

# Efficacy of methods for ovulation estimation and their effect on the statistical detection of ovulation-linked behavioral fluctuations

Joseph E. Gonzales<sup>1</sup> · Emilio Ferrer<sup>1</sup>

Published online: 25 July 2015  
© Psychonomic Society, Inc. 2015

**Abstract** Contentions of the ovulatory shift hypothesis are principally supported by failures to replicate previous findings; e.g., recent meta-analytic work suggests that the effects endorsing the hypothesis may not be robust. Some possible limitations in this and other ovulatory-effects research—that may contribute to such controversy arising—are: (a) use of error-prone methods for assessing target periods of fertility that are thought to be associated with behavioral shifts, and (b) use of between-subjects—as opposed to within-subjects—methods. In the current study we present both simulated and empirical research: (a) comparing the ability of between- and within-subject t-tests to detect cyclical shifts; (b) evaluating the efficacy of correlating estimated fertility overlays with potential behavioral shifts; and (c) testing the accuracy of counting methods for identifying windows of cycle fertility. While this study cannot assess whether the ovulatory shift hypothesis or other ovulatory-based hypotheses are tenable, it demonstrates how low power resulting from typical methods employed in the extant literature may be associated with perceived inconsistencies in findings. We conclude that to fully address this issue greater use of within-subjects methodology is needed.

**Keywords** Fertility estimation · Ovulatory effects · Ovulatory shift hypothesis · Statistical analysis · Data simulation

---

✉ Joseph E. Gonzales  
joegonzales@ucdavis.edu

<sup>1</sup> Department of Psychology, University of California, Davis, One Shields Ave., Davis, CA 95616-8686, USA

## Introduction

Research on ovulatory effects in humans spans a broad range of behaviors. For example, there is evidence that women's fertility status predicts differences in their self-ornamentation (e.g., Beall & Tracy, 2013; Haselton, Mortezaie, Pillsworth, Bleske-Rechek, & Frederick, 2007), perceived attractiveness and desirability (e.g., Roberts et al., 2004; Schwarz & Hassebrauck, 2008), earnings from lap dances (Miller, Tybur, & Jordan, 2007), frequency of sexual intercourse (e.g., Bullivant, et al., 2004; Wilcox et al., 2004), and shifts in attraction to traits indicative of men's masculinity (e.g., Penton-Voak & Perrett, 2000). Similarly, men have been found to engage in more mate-guarding behavior (e.g., Gangestad, Thornhill, & Garver, 2002; Pillsworth & Haselton, 2006) and have differential testosterone production (e.g., Miller & Maner, 2010) in response to women's fertility status. However, despite this breadth of effects presumably related to fertility, criticisms of the field, its research and its theories, abound.

For example, ongoing criticisms of the ovulatory shift hypothesis (Harris, 2011, 2013; Harris, Chabot, & Mickes, 2013; Harris, Pashler, & Mickes, 2014) contend that inconsistent methodologies, failures to replicate findings (e.g., Peters, Simmons, & Rhodes, 2009), and recent meta-analytic work concluding ovulatory effects are not robust (i.e., Wood & Joshi, 2011; Wood, Kressel, Joshi, & Louie, 2012a, b, 2014) indicate that research supporting the ovulatory shift hypothesis are the result of spurious findings or inflated researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). These accusations are extreme, and probably false given a rebuttal meta-analysis by Gildersleeve, Haselton, and Fales (2014a) and *p*-curve analysis by Gildersleeve, Haselton, and Fales (2014b). However, these criticisms were not completely unfounded. That is, despite recommendations

of ideal methodological practices when addressing failures to replicate (e.g., DeBruine et al., 2010), there is generally an inconsistent use of ovulation estimation methods, between- versus within-subjects designs, definitions of fertility windows, and methods employed to estimate fertility in the literature (Harris, 2013; Harris et al., 2013, 2014; Gildersleeve et al., 2014b). There can be very good reasons for methodological variability in this area of research (e.g., Gildersleeve et al., 2013), but such variability in lieu of recommended practices can give the appearance of post-hoc methodological modification.

Consider Gildersleeve et al.'s (2014a) sample of studies specific to the ovulatory shift hypothesis—the Gildersleeve et al. and Wood et al. (2014) samples are very similar so we give preference to Gildersleeve et al. as they are expert contributors and proponents of this body of research. Despite notions of methodological best practices, between-subjects group designs were used in 43.7 % of studies, and between-subjects continuous measures were used in 18.3 %. Thus, 62 % of all studies reviewed employed a between-subjects design. This means that, despite the supposed strength of repeated measure methods, within-subjects designs were only employed in 38 % of the studies reviewed.

Furthermore, with respect to fertility estimation, both forward-counting and fertility-likelihood overlay methods (e.g., Wilcox, Dunson, Weinberg, Trussell, & Baird, 2001) were employed heavily in the reviewed literature. Specifically, continuous overlays were used in 18.3 % of all studies, and account for 23.9 % of studies that used forward counting and 11.1 % of studies that used some form of backward counting. Of those studies that employed backward counting (25.4 % of all studies), only 33.3 % confirmed menstrual onset to ensure that backward estimates were derived from the correct start date; i.e., only 8.5 % of all reviewed studies confirmed menstrual onset. Finally, only six studies (8.5 % of all studies) employed some sort of assay of hormones, five of which used luteal hormone (LH) testing (7.1 % of all studies).

It may be unfair to evaluate the methodology of the entire sample of the ovulatory-shift literature given as a single period given that recommended practices emerge as the field develops. For example, treating DeBruine et al.'s (2010) commentary on the methodological limitations of Harris' (2011) replication attempt as an indicator of the prevalence of methodological ideals in ovulatory-shift research suggests some shifts in methodological focus. Specifically, before 2010, forward-counting methods were used in 69.8 % of studies and backward-counting methods were used in 22.6 % of studies. From 2010 onward, forward-counting utilization declined to 50 % of studies, and backward-counting methods rose to 33.3 % of studies. Similarly, between-subjects designs were used more often than within-subjects designs (64.2 % and 35.8 %, respectively) before 2010, but from 2010 onward

the use of between-subjects designs declined and within-subject designs increased (55.6 % and 44.4 %, respectively). Finally, comparing pre-2010 studies with 2010 and onward studies there were small increases in the relative frequency of studies confirming menstruation onset when using backward counting (7.5 % and 11.1 %, respectively), and in the use of hormonal estimation of ovulation (7.5 % and 11.1 %, respectively).

