Psychological Bulletin
2014, Vol. 140, No. 5, 1260–1264

0033-2909/14/$12.00 DOI: 10.1037/a0036478

COMMENT

Elastic Analysis Procedures: An Incurable (but Preventable) Problem in the Fertility Effect Literature. Comment on Gildersleeve, Haselton, and Fales (2014)

Christine R. Harris and Harold Pashler
University of California, San Diego

Laura Mickes
Royal Holloway, University of London

Gildersleeve, Haselton, and Fales (2014) presented a meta-analysis of the effects of fertility on mate preferences in women. Research in this area has categorized fertility using a great variety of methods, chiefly based on self-reported cycle length and time since last menses. We argue that this literature is particularly prone to hidden experimenter degrees of freedom. Studies vary greatly in the duration and timing of windows used to define fertile versus nonfertile phases, criteria for excluding subjects, and the choice of what moderator variables to include, as well as other variables. These issues raise the concern that many or perhaps all results may have been created by exploitation of unacknowledged degrees of freedom ("p-hacking"). Gildersleeve et al. sought to dismiss such concerns, but we contend that their arguments rest upon statistical and logical errors. The possibility that positive results in this literature may have been created, or at least greatly amplified, by p-hacking receives additional support from the fact that recent attempts at exact replication of fertility results have mostly failed. Our concerns are also supported by findings of another recent review of the literature (Wood, Kressel, Joshi, & Louie, 2014). We conclude on a positive note, arguing that if fertility-effect researchers take advantage of the rapidly emerging opportunities for study preregistration, the validity of this literature can be rapidly clarified.

Keywords: menstrual cycle, ovulatory cycle, p-hacking, fertility effects, mate preferences

Meta-analysis can often be very useful in allowing the scientific community to rationally aggregate the information contained in a complex literature, especially one that contains conflicting findings. However, it has been generally acknowledged that the credibility of a meta-analysis can be undermined in some situations. One widely recognized threat arises when unpublished studies are omitted. Another major threat, rarely discussed until the past few years, is at least as injurious to the interpretability of literature syntheses and also much harder to mitigate. This threat involves the bias that arises when the data in the studies being surveyed were originally analyzed in a tendentious fashion by experimenters wittingly or unwittingly exploiting unacknowledged degrees of freedom in order to obtain positive results (Simmons, Nelson, & Simonsohn, 2011). Such biased analyses have come to be referred to as "p-hacking."

We recently pointed out that the evolutionary psychology literature on fertility effects may be unusually prone to p-hacking partially because of the high degree of variability that exists from study to study in the method used to classify different women as fertile or infertile (Harris, 2013; Harris, Chabot, & Mickes, 2013). The idea that positive results in this literature may have been created, or at least greatly amplified, by p-hacking receives support from two observations: (a) recent attempts at exact replication of fertility results have mostly failed, and (b) an examination of just the unpublished studies in the literature revealed no effects[1] (Wood, Kressel, Joshi, & Louie, 2014).

In their review, Gildersleeve, Haselton, and Fales (2014) mentioned the problem of analytical elasticity, but they attempted to dismiss it. The purpose of this commentary is to show why this dismissal is unfortunately unpersuasive. We argue that Gildersleeve et al. have made logical errors in their analysis of how p-hacking would be expected to manifest itself in the literature. Our comment concludes on a positive note, however, pointing out that even if the current literature is inconclusive because of analytical flexibility and the potential for p-hacking, a solution is rapidly emerging that will allow future studies to definitively avoid these problems and test existing claims.

The arguments offered by Gildersleeve et al. (2014) to dismiss the role of p-hacking involve several misleading sets of claims,

Christine R. Harris and Harold Pashler, Department of Psychology, University of California, San Diego; Laura Mickes, Department of Psychology, Royal Holloway, University of London.

