$$r = .124 \quad (2)$$

As shown in Equations 1 and 2, assuming a true effect size,  $d$ , in the “woman-in-red” condition at  $d = .50$  and the existence of an attenuated interaction, the true interaction effect size expected by the design of Netchaeva and Kouchaki (2018) is  $r = .124$ . Sample sizes calculated to yield, say, 80% power to detect the true effect, based on the incorrect effect size estimate of  $d = .50$ , will at best result in a study with one-quarter the required sample size to achieve adequate statistical power.<sup>1</sup>

Although we focus on factorial designs, we note that the situation is not necessarily improved when one of the moderators is continuous. Simulation studies have demonstrated that in comparison with a  $2 \times 2$  design, normally distributed continuous moderators can result in an eight-fold decrease in statistical power (McClelland & Judd, 1993). This decrease eventuates because the ability of a continuous moderator to detect a genuine effect depends on the variance in the distribution of the moderator and its co-occurrence with the other moderator. In short, the greatest statistical power results from a large number of co-occurring values in the tail ends of each moderator's distribution. Although further focus on these subtleties is outside the scope of this article, McClelland and Judd (1993) provide a compelling explanation and should be consulted by interested readers.

The primary point of this section is worth repeating: although a sample size may be adequate for detecting main

effects, it may be grossly inadequate for detecting moderator effects of similar importance. The situation worsens when one deals with measurement error, which, in Netchaeva and Kouchaki's (2018) case, pertained to conceptive status. In what follows, we show that when accounting for measurement error, the expected true effect size diminishes even further.

### Adjusting the Expected True Effect Size for Measurement Error

In five of six studies, Netchaeva and Kouchaki (2018) estimated current fertility using two different variations of counting methods, both requiring women to estimate their typical menstrual cycle length. The first variation (“Method 1”) used this length estimate in conjunction with counting backward from the last day of the cycle to estimate women's current cycle day, classifying those women sampled on Days 8 to 14 (a 7-day window) into the high fertility group, and the remainder to the low fertility group. The second variation (“Method 2”) used the cycle length estimate in conjunction with counting forward from the first day of the cycle, and actuarial data from Wilcox et al. (2001), to provide a continuous estimate of conception probability based on the forward-counted current cycle day. How do these two methods stack up, in terms of measurement error? In addition to recalled menstrual cycle length being associated with around 21% error in and of itself (Small et al., 2007), both Method 1 and Method 2 are independently fraught with considerable error and low validities in their own right.

Using luteinizing hormone (LH) tests to indicate fertility, Blake et al. (2016) provide one estimate of measurement error for both methods. When using a 6-day fertile window, Blake et al. (2016) show that Method 1 classifies women into the “fertile” group with an error rate of 39.5%; when using the 7-day window from Netchaeva and Kouchaki (2018), the error rate increases to 45.2%. For Method 2, Blake et al. (2016) show that this method yields an error rate of 59.4%. Subsequent work supports modest degrees of measurement error for counting methods (Arslan et al., 2018), though they yield less dire rates of error than Blake et al. (2016), with an upper bound of around 30%. Taking all of the available evidence into account, it is reasonable to say that at least one—and possibly up to two—of every three subjects from the “fertile” group in Netchaeva and Kouchaki (2018) were very likely to be non-fertile at the time of measurement.

These error rates are sizable but do not present an insurmountable problem, assuming researchers account for reduced validity when conducting a priori sample size calculations. Gangestad et al. (2016) use simulated data to show how to do so, and we use their results to provide a second estimate of the validity of methods employed by Netchaeva and Kouchaki (2018). For Method 1—the binary 7-day fertile window estimate—the correlation between true fertility status and measured fertility status in Gangestad et al. (2016)

is very close to  $r = .50$ . This estimate assumes a 7-day window covering Days 10 to 16, but we note that Netchaeva and Kouchaki's (2018) 7-day window has lower validity than this window, and we approximate it to be  $r = .45$ . For Method 2—the continuous estimate based on forward-counted conception probability—the validity in Gangestad et al. (2016) is close to  $r = .50$ . Given Netchaeva and Kouchaki (2018) assigned the first day of the cycle to be Day 0, not Day 1, however, again this leads to a slightly less valid measure (also approximately  $r = .45$ ).