While these shifts in methodological practices are encouraging, it is clear that there is still a heavy reliance on counting methods (generally), despite evidence and arguments that other approaches (e.g., LH testing) may be more effective (e.g., Bullivant et al., 2004; DeBruine et al., 2010), and on both forward and pseudo backward counting (specifically), despite evidence that these procedures are likely more error prone than backward counting with confirmation of next menstrual cycle onset (e.g., Fehring, Schneider, & Raviele, 2006). Utilization of these methods is problematic because increased error in measurement due to misidentification of ovulation and fertile periods may diminish power to detect phase-related behavioral shifts. Since both the follicular and luteal phases are variable (e.g., Fehring et al., 2006), both between and within women (Creinin, Keverline, & Meyn, 2004; Fehring et al., 2006), static assumptions of cycle length—such as when using counting methods—reduces ovulation estimation accuracy by 37–57 % (Howards et al., 2008). This problem is further compounded when estimation relies on self-reported cycle length (i.e., pseudo backwards counting) since self-report of cycle length has been shown to be inaccurate in women (e.g., Small, Manatunga, & Marcus, 2007).

## The present study

While continuing the academic debate over which meta-analysis is correct (Gildersleeve et al., 2014b; Wood & Carden, 2014) makes for “good theater” (Ferguson, 2014), it is plausible that methodological limitations can account for many of the inconsistencies in these findings (e.g., Harris, 2011, 2013; Wood et al., 2014). Comments regarding methodological limitations and recommendations (e.g., DeBruine et al., 2010; Gildersleeve et al., 2013) reflect the following: (a) within-subjects designs are better than between-subjects designs; (b) forward counting is less precise than backward counting, and both are less accurate than hormonal assessments; (c) optimum methods are more expensive with respect to time and money; and (d) with a sufficiently large sample less optimum methodology is expected to overcome method-based reductions in power. However, it is not immediately clear how these reasonable expectations may actually function in research. Specifically, it is not well understood how much of a decrement in power is observed when using different methods (e.g., between-subjects vs. within-subjects; counting

vs. LH testing) or when these differences are negligible rather than pronounced.

To better understand the relation of power and methodology as it pertains to ovulatory effects we evaluate the following conditions for group-based mean comparisons using simulated data: (a) ovulation estimation method (LH testing vs. forward counting vs. unconfirmed backward counting vs. confirmed backward counting); (b) between- vs. within-subject designs; (c) fertile window length (day of peak fertility vs. sample from six-day window vs. sample from nine-day window); and (d) sample size ( $N = 10, 20, \dots, 200$ ). Additionally, we evaluate the ability of fertility overlays (i.e., Wilcox et al., 2001) to detect behavioral variability correlated with predicted fertility fluctuations as a function of sample size. Finally, we evaluate differences in fertile days selected as a function of ovulation estimation method using counting methods in both simulated and empirical data sets.

## Method

### Simulation study

**Data simulation procedure** In the present study we simulated a population of 20,000 ovulatory cycles. Each simulated cycle represented a single cycle of an individual woman. All cycles were made up of daily scores. Cycles were generated using estimates of the mean, variance, and covariance of cycle-phase lengths (i.e., menses, follicular, and luteal) previously reported (Fehring et al., 2006; R. J. Fehring, personal communication, 7 April 2014) (see Table 1). The Fehring et al. (2006) sample statistics were used to inform the present simulation for several reasons: (a) its frequency of citation in ovulatory research generally; (b) its frequency of citation in ovulatory-shift research specifically; (c) its citation as evidence of average cycle phase length, variability between and within women, and relative stability of luteal phase lengths compared to the follicular phase (e.g., Garver-Apgar et al., 2008; Gildersleeve et al., 2014a; Larson et al., 2013; Lukaszewski & Roney, 2009; Miller & Maner, 2010; Oinonen & Mazmanian, 2007; Prokosch, Coss, Scheib, & Blozis, 2009; Roney et al., 2011; Rosen & López, 2009; Schwarz & Hassebrauck, 2008); and (d) Its specific use as justification for using backward counting, rather than forward, to estimate ovulation and the fertile window (e.g., Oinonen & Mazmanian, 2007; Prokosch et al., 2009; Roney et al., 2011; Rosen & López, 2009; Schwarz & Hassebrauck, 2008). Taken together, it is apparent that researchers in the field value Fehring et al.'s work as sufficiently reliable to inform their own research methodology decisions. We therefore decided it represented a sound source for the present study's cycle simulation parameters.