Correspondence concerning this article should be addressed to Christine R. Harris, Department of Psychology, University of California, San Diego, 9500 Gilman Drive #0109, La Jolla, CA 92093-0109. E-mail: charris@ucsd.edu

[1] One potential way that p-hacking might explain such null findings would be if investigators who failed to p-hack usually found no results, and therefore, did not publish their results, whereas those who used p-hacking tended to succeed in getting positive findings, which they then published.

which we discuss and rebut in the following text. First, they downplayed the actual amount of variation in fertility definitions found in the literature. We point out that the flexibility is quite dramatic—sometimes even enough to reverse the direction of effects. Second, they argued that this potential flexibility is not actually leading to biased effects because investigators decide in advance how they will analyze their data. We acknowledge that the frequency of "fishing expeditions" in this literature is obviously not directly observable, but we argue that the fact that fertile periods often vary from one article to another within the same lab strongly suggests that p-hacking is more than a theoretical possibility (see Harris et al., 2013, Figure 1). Third, they maintained that the use of continuous fertility methods (as in a relatively small subset of the studies that they reviewed) gets around the problem of flexibility; we explain why it does not solve the problem. Finally, Gildersleeve et al. (2014) contended that specific quantitative findings from their meta-analysis argue against the idea that results are being inadvertently manufactured through p-hacking. We show that these arguments rest on logical and statistical confusions. We turn now to the first issue.

How Much Flexibility Is There in the Literature?

In attempting to argue that p-hacking is not a problem in this literature, Gildersleeve et al. (2014) played down the amount of flexibility in analyses that a typical menstrual cycle study has. They wrote, "Most aspects of study design are determined in advance of data collection, eliminating concerns about researcher degrees of freedom" (p. 1249). They acknowledged that definition of high- and low-fertility windows are "not always determined in advance of data collection" (p. 1249), seemingly implying that they are usually predetermined. In a contemporaneous piece, Gildersleeve et al. (2013) boldly stated, "It is implausible that these findings are a mere artifact of 'researcher degrees of freedom'" (p. 520).

In this literature, there are an unusual number of analytical choices that investigators must make. Ironically, Gildersleeve et al.'s (2014) own figures illustrated some of these choices (see Figure 3, p. 1250).[2] For example, consider the assignment of women into the high- and low-fertile groups. The experimenter chooses (a) the number of days to be counted as fertile, (b) the specific days considered fertile, (c) the number of days to be counted as infertile, (d) the specific days considered infertile, (e) which days will be thrown out of analyses all together (e.g., excluding all women who have cycle days longer than 28 days), and (f) whether to assess fertility with a forward method, backward method, or some mixture of the two. Together, these decisions can dramatically alter fertility classifications.

For example, based on the numbers in Gildersleeve et al.'s (2014) Figure 3, the number of days considered as fertile ranges from 3 to 15 days and the number of days considered infertile ranges from 3 to 22 days.[3] Thus, if experimenters simply chose from the range of options that already exist in the field, they would have at least 13 options for the number of days to include as high fertility and 20 options for low-fertility days. The total number of options available explodes when one considers all of the choices created by different combinations of these variables, along with the many options for the specific placement of these windows.

The flexibility in analyses, however, does not end with determining the fertile and infertile windows. Researchers often commonly make additional decisions about which subjects to exclude (e.g., women over a particular age, single women) and which moderators to examine (e.g., relationship status; primary mate's characteristics). Another avenue for flexibility comes from using different transformations of dependent variables.