These validity estimates have large effects on the true manifest interaction effect size. Assuming validity estimates of  $r = .45$ , the true manifest interaction effect size for Studies 2 to 6 in Netchaeva and Kouchaki (2018) under assumptions stated above, in  $r$ , is exceptionally small and shown in Equation 4:

$$r = .124 \times .45 \quad (3)$$

$$r = .059 \quad (4)$$

Thus, taking into account measurement error, the sample size estimated to yield 80% power is now at least 16 times larger than that estimated to yield 80% power assuming  $d = .50$ . The one exception is Study 1, which used LH tests to measure fertility status. As acknowledged by the researchers, LH tests are the most validated and least invasive method for estimating ovulation (Guida et al., 1999). If we assume that validity of that method is  $r = .9$ , then the manifest interaction effect size in that study, in  $r$ , is somewhat larger and shown in Equation 6:

$$r = .124 \times .9 \quad (5)$$

$$r = .111 \quad (6)$$

This still results in a sample size at least four times as large as that required when  $d = .50$ , but is nevertheless a big improvement on the sample required when using counting methods with such high rates of error.

In fairness to Netchaeva and Kouchaki (2018), they conducted power analyses based on published effect sizes. Hence, for instance, for Study 2, they calculated the necessary sample size to achieve 80% power to detect an effect size of  $f = .2$  (i.e.,  $r = \sim .2$ ), as reported in a previous study by Durante et al. (2014), resulting in targeted sample size of 200. As Durante et al. used a measure of fertility status similar to Netchaeva and Kouchaki, effect sizes in their study should have also been affected by poor measurement validity. Yet, Durante et al. did not examine an attenuated interaction, meaning the effect size that could be expected in Netchaeva and Kouchaki's studies was lower than what Durante et al. could expect. Furthermore, the true effect size of fertility status that a manifest effect size of  $r = .2$  assumes (i.e., disattenuated for measurement error) is closer to  $r = .44$  (i.e.,  $.2/.45$ )—an unrealistically large effect size. Due to publication bias, published effect sizes tend to overestimate true effect sizes, even when the latter are non-zero. Hence, we suggest, researchers should base power calculations on realistic effect sizes, partly in light of measurement error, and not

merely take published effect sizes at face value. We also note that, while Netchaeva and Kouchaki achieved sample sizes close to or greater than 200 in Studies 2, 5, and 6, sample size was 129 in Studies 3 and 4, and hence fell short of the sample size their own power estimate called for.

## Improper True Effect Size Calculations Dramatically Reduce Power

It is probably not too surprising for most researchers that improperly calculating true effect sizes can dramatically affect estimated statistical power. What we suspect is less intuitive to researchers is the extent to which power can be reduced, and how large sample sizes may need to achieve adequate power. Sample sizes in Netchaeva and Kouchaki (2018) ranged from  $N = 129$  to  $N = 537$ , so while they were not impressively large, neither were they intuitively much too small for a  $2 \times 2$  design. Researchers may ask: surely sample sizes in a  $2 \times 2$  design that exceed 500 approach adequate power, or at the very least, effects in the aggregate do? Here, we show that this intuition—that sample sizes around  $N = 500$  in  $2 \times 2$  designs are fairly defensible—is not the case. In fact, even aggregate sample sizes closer to  $N = 1,000$  can be severely underpowered to detect meaningful effects.

To examine the degree to which this intuition is misleading, we take our estimates of true manifest interaction effect size and sample sizes from Netchaeva and Kouchaki (2018) and calculate the power obtained in their studies to detect a pattern of meaningful effects. We assume a meaningful effect is of medium size ( $d = .50$ ) in the “woman in red” condition, and of zero size ( $d = 0$ ) in the “woman in blue/gray” condition. Table 1 documents our findings, which shows that the actual power for the individual studies in Netchaeva and Kouchaki (2018) ranges from just under 10% (when  $N = 129$ ; Studies 3 and 4) to about 25% (when  $N = 537$ ; Study 6). Even with a sample size of over 500, then, the power to detect a real substantive effect is seriously inadequate. Indeed, power to achieve even a “marginally significant” result ( $p < .10$ , two-tailed) ranges from about 17% to 36%. In even the most highly powered study, where  $N = 537$ , the probability of achieving even a marginally significant result is just over one in three. On average, across the individual studies, power to obtain significant and marginally significant results is 14% and 22%, respectively, and a study with 80% power would have required a total sample size of about  $N = 2,500$ . Power estimates are similarly inadequate in the aggregate where total  $N$ s =  $\sim 1,000$ : as shown in Table 2, power to detect true medium-sized effects in the aggregate data ranges from about 35% to 42%.