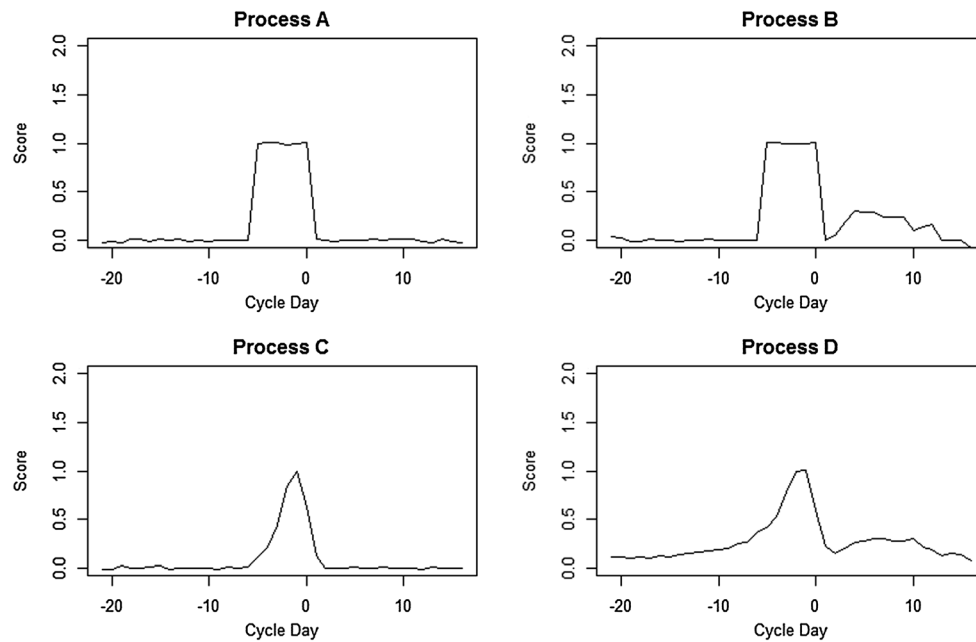
**Table 1** Simulation target values and observed simulation values

Fehring et al. (2006)				
Cor matrix	Length	Follicular	Luteal	Menses
Length	1	-	-	-
Follicular	0.826	1	-	-
Luteal	0.247	-0.335	1	-
Menses	0.194	0.237	-0.110	1
<i>M</i>	28.9	16.5	12.4	5.9
<i>SD</i>	3.3	3.4	2	2.9
95 % cases	22–36	9–23	8–17	0–12
Current simulation				
Cor matrix	Length	Follicular	Luteal	Menses
Length	1	-	-	-
Follicular	0.836	1	-	-
Luteal	0.227	-0.345	1	-
Menses	0.180	0.239	-0.116	1
<i>M</i>	28.97	16.47	12.5	5.59
<i>SD</i>	3.35	3.47	1.96	1.33
95 % cases	22–36	10–23	9–16	3–8

*Note:* Fehring et al. (2006) values were the simulation target values, and current simulation values are the values observed in the simulated data

Shifts in a hypothetical outcome (e.g., attraction to male traits) were then simulated for each cycle following four distinct but similar hypothetical process trajectories (see Fig. 1).

Process A consists of a single stepwise (i.e., on-off) behavioral shift corresponding to a six-day peak fertility window ending on the day of ovulation (e.g., Dunson, Baird, Wilcox, & Weinberg, 1999; Dunson, Colombo, & Baird, 2002; Wilcox, Weinberg, & Baird, 1995; 1998); across the six-day window the degree of behavioral change is equal. Process B is identical to Process A, except it has a second stepwise behavioral shift during the mid-luteal phase. This mid-luteal shift is meant to represent a secondary shift sometimes observed in ovulatory effects research (e.g., Miller et al., 2007) and thought to reflect dependency on hormonal processes observed across the ovulatory cycle (e.g., Roney & Simmons, 2013). The degree of the secondary shift, relative to the first, was determined by dividing the predicted mid-luteal estrogen to progesterone ratio by the average peak fertility estrogen to progesterone ratio—predicted hormone values reported by Stricker et al. (2006). This results in a mid-luteal secondary shift that is 30 % of the fertile phase maximum value. Process C is like Process A in that it has a single shift corresponding to the six-day peak in fertility. However, unlike Process A, it has a continuous curvilinear trajectory such that each day's weight of behavioral score increase is equivalent to the daily projected estrogen value divided by the predicted estrogen value for the day of peak fertility—this is calculated across the six-day window (e.g., Dunson et al., 2002; Wilcox et al., 1995; 1998). Finally, Process D is similar to Process B in that



**Fig. 1** Examples of Processes A–D

it has two surges in change, a maximum corresponding to the fertile phase and a secondary minor shift corresponding to the mid-luteal phase. Additionally, weighting of behavioral shifts are based on hormonal ratios across the entire ovulatory cycle, and not just restricted to peak fertility and mid-luteal phases. It is unclear what hormonal processes may be associated with shifts in behavior, and whether any single hormonal process is equally predictive across behavior types. To this end, we considered the average trajectory of two hormonal processes: (a) the ratio of each day's predicted estrogen to progesterone level divided by the predicted estrogen to progesterone level for the day of peak fertility; and (b) the predicted daily level of estrogen divided by the predicted level of estrogen on the day of peak fertility. The daily weights generated using these two approaches were averaged together and the resulting aggregate daily weights were used to model average daily behavioral scores.

For each trajectory we created six conditions of maximum mean value ( $M_{max} = 0.15, 0.25, 0.5, 1.0, 1.5,$  and  $2.0$ ) for days of peak fertility. Maximum mean values represent the maximum behavioral fluctuation value possible based on the daily weight. For non-fertile days, including mid-luteal days without secondary behavioral shifts,<sup>1</sup> the mean expected score was zero. Across all participants and days, variability of behavioral scores was held constant ( $SD = 1.0$ ), with the goal being that within-cycle effect sizes would then be equal to the

<sup>1</sup> Expected mid-luteal scores with secondary shifts are their proportion change score, relative to the day of peak fertility score, multiplied by the maximum mean value.

difference between the sampled high and low fertility day behavioral scores (i.e.,  $[\text{High Fertility Score} - \text{Low Fertility Score}] / 1$ ). This generating procedure resulted in an average within-cycle variability of 1.05 ( $SD = 0.14$ , 95 % CI 0.78–1.32) for behavior scores. In addition to generating behavioral shifts, a null behavioral change score was generated where no change across the cycle occurred ( $M_{max} = 0.00$ ,  $SD = 1.0$ ). Due to variability in data generation (i.e.,  $SD = 1.0$ ), and variability in degree of relative differences between peak fertility and mid-luteal phases (e.g., Processes A & C compared to Processes B & D) observed effect sizes in the population are attenuated from generating  $M_{max}$  values (see Table 2). Finally, for between-subjects designs each case had a random mean score ( $M = 5.5$ ,  $SD = 1.0$ ) added to each of their daily scores to reflect individual differences in average scores.<sup>2</sup>

**Estimation of fertility** For each simulated ovulatory cycle the true phase lengths and corresponding ovulatory days (e.g. days  $-2, -1, 0, 1,$  and  $2$ ) were known, and correspond with days of ovulation predicted using fertility monitors (i.e., Fehring et al., 2006). We estimated day of ovulation for each simulated cycle using forward counting and confirmed

<sup>2</sup> The mean of these scores has no impact on mean differences as this merely transposes the average value location. What is potentially impactful is the variability in scores between individuals as this additional variability cannot be separated from behavioral variability when only using between-subjects designs.