If changes in fertility definitions are large, one might still wonder whether they are large enough to alter results. Although this question could benefit from detailed study applying simulated classifications to real data sets, there is little doubt that changing fertility windows can transform the results of a study. For example, Harris (2011) attempted to replicate Penton-Voak and colleagues' findings that women in the fertile phase preferred more masculine-faced men than women in nonfertile phases (Penton-Voak & Perrett, 2000; Penton-Voak et al., 1999) using stimuli obtained from Penton-Voak. Harris' primary analyses were performed exactly as Penton-Voak and Perrett (2000) had performed them: Women who had more than 28 cycle days were excluded, women who were on Cycle Days 6–14 were placed in the high-fertility group, and all others were placed in the low-fertility group. Harris found a significant effect ($p < .03$), but in the *opposite* direction to that reported by Penton-Voak and colleagues: women in the fertile phase preferred *less* masculine faces relative to women in the nonfertile phase. In subsequent analyses, Harris showed how this effect could be changed by relatively small shifts in the fertile window. When Days 8–16 were considered the fertile days (a shift of only 2 days from the previous analysis) and all remaining days were counted as not fertile, the effect disappeared ($p > .56$). Interestingly, Gildersleeve et al. (2014) chose to include the latter analysis, rather than the former, in their meta-analysis.[4] In other work, Wood et al. (2014) reanalyzed data from Frost (1994) and showed that reducing the fertile phase window by 1 day could make a previously significant result no longer significant ($p = .283$). Thus, even slight variations in analytic strategies sometimes

[2] Perusal of this figure may actually leave the reader with an *overestimation* of the consistency in the literature, given that a single study often has multiple separate entries. It also should be noted that their figure has errors, of which we note just two: (a) Little, Jones, Burt, and Perrett (2007) were listed as having excluded women who were on Day 15 or higher in their cycle when in fact these women were included in the nonfertile group. (b) Rupp et al. (2009) were listed as having used a continuous fertility method when they actually placed participants in low-fertility (Days 1–5 and 17–35) and high fertility (Days 6–16) groups.

[3] This number is actually greater since some studies include cycle lengths as long as 40 days, but the Gildersleeve et al. (2014) figure only displays data for 28 days. Extreme variability also has been documented in other reviews (Harris et al., 2013; Wood et al., 2014).

[4] Gildersleeve et al. (2014) justified their selection by claiming that they chose to include the Harris (2011) analysis that was based on the high-fertility window with the highest estimated average conception probability according to the values reported by Wilcox, Dunson, Weinberg, Trussell, and Baird (2001). Previously, Gildersleeve et al. (2013) claimed "the backward counting method is generally regarded as a more accurate method of estimating cycle position and fertility" (pp. 519). Therefore, it is odd that they did not choose to use the analyses in which Harris used a backward fertility estimate, especially since this is the analysis for which the high-fertility window had the highest estimated average conception probability (contrary to the claims of Gildersleeve et al., 2014, in footnote 8).

have drastic effects on whether an analysis produces significant effects.

Is Potential Analytical Flexibility Being Exploited to Seek Positive Results?

We contend that when analytical flexibility is present, it is only sensible to assume that some experimenters (whether consciously or unconsciously) will be exploiting this flexibility in order to find positive publishable effects, and more specifically to arrive at findings they find theoretically agreeable. In a series of simulations, Simmons et al. (2011) examined the consequences of a number of research practices such as excluding subsets of participants, exploring different transformations of dependent measures, not reporting all analyses or conditions, and so on. They showed that such tactics, especially when adopted in combination, can increase false alarm rates far above the nominal 5% that is assumed to exist in published research. For example, Simmons et al. found that performing four such practices in combination resulted in a 60% likelihood of finding an effect that was significant at $p = .05$ in the absence of any real effect.

The clearest evidence for exploitation of analytical flexibility arises when the same researchers adopt and then discard various transformations of the same dependent variables from one study to the next, without any justification being provided for these shifts. For example, both Pillsworth and Haselton (2006) and Haselton and Gangestad (2006) examined a mate's sexual and investment attractiveness. In Pillsworth and Haselton (2006), the types of attractiveness were analyzed separately, but in Haselton and Gangestad (2006), a difference score of sexual versus investment attractiveness was calculated and then used in analyses.

Of course, it should go without saying that many investigators undoubtedly do not engage in p-hacking, even if common practices would have allowed them to do so. To produce the slight preponderance of positive findings that Gildersleeve et al. (2014) contended the literature shows, it would probably only take a fairly modest number of biased analyses. Moreover, such behaviors may be undertaken in good faith, related to ignorance of their consequences. The outpouring of interest that the Simmons et al. (2011) article has drawn within the scientific community probably shows that many investigators have been unaware of how small choices can cumulate to easily produce statistically significant findings built out of sampling error.