### On the Pitfalls of Interpreting Results When Power Is Inadequate

Netchaeva and Kouchaki (2018) conclude that, “Across six studies, our research fails to provide support for the prediction

**Table 1.** Estimates of Power to Detect Medium Attenuated Interaction Effect Sizes From Netchaeva and Kouchaki (2018).

Study	N	Power	
		Two-tailed, $p = .05$	Two-tailed, $p = .10$
Study 1	66	.145	.233
Study 2	209	.126	.207
Study 3	129	.096	.166
Study 4	129	.096	.166
Study 5	192	.120	.198
Study 6	537	.253	.363
Total N	1,262	Mean power	.139

that an ovulating woman is less likely to trust another woman wearing red compared with a nonovulating woman” (p. 1). Yet, based on the power estimates from their individual studies, mostly non-significant results were to be expected. Specifically, about one in seven of the results should have been statistically significant, and one in four should have been marginally significant (if true effects were medium in size). In fact, across the two different fertility estimation methods and 20 different specific outcomes examined in their six studies (a total of 39 effects), the number of effects with  $p < .05$  (6/39; 15%) and effects with  $p < .10$  observed (14/39; 36%) equals or exceeds the numbers that could have been expected based on these power estimates.

This pattern is also evident in the  $p$ -values associated with the effect sizes when data are analyzed in the aggregate. The mean interaction effect for the primary effect of interest—trust of the target women—is significant, with  $p = .008$  and  $p = .004$  for the continuous and binary measures of fertility status, respectively. Other effects are significant, in aggregate, for prosocial giving toward the target ( $p = .004$  and  $p = .054$ ) and perceived warmth of the target ( $p = .119$  and  $p = .002$ ). Tests for other outcome variables vary (e.g., they approach significance or are significant for perceived dominance of the target [ $p = .119$  and  $p = .048$ ] and perceived attractiveness [ $p = .084$  and  $p = .051$ ]). Thus, of 14 different tests, five  $p$ -values—or 36% of tests—were less than .05 and eight  $p$ -values—57% of tests—were less than .10, values greater than expectation if true effects are medium in size (30% and 42%, respectively).

We can also use the observed manifest mean effect sizes to estimate the true interaction effect size in  $d$ . Once again,  $d$  for a medium attenuated interaction effect is equal to .25. If validity of measurement is assumed to be .45, the mean estimated interaction effects across all outcomes are  $d = .33$  and  $d = .36$  (using continuous and binary fertility status measures, respectively). The mean effects for the primary outcome of interest, trust in the target woman, are  $d = .39$  and  $d = .43$ . More generally,  $d$  is above .3 for prosocial responses possibly related to trust (perceptions of warmth, prosocial

giving, and liking), though lower ( $< .3$ ) for other perceptions (attractiveness, competence, and dominance).

In aggregate, these data appear to be compatible with larger-than-medium effect sizes, at least in some domains. Nonetheless, our point is *not* that meaningful true effects do exist; further investigation is needed to assess their size and robustness. Rather, our primary point is that, in light of weak power to detect true effects, Netchaeva and Kouchaki’s (2018) conclusion that there is scant evidence for non-null effects is potentially misleading. Within their data, predictions performed as well as could be expected if true medium-sized effects exist, such that the conclusion that the research failed to find support for them is not warranted. The conclusion only appears reasonable because the weak power in Netchaeva and Kouchaki’s studies is vastly underappreciated.

### Tests of Attenuated Interactions Are Common and Typically Underpowered

We have used Netchaeva and Kouchaki (2018) as our case study to illustrate issues of power in tests of attenuated interactions, but the problem is not confined to this one example. Tests of attenuated two-way interactions are commonly used by many researchers in social and personality psychology, and issues with statistical power often result. To demonstrate, we used Web of Knowledge to identify all papers published in PSPB in 2019, including online. Of 147 papers, we found 12 that explicitly sought to test attenuated interactions.<sup>2</sup> For instance, Nelson-Coffey et al. (2019) predicted that parenthood enhances men’s well-being more than women’s well-being. Wang and Ackerman (2019) predicted that infectious disease primes would increase the impact of individual differences in germ aversion on perceptions of social crowdedness. Townsend et al. (2019) examined whether an intervention designed to educate students about the college experience would close the gap in achievement across social classes. Across these papers, 34 different studies examined attenuated interactions.