**Table 2** Observed population level effect sizes by process type, max mean, estimation method, and fertile window size

Process	$M_{max}$	TRUE windows			Forward windows			Pseudo backward windows			Backward confirmed windows		
		1	6	9	1	6	9	1	6	9	1	6	9
Null	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	-0.01	0.01	0.01	0.00
A	0.15	0.10	0.10	0.06	0.03	0.03	0.01	0.08	0.06	0.04	0.09	0.07	0.05
	0.25	0.18	0.17	0.11	0.05	0.04	0.03	0.14	0.13	0.10	0.14	0.13	0.10
	0.50	0.35	0.35	0.24	0.12	0.10	0.08	0.28	0.24	0.18	0.28	0.24	0.18
	1.00	0.70	0.70	0.45	0.20	0.18	0.12	0.53	0.48	0.35	0.54	0.49	0.36
	1.50	1.06	1.05	0.64	0.28	0.24	0.17	0.77	0.69	0.51	0.78	0.69	0.52
	2.00	1.40	1.41	0.79	0.33	0.29	0.19	1.00	0.87	0.65	0.98	0.86	0.64
B	0.15	0.07	0.07	0.03	0.01	0.01	-0.01	0.06	0.05	0.03	0.05	0.05	0.02
	0.25	0.12	0.12	0.06	0.02	0.02	0.00	0.08	0.07	0.05	0.09	0.07	0.05
	0.50	0.25	0.24	0.13	0.05	0.04	0.00	0.18	0.14	0.08	0.18	0.15	0.09
	1.00	0.50	0.50	0.25	0.11	0.08	0.01	0.34	0.28	0.16	0.35	0.29	0.17
	1.50	0.75	0.75	0.35	0.14	0.10	0.01	0.48	0.39	0.22	0.49	0.39	0.22
	2.00	0.98	0.98	0.43	0.18	0.13	0.02	0.62	0.51	0.29	0.61	0.50	0.28
C	0.15	0.11	0.05	0.04	0.03	0.02	0.00	0.04	0.03	0.02	0.05	0.04	0.03
	0.25	0.18	0.10	0.06	0.02	0.02	0.01	0.09	0.07	0.05	0.08	0.07	0.05
	0.50	0.35	0.19	0.12	0.04	0.03	0.02	0.16	0.14	0.10	0.17	0.15	0.10
	1.00	0.70	0.37	0.24	0.09	0.06	0.03	0.34	0.26	0.18	0.35	0.27	0.19
	1.50	1.06	0.54	0.34	0.11	0.08	0.02	0.49	0.38	0.27	0.49	0.38	0.27
	2.00	1.41	0.70	0.45	0.15	0.10	0.03	0.62	0.48	0.34	0.64	0.49	0.36
D	0.15	0.07	0.05	0.02	0.00	0.01	0.01	0.03	0.02	0.01	0.03	0.02	0.01
	0.25	0.13	0.09	0.06	0.02	0.02	0.00	0.06	0.06	0.03	0.07	0.06	0.04
	0.50	0.26	0.16	0.10	0.04	0.04	0.01	0.12	0.10	0.06	0.13	0.11	0.07
	1.00	0.51	0.30	0.20	0.09	0.07	0.02	0.26	0.22	0.15	0.26	0.22	0.15
	1.50	0.75	0.44	0.29	0.14	0.11	0.04	0.37	0.33	0.23	0.37	0.32	0.22
	2.00	1.02	0.59	0.40	0.16	0.12	0.05	0.47	0.42	0.29	0.48	0.42	0.29

Note: Shaded cells indicate a maximum effect size value for the row. Each row represents a unique process by max mean value

backward counting—when the date of next menstrual onset is known. For these two methods we assume accuracy of prior and later menstrual onset for forward and backward counting, respectively. We also estimated day of ovulation using a pseudo backward counting method where participants had a 50 % chance of accurately predicting next menstrual onset, a 40 % chance of incorrectly indicating next menstrual onset by  $\pm 1$  day, and a 10 % chance of incorrectly indicating next menstrual onset by  $\pm 2$  days. This method adds a relatively small error component to self-report or prediction of next menstrual onset—as is often observed in studies that evaluate discrepancies between expected and observed future menstrual onset dates (e.g. Wilcox, Dunson, & Baird, 2000).

Wood et al. (2014) and Gildersleeve et al. (2014a, b) came to different conclusions about the effect of fertile phase window width on statistical conclusions in the extant literature. By using very precise to less precise windows under controlled conditions we assess what effect, if any, fertile window width can have on statistical power. Three

different peak fertility sampling methods were applied using each of the fertility estimation method (e.g. forward and pseudo backward) ovulatory calendars: (a) the day of peak fertility (cycle day  $-1$ ); a random day from the six-day fertile window (cycle days  $-5$  through  $0$ ); and (c) a random day from a nine-day fertile window (cycle days  $-8$  through  $0$ )<sup>3</sup> (see Table 2). Thus, we have a degree of precision reflecting the day of greatest expected change, a random day from the entire six-day fertile window, and a random day from a nine-day fertile window to determine what influence fertility window size and estimation method have on power and effect size. To calculate a non-fertile