Does Using Continuous Calculations for Fertility Solve the P-Hacking Issue?

Gildersleeve et al. (2014) argued that restricting their analysis to just those studies that employ continuous fertility calculations (rather than dividing women into high- and low-fertility groups) can resolve concerns about p-hacking, because such studies do not require the investigator to specify windows. They suggested that some effects are still present in this subgroup of studies and, therefore, that experimenter degrees of freedom are not an issue.

However, while continuous fertility calculation methods generally reduce the number of choices to be made relative to the categorical classification methods, they are not necessarily as straightforward as Gildersleeve et al. (2014) suggested, and many such studies still report making complex and idiosyncratic analysis decisions. Most investigations cite Wilcox et al. (2001) as the source of their continuous fertility numbers. Wilcox et al. provided a table for estimated probability of pregnancy following a single act of unprotected intercourse, which provides a risk estimate for each day in a woman's cycle up to Day 40 (separately for women whose cycle lengths are consistent vs. irregular). The most straightforward way to use this table is to assign a woman a probability value based on the day in her cycle (as done by Morrison, Clark, Gralewski, Campbell, & Penton-Voak, 2010). However, other researchers take into account a woman's predicted cycle length and perform transformations to force the woman on a 28/29-day cycle (regardless of what her natural cycle length is) and then try to apply the Wilcox et al. table to these transformed numbers (e.g., Thornhill, Chapman, & Gangestad, 2013).[5] Other studies combine the 28/29-day conversions with additional transformations such as averaging forward and backward fertility calculations (e.g., Gangestad, Garver-Apgar, Simpson, & Cousins, 2007). Although such analytic strategies may be individually defensible, the fact that such a range of choices exists undercuts Gildersleeve et al.'s argument.

Quantitative Evidence About P-Hacking

Gildersleeve et al. (2014) also described some specific analyses they undertook to shed light on the possible role of p-hacking (pp. 45–48). In one such analysis, they took all the studies in which women were placed in discrete low- or high-fertility categories and attempted (using a continuous estimator of fertility) to estimate the fertility difference expected from this choice of window. That amounted to a grade of how successfully the study sorted high- from low-fertility women. Having graded studies in this way, they looked at whether there was a detectable correlation across studies between this grade and the effect size that the study reported for the outcome variable. They reported that there was no sign that studies that made a "good choice" of window produced any bigger effects.

The results of this analysis are problematic for the view that this literature reveals real fertility effects. If it does, then studies that better measure fertility should tend to show stronger effects. On the other hand, if fertility effects in the literature were all produced by Type-1 errors based upon sampling error, then the windows that are graded high and those graded low by Gildersleeve et al.'s (2014) procedure should not differ in the size of reported effects. This is what their analysis showed. (The failure of better "quality" studies to show effects also has been reported by Wood et al., 2014, e.g., studies that used hormonal assays did not find more robust effects.)

Having failed to find confirmation for real effects from their analysis, Gildersleeve et al. (2014) then displayed a figure with effect sizes for a subset of studies along with information specifying which days each study counted as fertile or infertile and which were excluded altogether. According to Gildersleeve et al.,

[5] This procedure could alter the data quite substantially. Take, for example, a woman who has a 30-day-cycle and is tested on Day 14. A straightforward application of Wilcox et al. (2001). would give her a pregnancy risk estimate of .085. However, transformations using her cycle length and assumptions that she is fertile 15 days prior to the start of her next cycle would produce a risk estimate of .059.

the reader should be able to see that there are no interesting correlations between larger effect sizes and any of three things:

> (a) more variable high- and low-fertility window definitions, b) more poorly placed high- and low-fertility windows (high-fertility windows that included true low-fertility days of the cycle and/or low-fertility windows that included true high-fertility days of the cycle), and (c) less frequent use of a continuous fertility variable. (p. 1249)

Thus, Gildersleeve et al. concluded, "We used multiple procedures to assess and adjust for various forms of potential bias. The results of these procedures do not suggest that these sources of bias account for the robust cycle shifts observed in this meta-analysis" (p. 1249).