To gauge the statistical power of these studies, we make a few simple assumptions. Namely, we assume (a) a fully attenuated interaction, with a medium true effect ( $d = .50$ ) in an “effect-present” condition and no effect ( $d = .00$ ) in an “effect-absent” condition, (b) a balanced research design, and (c) measurement validity at .80. Under these assumptions, the effect size of the interaction (represented in  $r$  or  $f$ ) is .10, and 80% power to detect it with a two-tailed test requires a sample size of  $N = 781$  (where  $\alpha = .05$  and there are no additional covariates; G\*power; Faul et al., 2009). As shown in Table 3, the actual sample size across the 34 studies ranged from  $N = 62$  to  $N = 13,007$  (median = 224, harmonic  $M = 210$ ), most of them (29 studies; 85%) fell below 781, and five sample sizes were  $N = 2,452$  or larger. Of the 29 studies failing to meet the criterion of  $N \geq 781$ , the median sample size was 206, the harmonic mean was 180,

**Table 2.** Significance Tests and Estimated Effect Sizes From Aggregate Data From Netchaeva and Kouchaki (2018).

Outcome	Measure of fertility status	N	Mean $z_r$	$z$	$p$	Mean $wtr$	Estimated $d$
Trust	Continuous	795	-0.136	-2.66	.008	-0.111	-.39
	Binary	795	-0.148	-2.89	.004	-0.121	-.43
Prosocial giving	Continuous	209	-0.202	-2.88	.004	-0.199	-.66
	Binary	209	-0.138	-1.93	.054	-0.202	-.44
Liking	Continuous	258	-0.078	-1.22	.222	-0.077	-.33
	Binary	258	-0.094	-1.43	.140	-0.093	-.40
Perceived warmth	Continuous	858	-0.064	-1.56	.119	-0.069	-.30
	Binary	858	-0.127	-3.08	.002	-0.113	-.47
Perceived attractiveness	Continuous	858	-0.071	-1.73	.084	-0.064	-.28
	Binary	858	-0.08	-1.95	.051	-0.069	-.30
Perceived dominance	Continuous	987	0.048	1.56	.119	0.041	.18
	Binary	987	0.063	2.04	.048	0.051	.22
Perceived competence	Continuous	858	-0.039	-0.95	.344	-0.044	-.19
	Binary	858	-0.053	-1.29	.198	-0.057	-.25

and the range was 62–670. This median sample size yields just 30% power under the stated assumptions.

All but three studies (31; 91%) report at least one interaction  $p$ -value as “significant” (26 of the 29 studies [90%] with sample size < 781), yet an examination of the distribution of reported  $p$ -values leads to the conclusion that mean power in these studies falls far short of 80%. If 80% power were achieved, then 70% of significant  $p$ -values should be  $p < .01$ . Yet, only 10 of the 31 studies yielded  $p < .01$  (and when the five large studies were excluded, just six studies yielded  $p < .01$ ). For these 31 studies, a  $p$ -curve calculated through  $p$ -curve app (Simonsohn, 2017) yielded a modest estimated power of 46% (95% CI = [22, 68]) across all studies, and a stunningly low power of 9% (95% CI = [5, 28]) across the 26 studies with sample size less than 781. Indeed, the distribution of these 26  $p$ -values is not significantly different from what one would expect if all effects were null ( $p = .17$ ).<sup>3</sup>

Many of these studies were underpowered despite authors conducting and reporting power analyses. Of the nine papers exclusively reporting studies with sample sizes less than 781, seven calculated a priori power analyses. Of these, four studies estimated power based on effect sizes calculated on a small pilot study ( $r = .26$ ; Wang & Ackerman, 2019), or effect sizes that were presumed to be medium to large ( $r = .32$ ; Carrier et al., 2019), medium (.25; Yao & Chao, 2019), or small to medium (.20; Martin et al., 2019). As these assumed *interaction* effect sizes imply true simple effects in an effect-present condition that is double the assumed interaction effect size (i.e.,  $r = .40$ –.64), the hypothesized interaction effect sizes were overly optimistic. Another two papers (Eck et al., 2019; Townsend et al., 2019) used prior data pertaining to effect size in an effect-present condition where  $d = .35$  or  $d = .70$ , respectively, but did not account for the effect size of a fully attenuated interaction effect halving that in the effect-present

condition. Just one paper reported a power analysis based on a putative true effect within an effect-present condition and a zero effect within an effect-absent condition (Voelkel & Brandt, 2019), leading to a required sample size that was double that of the other six papers.