<sup>3</sup> Based on the estimation method used (i.e., pseudo backward counting) and the length of a cycle, it is possible that the earliest follicular cycle days may not extend to cycle day  $-5$  or  $-8$ . In these cases where a 6- or 9-day window was used, the cycle days sampled ranged from the earliest follicular cycle day (e.g., cycle day  $-6$ ) through cycle day  $0$ .

period, we selected from the mid-luteal phase. Specifically, we used cycle day 7 for each method of estimating ovulation.<sup>4</sup>

Additionally, we applied Wilcox et al.'s (2001) fertility estimation overlays—all, regularly, and irregularly cycling women—to each cycle, yielding a daily score of expected conception likelihood for each simulated cycle. Consistent with the predominant application of fertility overlays in between-subjects studies, these fertility overlays were treated in a forward-counting manner. That is, the first day of each cycle corresponded to the first day of the overlay, and so on. However, Wilcox et al. (2000) indicate that variability in average cycle length relates to variability in occurrence of fertile days. This suggests that aligning fertility overlays based on other estimates of fertility—rather than treating it as a consistent forward-counting estimate of fertility—may improve correspondence of fertility overlays and person-specific fertility—as well as fertility-related behavioral shifts. That is, if days of estimated fertility are aligned with estimated fertile days (e.g., using backward counting or LH testing), then the overlay may better correspond with actual fertility status. To test this we aligned overlays for all simulated cycles such that peak overlay fertility scores corresponded with the estimated day of peak fertility using forward, backward, and pseudo backward-counting methods, and true estimation.

**Procedure for testing effects** For between-subjects *t*-tests of sample size *N* (i.e., 10, 20, ..., 200), *N* random participants' peak fertility scores were drawn. Then an independent and random sample of *N* additional participants' mid-luteal scores was drawn. Mean phase differences (i.e., peak fertility compared to mid-luteal), pooled *SD*, effect size (Cohen's *d*), *t*-value, and *p*-value were calculated, then determination of whether the correct statistical decision was made. This process was repeated across all behavioral shift trajectories, ovulation estimation methods, and fertile window-width conditions. The within-subject *t*-test procedure was identical to the between-subject *t*-tests except that *N* random participants' peak fertility scores were drawn and matched to the participants' corresponding mid-luteal scores.<sup>5</sup> The procedure for testing

<sup>4</sup> Based on the estimation method used (i.e., forward counting) and the length of a cycle, it is possible that the estimated luteal cycle days may not extend to cycle day 7. In these cases the latest luteal day (e.g., cycle day 6) was used.

<sup>5</sup> Note that total sample size for between-conditions is actually twice that in within-conditions for sample size *N*. This suggests that for between- and within-samples *t*-tests to be matched in total sample size *N* = 10 for within-subjects corresponds to *N* = 20 for between-subjects. However, this match is not perfect either as *df* would still differ (e.g., when *N* = 20, *df* = 18 and 19 for between- and within-subject designs, respectively).

between-subject correlations using fertility overlays consisted of sampling a random cycle day from *N* randomly selected cycles. Cycle-day behavior scores were matched with corresponding fertility overlay scores. The resulting samples were tested using the Pearson correlation where  $r(N-2)$  served as the effect size. Similar to the *t*-test procedure, we concluded by determining whether the correct statistical decision was made for each test of correlation.

**Evaluating power using simulated data** For estimates of power using between- and within-subjects *t*-tests a correct statistical decision was defined using two criteria: (a) mean differences were in the same direction as the population (i.e.,  $M_{\text{HighFertility}} > M_{\text{LowFertility}}$ ); and (b) the effect was statistically significant using a one-tailed test ( $\alpha = .05$ ). All other results were scored as incorrect decisions. Power was defined as the number of correct decisions divided by the total number of tests.<sup>6</sup> Power estimates using correlations of fertility overlays were defined using the same criteria as *t*-tests with the exception that criterion (a) refers to correlation direction (i.e., positively related) rather than mean differences and criterion (b) uses two-tailed tests (i.e.,  $\alpha = .05$  split above and below the estimated correlation value).<sup>7</sup>

### Empirical data

A publicly available dataset was used in the present study (Fehring, Schneider, Raviele, Rodriguez, & Pruszyński, 2013). In the present study, cycle inclusion from the public data set was contingent upon whether participants had been assigned to monitor fertility status using Clear Blue Easy Fertility Monitors (CBFM), and whether there was sufficient ovulatory cycle event data to characterize each cycle; i.e., CBFM indicated day of peak fertility, estimated day of ovulation, and both cycle onset and offset were recorded. These criteria resulted in the inclusion of 91 women's cycles, each with 1–43 ovulatory cycles ( $N_{\text{cycles}} = 906$ ,  $M_{\text{cycles per woman}} = 9.98$ ,  $SD_{\text{cycles per woman}} = 7.37$ ). For a full description of the study we refer readers to the original study and data hosting site (Fehring et al., 2013; e-Publications@Marquette, 2012).

<sup>6</sup> We had 6,000 replications for each *t*-test condition, and 125,000 replications for each overlay condition. The discrepancy in the number of replications was based on necessity to sufficiently represent sampling between cycles and cycle days using the overlay method. Fewer replications (i.e., 6,000) showed underrepresentation of both these dimensions with respect to the simulated data.

<sup>7</sup> For null conditions any significant effect was counted as a Type I error, non-significant effects were considered to be correct statistical decisions.

**Estimation of fertility** Estimates of ovulation based on CBFM results were used to determine six-day periods of peak fertility (e.g., Dunson et al., 1999, 2002; Wilcox et al., 1995, 1998) which were then used as a comparison to forward and confirmed backward counting methods.<sup>8</sup> These estimation methods were then compared with CBFM by determining the frequency with which each method corresponded with CBFM indicated days of peak fertility. These comparisons included when considering capturing the day of expected peak fertility and any of the six days of peak fertility. This was considered for each estimation method when using the predicted day of peak fertility, a six-day window of predicted peak fertility, and a nine-day window of peak fertility.