We find it odd that Gildersleeve et al. would suggest that readers can rely on "eyeball" judgments of complex multidimensional data—especially when the data are presented in tabular fashion. Clearly, this provides no reliable information about whether any relationships exist. But even if they *had* demonstrated the lack of any such relationships, the implications for possible p-hacking would not be self-evident. The patterns produced by p-hacking would depend upon how the p-hacking was carried out. For example, if there were no true effect of fertility on the outcome variables being measured, and investigators tried multiple different candidate windows selected at random until a significant result was obtained (if at all), one should not expect to see any relation between choice of a fertile window and the strength of the positive effects. In a universe composed exclusively of Type-1 errors, any measurement of the independent variable is as good (or bad) as any other.

The most reasonable way to see whether p-hacking could result in a correlation between window duration and effect size would be if the following conditions happen to hold: (a) researchers who p-hack begin their analyses with small windows and then try progressively larger windows until they get an effect, and (b) researchers vary in how much they persevere in p-hacking. Interestingly, another recent meta-analysis of this literature (Wood et al., 2014) revealed that larger fertility windows were *more* likely to show effects. This effect is exactly what one would predict if investigators tended to start with relatively narrow fertility windows and then some investigators (i.e., those engaging in more energetic p-hacking) expanded the window in search of effects. We are unable to think of any completely benign explanation for this pattern, and it would seem to trump Gildersleeve et al.'s (2014) failure to observe any relationships from looking at a figure.

Looking Forward to Better Research: How to Prevent Such Problems in the Future

The general point of the current comment is to say that excess analytical flexibility makes the literature less conclusive than it could and should be. Supporting this interpretation is that those few studies in which we can be sure that analytical flexibility was not present—namely, recent articles that attempted to perform *direct replications*—have generally reported negative results. For example, as described earlier, Harris (2011) carried out a fairly direct replication of the work by Penton-Voak and colleagues, using the same choices as the original investigators, and found no evidence whatsoever for the predicted cycle shift in facial masculinity preferences (with the effect running in the opposite direction). In two additional investigations in which these same methods, fertility classifications, and so on were used, researchers failed to find any effect of cycle phase on masculinity preferences, even when the relationship context was specified as short-term (Mickes & Harris, 2014). Another direct replication of shifts in religiosity, political attitudes, and voting preferences (Harris & Mickes, in press) failed to support most of the effects of fertility reported by Durante, Rae, and Griskevicius (2013).

Looking beyond these direct replications, what methodological improvements can help avoid the kinds of problems discussed here? Fortunately, the problem of flexibility and potential for p-hacking that bedevil the fertility effect literature can be prevented in a simple and decisive fashion. What is needed is for investigators to conduct future studies using preregistration (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) of their definition of fertile periods and all other analytical methods, including plans for excluding subjects and potential moderator variables to be included in analyses. Preregistration is now possible and convenient through the Open Science Framework (https://OSF.io). (The use of biological measures of fertility may also increase reliability, but the use of these tests does not eliminate many of the forms of elasticity described; so when used, they should be combined with preregistration.) Preregistration became the norm some years ago in clinical trials, and there is rapidly increasing awareness of its potential to advance basic research as multiple journals and organizations embrace it. New studies should routinely utilize preregistration, and key studies in the literature need to be replicated with prespecified procedures as well. If this plan is adopted, we will soon have a good idea of whether the findings in this literature are solid but relatively small (as Gildersleeve et al., 2014, contend) or whether many effects have been invented out of the whole cloth (as the Wood et al., 2014, meta-analysis would seem to suggest).