We emphasize that we are *not* saying that the attenuated interactions reported in these studies do not exist. Likewise, we emphasize that  $p$ -values in all studies with large sample sizes were close to or less than  $p = .01$  (Bahamondes et al., 2019; Nelson-Coffey et al., 2019; Sparks & Ledgerwood, 2019). In two of the papers, authors also conducted internal meta-analyses across multiple studies, which yielded evidence that suggests effects are real (Hasan-Aslih et al., 2019; Wang & Ackerman, 2019). Rather, we draw on these examples to illustrate that the issue of inadequate power in attenuated interactions is not just confined to our case study of Netchaeva and Kouchaki (2018), but common across personality and social psychology—even when power analyses are reported.

## Discussion

The first lesson that researchers can draw from our analyses is to beware effect size calculations for attenuated interactions. In particular, it is critical to be mindful that, if a meaningful main effect size in an “effect-present” condition is a medium  $d = .50$ , a meaningful effect size for a fully attenuated interaction is  $d = .25$ . When a moderating variable only partially attenuates the effect, even larger sample sizes are required. To detect an attenuated interaction, then, a study requires at the very least twice as many participants per cell to achieve the same statistical power as a study designed to detect a main effect in an “effect-present” condition.

Second, it is important to remember that mostly non-significant results are to be expected when studies are underpowered, and these results are not especially meaningful.

**Table 3.** Sample Characteristics and *t*-Statistic Summaries of PSPB 2019 Papers Reporting Attenuated Interactions.

Paper	Study	<i>N</i>	<i>df</i>	<i>t</i>	<i>p</i>
Wang and Ackerman (2019)	Pilot	62	58	2.09	.041
	Study 1	100	96	2.68	.009
	Study 2	206	202	2.24	.026
	Study 3	358	353	1.61	.108
	Study 4	222	218	2.06	.041
	Study 5	353	349	<sup>a</sup>	<sup>a</sup>
Yao and Chao (2019)	Study 1	195	191	2.42	.016
	Study 2	153	148	2.02	.045
Hasan-aslih et al. (2019)	Study 1a	152	148	2.58	.011
	Study 1b	153	149	1.54	.126
	Study 1c	225	221	2.00	.047
	Study 2	276	272	2.73	.007
Martin et al. (2019)	Study 2	148	144	2.15	.033
	Study 3	165	161	2.37	.019
	Study 4	212	204	2.46	.015
	Study 5	235	231	2.14	.033
	Study 6	156	152	2.57	.011
Adelman and Dasgupta (2019)	Study 1	392	388	2.13	.034
	Study 2	670	666	3.08	.002
	Study 3	551	543	2.29	.022
Voelkel and Brandt (2019)	Study 1	542	534	2.85	.005
	Study 2	416	410	2.17	.031
Townsend et al. (2019)	Pilot	124	116	2.74	.007
	Intervention	133	115	2.24	.027
Sparks and Ledgerwood (2019)	Study	2,452	2,439	2.53	.011
Eck et al. (2019)	Study 2	131	124	2.03	.044
	Study 3	414	398	2.44	.015
Bahamondes et al. (2019)	Study 1a	12,959	12,943	3.14	.002
	Study 1b	12,859	12,843	10.15	<.001 <sup>b</sup>
Nelson-Coffey et al. (2019)	Study 1a	13,007	13,003	4.00	<.001 <sup>c</sup>
	Study 1b	472	468	2.33	.020
	Study 2	4,930	4,926	8.03	<.001 <sup>d</sup>
Carrier et al. (2019)	Study 2a	106	102	4.50	<.001 <sup>e</sup>
	Study 2b	88	84	2.09	.040
All studies ( <i>N</i> = 34)	Mean	1,577	1,571		.020
	Median	224	220		
	Harmonic mean	210	202		
Underpowered studies ( <i>N</i> = 29)	Mean	256	250		.023
	Median	206	202		
	Harmonic mean	180	173		

Note. All reported statistics were transformed to *t*, necessary to allow ease of computation of *p*. Where moderation interaction effects were examined but *df* were not reported (e.g., using Hayes' [2012] process macro), we calculated *t* by dividing the reported effect by its SE. <sup>a</sup>No *t*-value or other test statistic was reported; *p* > .25. <sup>b</sup>*p* = 4.09E–24. <sup>c</sup>*p* = 6.37E–05. <sup>d</sup>*p* = 1.21E–15. <sup>e</sup>*p* = 1.81E–05.

Treating them as such can give the impression that an effect is absent, or that there is genuine inconsistency between studies, when those yielding null results may simply reflect a lack of statistical sensitivity (Vadillo et al., 2016). Likewise, when manifest observed effect sizes yield point estimates of highly meaningful true effects and yet significant tests yield mixed results, the conclusion should not be that true effects are likely near-null. Rather, in such instances, mixed significance tests may very well reflect

weak power to detect meaningful true effects, and demonstrably so in our case study.