## Results

### Predicting power

A total of 11,520 power estimates were generated for between- ( $n = 5,760$ ,  $M = .44$ ,  $SD = .36$ ) and within- ( $n = 5,760$ ,  $M = .51$ ,  $SD = .37$ ) subject  $t$ -tests. A  $t$ -test was used to determine that, generally, within-subjects designs, compared to between-subjects designs, had greater power to detect fertility-based differences ( $t(11518) = 11.17$ ,  $p < .001$ ,  $d = .21$ ).

To test and control other simulation conditions a Factorial ANCOVA model was fit using type II sums of squares and the following additional predictors: (a) six-day and nine-day windows compared against day of predicted peak fertility; (b) pseudo backward, confirmed backward, forward counting and true estimation; (c) sample size; (d) a four-way interaction of within- versus between-subjects designs, fertility estimation method, window size, and sample size as well as all three-way and two-way interactions of these terms; and (e) average effect size as a control. This second model (see Table 3) explained a significant proportion of variance in power ( $F(48, 11,471) = 710.1$ ,  $p < .0001$ ,  $R^2 = .748$ ). It is important to note that in this second model the increase in power when using within-subjects tests ( $B = .10$ ,  $SE = .02$ ,  $p < .001$ ) is consistent with our initial  $t$ -test results—after accounting for other simulation conditions.

Dummy-coded contrasts were used to investigate the interaction between fertility estimation method and within- versus between-subjects designs. Results indicate that the interaction was driven by a reduction in within-subjects designs' power, relative to between-subjects designs, when comparing forward counting with both pseudo backward- and backward-counting fertility estimation methods ( $B = -.07$  and  $-.07$ ,  $SE = .035$  and  $.35$ ,  $p = .046$  and  $.047$ , respectively)—also note

<sup>8</sup> Pseudo backward counting could not be adequately compared as participant estimation of menstrual onset was not a component of the study and its data.

that within-subjects designs were still more powerful than between-subjects designs, despite this reduction. This suggests that the benefit of using within-subjects designs is reduced when coupled with forward-counting estimation, but this effect was only significant when comparing forward- and backward-counting methods.

To understand the interaction between fertility estimation method and fertility window size, data was subset by fertility estimation type and a one-way ANOVA was conducted for each subset where power was predicted by window size. Results indicate that fertility window size was a significant predictor of power when using forward ( $F(2, 2,877) = 16.26$ ,  $p < .001$ ,  $R^2 = .01$ ), pseudo backward ( $F(2, 2,877) = 4.87$ ,  $p < .01$ ,  $R^2 = .003$ ), backward ( $F(2, 2,877) = 4.76$ ,  $p < .01$ ,  $R^2 = .003$ ), and true ovulation estimation methods ( $F(2, 2,877) = 75.01$ ,  $p < .001$ ,  $R^2 = .05$ ). It is evident that the effect of window size was larger for true ovulation estimation relative to the other estimation procedures. Pairwise comparisons using the Bonferroni correction were used to test if there were specific differences in the pattern of window effects within each ovulation estimation method. Results indicated several key conclusions: (a) 1-day estimates of peak fertility yielded significantly more power than either six-day or nine-day windows; (b) there is a significant difference in power between six-day ( $M = .63$ ,  $SD = .37$ ) and nine-day ( $M = .50$ ,  $SD = .37$ ) windows for true ovulation estimation ( $t(1,918) = 7.71$ ,  $p < .001$ ), but not for any of the other ovulation estimation methods; and (c) effect size patterns were larger for the true ovulation estimation condition relative to all other estimation methods, and for forward counting estimation relative to pseudo backward and backward counting (see Table 4).

Investigation of the interaction between sample size and fertility estimation method was completed using dummy-coded estimation variables. The interaction effect was driven by a difference in the rate of change in power as a function of sample size between pseudo backward counting ( $B = .007$ ,  $SE = .002$ ,  $p < .001$ ), backward counting ( $B = .007$ ,  $SE = .002$ ,  $p < .001$ ), and true ovulation ( $B = .004$ ,  $SE = .002$ ,  $p = .04$ ) with forward counting. These results indicate that power increases faster for pseudo backward, backward, and true ovulation methods. For example, with sample sizes of 20, 50, and 100 pseudo backward and backward (additional power = .14, .35, .70) and true ovulation (additional power = .08, .20, .40) would have more power relative to forward estimation.

### Comparing estimation methods

To understand differences in power between estimation methods we evaluated the frequency that peak fertility days, defined by true ovulation, were identified using different fertility estimation counting methods when applied to both the simulated population data and the Fehring et al. (2013) data (see Table 5). Comparisons were made with respect to correct

**Table 3** Factorial ANCOVA table predicting power to detect ovulatory effects

Conditions	Type II SS	df	F-value	p-value
Within	20.04	1.00	577.30	< .001
Fertility estimation method	22.18	3.00	212.91	< .001
Window size	1.04	2.00	14.96	< .001
N	111.93	1.00	3,223.71	< .001
Effect size	773.82	1.00	22,285.99	< .001
Within:fertility estimation method	0.66	3.00	6.29	< .001
Within>window size	0.01	2.00	0.17	0.84
Fertility estimation method: window size	5.27	6.00	25.31	< .001
Within:N	0.02	1.00	0.72	0.40
Fertility estimation method: n	3.73	3.00	35.77	< .001
Window size: n	0.02	2.00	0.30	0.74
Within: fertility estimation method: window size	0.11	6.00	0.51	0.80
Within: fertility estimation method: n	0.23	3.00	2.20	0.09
Within: window size: n	0.00	2.00	0.01	0.99
Fertility estimation method: window size: n	0.42	6.00	2.01	0.06
Within: fertility estimation method: window size: n	0.01	6.00	0.04	1.00
Residuals	398.30	11,471.00		

Note: Type II sums of squares were used to account for covariance between predictors regardless of predictor entry into the model

and incorrect selection of the day of peak fertility, and then of any six fertile days in a cycle. Correct percentages were based on how many target (i.e., true fertile) days were captured divided by the total possible (e.g., 20,000 possible days of peak fertility in the simulated data). Incorrect percentages were based on how many non-target days were captured divided by the total number of captured days (e.g., 180,000 total days when using a nine-day window). To evaluate efficacy, the one-day predicted peak fertility and both six- and nine-day windows of peak fertility were used.