References

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science, 24,* 1007–1016. doi:10.1177/0956797612466416

Gangestad, S. W., Garver-Apgar, C. E., Simpson, J. A., & Cousins, A. J. (2007). Changes in women's mate preferences across the ovulatory cycle. *Journal of Personality and Social Psychology, 92,* 151–163. doi:10.1037/0022-3514.92.1.151

Gildersleeve, K., DeBruine, L., Haselton, M. G., Frederic, D. A., Penton-Voak, I. S., Jones, B. C., & Perrett, D. I. (2013). Shifts in women's mate preferences across the cycle: A critique of Harris (2011) and Harris (2012). *Sex Roles, 69,* 516–524. doi:10.1007/s11199-013-0273-4

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,* 1205–1259. doi:10.1037/a0035438

Harris, C. R. (2011). Menstrual cycle and facial preferences reconsidered. *Sex Roles, 64,* 669–681. doi:10.1007/s11199-010-9772-8

Harris, C. R. (2013). Shifts in masculinity preferences across the menstrual cycle: Still not there. *Sex Roles, 69,* 507–515. doi:10.1007/s11199-012-0229-0

Harris, C. R., Chabot, A., & Mickes, L. (2013). Shifts in methodology and theory in menstrual cycle research on attraction. *Sex Roles, 69,* 525–535. doi:10.1007/s11199-013-0302-3

Harris, C. R., & Mickes, L. (in press). Women can keep the vote: No evidence that hormonal changes during the menstrual cycle impact political and religious beliefs. *Psychological Science*.

Haselton, M. G., & Gangestad, S. W. (2006). Conditional expression of women's desires and men's mate guarding across the ovulatory cycle. *Hormones and Behavior, 49,* 509–518. doi:10.1016/j.yhbeh.2005.10.006

Little, A. C., Jones, B. C., Burt, D. M., & Perrett, D. I. (2007). Preferences for symmetry in faces change across the menstrual cycle. *Biological Psychology, 76,* 209–216. doi:10.1016/j.biopsycho.2007.08.003

Mickes, L., & Harris, C. R. (2013). [No ovulatory cycle effects on masculinity preferences]. Unpublished data.

Morrison, E. R., Clark, A. P., Gralewski, L., Campbell, N., & Penton-Voak, I. S. (2010). Women's probability of conception is associated with their preference for flirtatious but not masculine facial movement. *Archives of Sexual Behavior, 39,* 1297–1304. doi:10.1007/s10508-009-9527-1

Penton-Voak, I. S., & Perrett, D. I. (2000). Female preference for male faces changes cyclically: Further evidence. *Evolution and Human Behavior, 21,* 39–48. doi:10.1016/S1090-5138(99)00033-1

Penton-Voak, I. S., Perrett, D. I., Castles, D. L., Kobayashi, T., Burt, D. M., Murray, L. K., & Minamisawa, R. (1999, June 24). Menstrual cycle alters face preference. *Nature, 399,* 741–742. doi:10.1038/21557

Pillsworth, E. G., & Haselton, M. G. (2006). Male sexual attractiveness predicts differential ovulatory shifts in female extra-pair attraction and male mate retention. *Evolution and Human Behavior, 27,* 247–258. doi:10.1016/j.evolhumbehav.2005.10.002

Rupp, H. A., Librach, G. R., Feipel, N. C., Ketterson, E. D., Sengelaub, D. R., & Heiman, J. R. (2009). Partner status influences women's interest in the opposite sex. *Human Nature, 20,* 93–104. doi:10.1007/s12110-009-9056-6

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,* 1359–1366. doi:10.1177/0956797611417632

Thornhill, R., Chapman, J. F., & Gangestad, S. W. (2013). Women's preferences for men's scents associated with testosterone and cortisol levels: Patterns across the ovulatory cycle. *Evolution and Human Behavior, 34,* 216–221. doi:10.1016/j.evolhumbehav.2013.01.003

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. doi:10.1177/1745691612463078

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,* 211–215.

Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review, 6,* 229–249. doi:10.1177/1754073914523073

Received December 17, 2013
Revision received February 4, 2014
Accepted February 10, 2014 ■