Third, a series of underpowered studies, while perhaps creating a veneer of reliability and coherence, do not provide the same evidence of an effect as one adequately powered study. This point has been convincingly made already (Schimmack, 2012), but its uptake has been disappointingly slow, especially in personality and social psychology where some of the top journals require multi-study papers.

It is likely that non-significant results will eventuate from multiple small, separate studies with low power (Maxwell, 2004), but it is quite possible that these effects are “significant” (i.e., in the aggregate, unlikely under the null hypothesis), as we found with Netchaeva and Kouchaki (2018). Designing a single, well-powered study will almost always exceed the evidentiary value of a series of small, underpowered studies, even if the latter bolsters its credibility by employing varying procedures and samples. Alternatively, an analysis that combines results across studies may yield a conclusion that differs from analyses of individual studies (Scheibehenne et al., 2016).

Fourth, when power is low, even positive results can be misleading. Studies that do detect effects will tend to overestimate their size, and in conjunction with publication bias, the published literature will tend to overestimate true effect sizes. In addition, when power is low, positive results are less compelling evidence for true effects, as the proportion of positive effects that result when true effects exist—as opposed to when no effects exist—is smaller, all else equal (e.g., Christley, 2010). Power analyses based on the effect sizes of past studies should account for these statistical facts, and may do well to estimate effect sizes conservatively rather than optimistically.

Fifth, researchers should keep in mind that measurement error of independent or predictor variables affects statistical power. Just as *p*-hacked studies with inflated alpha criteria can litter a field with false claims, so too can studies with excessive Type II errors offering null conclusions. Although much attention is paid to the harmful consequences of elevated Type I errors, underpowered studies are a major contributing factor to false positives and false negatives (Button et al., 2013). There are several reasons to suspect that deceptive null results are just as harmful to scientific progress as unreliable positive findings (Fiedler et al., 2012; Vadillo et al., 2016). Although the replication crisis has led scholars to focus on the latter, the former too should be avoided.

For the field examining shifts across the ovulatory cycle specifically, progress depends on careful attention to estimating fertility using methods that meet a minimum acceptable criterion for validity. Estimating fertility through LH tests is the most cost-effective, well-validated, and reliable method, and Blake et al. (2016) provide a protocol for utilizing this method in research designs. However, even this method has important limitations that must be considered in study designs (Blake, 2018; Roney, 2018), and researchers are advised to choose a method best suited to answer their research question. Steroid hormones (notably estradiol and progesterone) may well be the primary signals that mediate psychological shifts across the cycle, and for this reason, it may be advisable to directly examine associations with steroid hormone levels. Sample sizes determined by properly informed power analyses that account for error in the chosen fertility estimation method, and estimate true effect sizes accordingly, are a critical step forward (Gangestad et al., 2016).

Finally, we join calls made by other researchers to design studies that yield informative results about the presence—and absence—of meaningful effects (Amrhein et al., 2019; Funder et al., 2014; Lakens et al., 2018). Null hypothesis significance testing by itself can never provide information whether meaningful effects are absent, instead allowing researchers only to reject the null hypothesis (e.g., Rogers et al., 1993). Concluding that a meaningful effect is absent requires quantifying what a meaningful effect looks like (Anderson & Maxwell, 2016). Equivalence testing (Lakens, 2017; Schuirmann, 1987), inference by confidence intervals (Amrhein et al., 2019; Westlake, 1972), and Bayesian statistics (Jeffreys, 1939; Maxwell et al., 2015) all allow the absence of meaningful effects to be detected and can be used to inform the minimum sample size required to obtain sufficient statistical evidence.

## Conclusion

Researchers are generally aware of the pitfalls of low power and of measurement error. Yet, they are sometimes insufficiently aware that the true effect sizes of meaningful attenuated interactions appear relatively small and hence require substantial sample sizes to detect. Researchers may also neglect the considerable impact that measurement error can have on manifest sizes and, as a consequence, statistical power to detect a meaningful true effect. Incorrectly estimated sample sizes needed to detect meaningful effects in these designs can dramatically weaken power. By following some simple recommendations, researchers can avoid these pitfalls.


## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Khandis R. Blake  <https://orcid.org/0000-0003-4834-4120>

## Notes

1. A reviewer rightfully noted that although our point holds when one tests an interaction with traditional contrasts (comparing cells along the two diagonals in a  $2 \times 2$  factorial design), one could assess an attenuated interaction with a +3 (assigned to the cell expected to differ from the other three) versus -1 (assigned to each of the other cells) contrast. This contrast is equally confounded with the two main effects and the interaction contrast in a traditional design and, for that reason, is typically considered inadequate to test whether effects of a variable in one condition on a moderating variable differs from the effects of the variable in the other condition of the moderating variable. If the two main



effects are sufficiently strong, the 3 versus -1 contrast could be non-zero even in total absence of any moderation effect, attenuated or otherwise. If one does use this contrast, however, true effect size for the contrast is *not* halved relative to the effect in, say, the “woman in red” condition as stated here; it is equal to that effect. Netchaeva and Kouchaki (2018) used a traditional interaction contrast, as is typical.

2. Of course, additional studies examined interactions predicted to be cross-over interactions, and some may have examined attenuated interactions that were not a primary focus apparent to us.
3. Some studies reported more than one  $p < .05$ . In such cases, we took just the first significant interaction effect reported. We note that effect sizes estimated from  $p$ -curve assume a homogeneous true effect across studies, which may bias  $p$ -curve effect size estimation (McShane et al., 2016); the same criticism may apply to power estimates. Still, one can refrain from placing weight on the *exact* quantitative power estimates from  $p$ -curve while appreciating the qualitative conclusion that, based on a  $p$ -value distribution, these studies are, on average, woefully underpowered to detect true effects.

## References

- Adelman, L., & Dasgupta, N. (2019). Effect of threat and social identity on reactions to ingroup criticism: Defensiveness, openness, and a remedy. *Personality and Social Psychology Bulletin*, *45*(5), 740–753.
- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions* [Leona S. Aiken, Stephen G. West with contributions by Raymond R. Reno]. SAGE.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature* *567*(7748), 305–307.
- Anderson, S. F., & Maxwell, S. E. (2016). There’s more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12.
- Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2018). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*. Advance online publication. <https://doi.org/10.1037/pspp0000208>
- Bahamondes, J., Sibley, C. G., & Osborne, D. (2019). “We look (and feel) better through system-justifying lenses”: System-justifying beliefs attenuate the well-being gap between the advantaged and disadvantaged by reducing perceptions of discrimination. *Personality and Social Psychology Bulletin*, *45*(9), 1391–1408.
- Blake, K. R. (2018). Resolving speculations of methodological inadequacies in the standardized protocol for characterizing women’s fertility: Comment on Lobmaier and Bachofner (2018). *Hormones and Behavior*, *106*, A4–A6.
- Blake, K. R., Dixon, 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*, 74–83.
- Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*(3), 549–562.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafó, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.
- Carrier, A., Dompnier, B., & Yzerbyt, V. (2019). Of nice and mean: The personal relevance of others’ competence drives perceptions of warmth. *Personality and Social Psychology Bulletin*, *45*(11), 1549–1562.
- Christley, R. (2010). Power and error: Increased risk of false positive results when power is low. *Open Epidemiology Journal*, *3*, 16–19.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Revised ed.). Academic Press.
- Durante, K. M., Griskevicius, V., Cantú, S. M., & Simpson, J. A. (2014). Money, status, and the ovulatory cycle. *Journal of Marketing Research*, *51*, 27–39.
- Eck, J., Schoel, C., Reinhard, M. A., & Greifeneder, R. (2019). When and why being ostracized affects veracity judgments. *Personality and Social Psychology Bulletin*, *46*, 454–468.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.
- Fehring, R. J., & Schneider, M. (2008). Variability in the hormonally estimated fertile phase of the menstrual cycle. *Fertility and Sterility*, *90*(4), 1232–1235.
- Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from  $\alpha$ -error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*(6), 661–669.
- Francis, G. (2013). Publication bias in “red, rank, and romance in women viewing men,” by Elliot et al. *Journal of Experimental Psychology: General*, *142*, 292–296.
- Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*(1), 3–12.
- Gangestad, S. W., Haselton, M. G., Welling, L. L., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., ... & Puts, D. A. (2016). How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications. *Evolution and Human Behavior*, *37*(2), 85–96. <https://doi.org/10.1016/j.evolhumbehav.2015.09.001>
- Guermendi, E., Vegetti, W., Bianchi, M. M., Uglietti, A., Ragni, G., & Crosignani, P. (2001). Reliability of ovulation tests in infertile women. *Obstetrics and Gynecology*, *97*(1), 92–96.
- Guida, M., Tommaselli, G. A., Palomba, S., Pellicano, M., Moccia, G., Di Carlo, C., & Nappi, C. (1999). Efficacy of methods for determining ovulation in a natural family planning program. *Fertility and Sterility*, *72*(5), 900–904.
- Hasan-Aslih, S., Pliskin, R., van Zomeren, M., Halperin, E., & Saguy, T. (2019). A darker side of hope: Harmony-focused hope decreases collective action intentions among the disadvantaged. *Personality and Social Psychology Bulletin*, *45*(2), 209–223.
- Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling* [White paper]. [https://pdfs.semanticscholar.org/e9bb/7b23993113a73ee1ff6cde5ff9a4164f946e.pdf?\\_ga=2.138693107.1964489091.1583384140-607929785.1576734677](https://pdfs.semanticscholar.org/e9bb/7b23993113a73ee1ff6cde5ff9a4164f946e.pdf?_ga=2.138693107.1964489091.1583384140-607929785.1576734677)

- Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
- Lakens, D., McLatchie, N., Isager, P., Scheel, A., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*, 75, 45–57.
- Lehmann, G. K., & Calin-Jageman, R. J. (2017). Is red really romantic? Two pre-registered replications of the red-romance hypothesis. *Social Psychology*, 48, 174–183.
- Mackinnon, D. P. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice*, 21(6), 675–681.
- Martin, A. E., North, M. S., & Phillips, K. W. (2019). Intersectional escape: Older women elude agentic prescriptions more than older men. *Personality and Social Psychology Bulletin*, 45(3), 342–359.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163.
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, 70(6), 487–498.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114(2), 376–390.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives of Psychological Science*, 11, 730–749.
- Nelson-Coffey, S. K., Killingsworth, M., Layous, K., Cole, S. W., & Lyubomirsky, S. (2019). Parenthood is associated with greater well-being for fathers than mothers. *Personality and Social Psychology Bulletin*, 45(9), 1378–1390.
- Netchaeva, E., & Kouchaki, M. (2018). The woman in red: Examining the effect of ovulatory cycle on women’s perceptions of and behaviors toward other women. *Personality & Social Psychology Bulletin*, 44(8), 1180–1200.
- Porterfield, S. P. (2001). *The Mosby physiology monograph series. Endocrine physiology* (2nd ed.). Mosby.
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553–565.
- Roney, J. R. (2018). Hormonal mechanisms and the optimal use of luteinizing hormone tests in human menstrual cycle research. *Hormones and Behavior*, 106, A7–A9.
- Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, 27(7), 1043–1046.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), 551–566.
- Schuirman, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.
- Simonsohn, U. (2014, March 12). *No-way interactions* [Data colada]. <http://doi.org/10.15200/winn.142559.90552>
- Simonsohn, U. (2017). *P-curve app 4.06*. <http://www.p-curve.com/app4/>
- Small, C. M., Manatunga, A. K., & Marcus, M. (2007). Validity of self-reported menstrual cycle length. *Annals of Epidemiology*, 17(3), 163–170.
- Sparks, J., & Ledgerwood, A. (2019). Age attenuates the negativity bias in reframing effects. *Personality and Social Psychology Bulletin*, 45(7), 1042–1056.
- Townsend, S. S., Stephens, N. M., Smallets, S., & Hamedani, M. G. (2019). Empowerment through difference: An online difference-education intervention closes the social class achievement gap. *Personality and Social Psychology Bulletin*, 45(7), 1068–1083.
- Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, 23(1), 87–102.
- Voelkel, J. G., & Brandt, M. J. (2019). The effect of ideological identification on the endorsement of moral values depends on the target group. *Personality and Social Psychology Bulletin*, 45(6), 851–863.
- Wang, I. M., & Ackerman, J. M. (2019). The infectiousness of crowds: Crowding experiences are amplified by pathogen threats. *Personality and Social Psychology Bulletin*, 45(1), 120–132.
- Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, 61(8), 1340–1341.
- Wilcox, A. J., Dunson, D. B., Weinberg, C. R., Trussell, J., & Baird, D. D. (2001). Likelihood of conception with a single act of intercourse: Providing benchmark rates for assessment of post-coital contraceptives. *Contraception*, 63(4), 211–215.
- Yao, D. J., & Chao, M. M. (2019). When forgiveness signals power: Effects of forgiveness expression and forgiver gender. *Personality and Social Psychology Bulletin*, 45(2), 310–324.