**Table 4** Effect sizes comparing power between window size conditions within fertility estimation methods

Fertility estimation method	1-day vs. 6-day	1-day vs. 9-day	6-day vs. 9-day
Forward counting	0.219	0.220	0.001
Pseudo backward counting	0.123	0.124	0.001
Backward counting	0.121	0.122	0.001
True ovulation	0.195	0.555	0.352

Note: Computed effects sizes are Cohen’s d. For forward, pseudo backward, and backward methods, there was essentially no difference between 6-day and 9-day windows. For true ovulation, there was an effect, and this effect was larger than any of the calculated effects with respect to detected differences between 1-day and either 6- or 9-day windows for forward, pseudo backward, and backward counting methods. Generally, true ovulation effects were larger than the other methods, suggesting greater sensitivity of true ovulation with respect to window size effects. Similarly, forward counting demonstrated greater sensitivity to window size effects with respect to pseudo backward and backward counting, though this was not as pronounced

Generally, accuracy at predicting the true day of peak fertility was very poor when using the one-day and six-day windows. While inclusion of the true day of peak fertility did increase considerably using the six-day window, these estimates were accompanied by a comparable proportion of incorrectly identified days (compared with the one-day window). However, it should be noted that when considering the number of total true fertile days (six days total), the error rates drop by 30–50 %; error declined greatest for pseudo backward and backward. Unsurprisingly, moving to a nine-day window necessarily caused both correct target identification and non-target identification percentages to increase relative to the six-day window. Forward counting consistently underperformed compared to both backward methods, and pseudo backward slightly underperformed relative to backward.

**Efficacy of fertility estimation overlays**

Generally, the fertility overlay for all participants (i.e., both regularly and irregularly cycling participants) yielded the largest effect sizes, but these gains were minor with respect to the overlay for regularly cycling participants (see Table 6). Nonetheless, we focus our interpretation on the overlay that had the best ability to detect the effect of fertility (i.e., the overlay for all participants). Across all processes, power only exceeded 80 % for the greatest maximum (max) mean condition ( $M_{max} = 2.0$ ) for Processes A and B (8.3 % of conditions; see Fig. 2). Even at a reduced power threshold of 50 %, only seven (29.2 %) of the considered conditions (i.e.,  $M_{max} = 2.0$



**Table 5** Percent of true fertile days captured using different fertility windows and estimation methods

Condition	Simulated data		Fehring et al. (2013) data	
	Correct %	Incorrect %	Correct %	Incorrect %
Forward: 1-day in 1-day	8.10	91.91	13.25	86.76
Pseudo back: 1-day in 1-day	14.77	85.24	-	-
Backward: 1-day in 1-day	15.06	84.94	20.09	79.91
Forward: 1-day in 6-day	50.17	81.73	66.89	80.56
Pseudo back: 1-day in 6-day	75.87	81.63	-	-
Backward: 1-day in 6-day	79.85	81.46	90.95	73.57
Forward: 6-day in 6-day	45.76	54.24	57.36	42.64
Pseudo back: 6-day in 6-day	68.85	31.15	-	-
Backward: 6-day in 6-day	71.78	28.22	72.94	27.06
Forward: 6-day in 9-day	54.24	63.84	63.32	57.79
Pseudo back: 6-day in 9-day	79.85	46.69	-	-
Backward: 6-day in 9-day	81.76	45.45	76.67	48.88

*Note:* The Correct % 1-day in 1-day indicates the percent of true days of peak fertility that fell within the estimation method's predicted day of peak fertility; 1-day in 6-day indicates the percent of true days of peak fertility fell within the 6-day window of the estimation method; 6-day in 6-day indicates the percent of true fertile days (six per cycle) that fell within the 6-day window of the estimation method; and 6-day in 9-day indicates the percent of true fertile days (six per cycle) that fell within the 6-day window of the estimation method. The Incorrect % corresponds with the Correct %; however, these scores represent the percent of days incorrectly selected as corresponding with fertility. Pseudo backward percentages could not be estimated using empirical data as sufficient information for those estimates were not a part of the study and resulting data set

for Processes A–D,  $M_{\max} = 1.5$  for Processes A and B, and  $M_{\max} = 1.0$  for Process A; see Fig. 3) met or exceeded the threshold. This means that for 70.8 % of the conditions considered, detection of the true effect would be less likely than a coin toss, even with a sample of 200 participants.

This can be better understood by considering the average effects observed within individuals. A second set of analyses considered the relation of fertility overlay values with simulated behavioral shifts for each overlay type, effect size, and

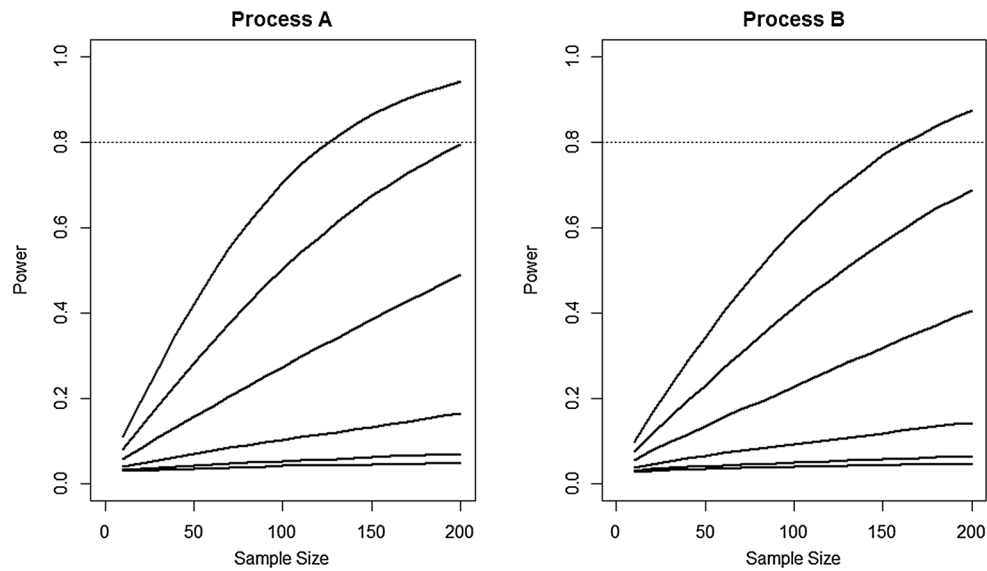
process type (see Table 7). Average cycle effects and power were relatively consistent across overlay types, but as with the between-cycle results, effect size and power diminished with process complexity. Even when power was greatest (i.e., Process A), the effect was not successfully detected in just over 50 % of the cycles—this disparity maximized at 78–80 % for Processes C and D.

As previously mentioned, it may be beneficial to consider a cycle-centered approach with respect to overlay use. That is, if

**Table 6** Average Pearson correlation values by maximum difference, process type, and fertility overlay

Process	$M_{\max}$	All	Regular	Irregular	Process	$M_{\max}$	All	Regular	Irregular
A	0.15	0.020	0.020	0.015	C	0.15	0.012	0.012	0.011
	0.25	0.033	0.033	0.025		0.25	0.018	0.017	0.016
	0.50	0.069	0.067	0.050		0.50	0.038	0.036	0.034
	1.00	0.134	0.131	0.099		1.00	0.075	0.072	0.068
	1.50	0.193	0.189	0.141		1.50	0.111	0.107	0.099
	2.00	0.245	0.240	0.178		2.00	0.145	0.139	0.130
B	0.15	0.019	0.018	0.020	D	0.15	0.010	0.010	0.010
	0.25	0.030	0.028	0.031		0.25	0.020	0.019	0.019
	0.50	0.061	0.058	0.064		0.50	0.039	0.038	0.035
	1.00	0.118	0.111	0.119		1.00	0.078	0.075	0.070
	1.50	0.168	0.159	0.170		1.50	0.115	0.110	0.104
	2.00	0.212	0.201	0.216		2.00	0.151	0.145	0.135
Null	0.00	0.000	0.000	0.000					

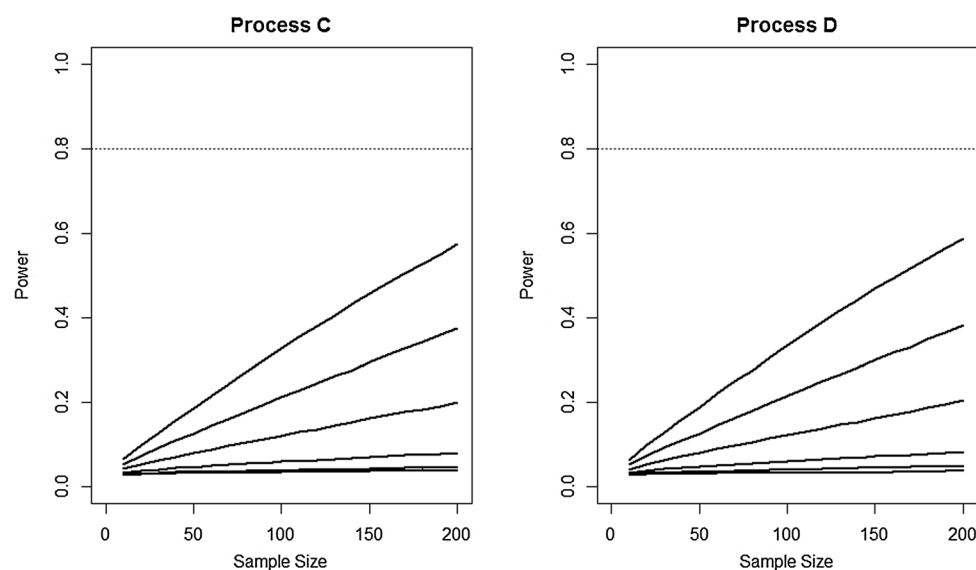
*Note:* Note that effect size generally increased with sample size, but decreased with process complexity. Also, while the fertility overlay for both regular and irregular cycles was best related to ovulatory fluctuations, the overlay for regular cycles performed nearly as well



**Fig. 2** Power trajectories for Processes A–B using Wilcox et al.’s (2001) fertility overlay for all participants across sample size. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

individual cycle-estimated fertility is used to center overlays then overlay efficacy (e.g., power) may increase. This approach was tested with the overlay for both regular and irregular cycles, using forward, pseudo backward, backward, and true ovulation to center overlays. Improvements in power across estimation types are striking. For Processes A and B,  $M_{\max}$  values of 2.0 and 1.5 exceed 80 % power with a sample size of 200 for pseudo backward, backward, and true ovulation estimation. Given a  $M_{\max}$  value of 2.0, forward estimation also exceeds 80 % with a sample size of 200 for Processes A and B (see Figs. 4, 5, 6, and 7).

Similarly, for Processes C and D, increased power has resulted in exceeding the 80 % power threshold for the  $M_{\max}$  value of 2.0 using both backward and true ovulation (see Figs. 8, 9, 10, and 11). Further, 45.8 % of true ovulation conditions, 41.7 % of pseudo backward and backward conditions, and 25 % of forward conditions now exceed 50 % power given a sample size of 200. While these represent marked improvements in power, the number of conditions that can achieve power of 80 % given a maximum sample size of 200 is, at best, almost 50 %, and is favored by the largest effect and sample sizes considered.

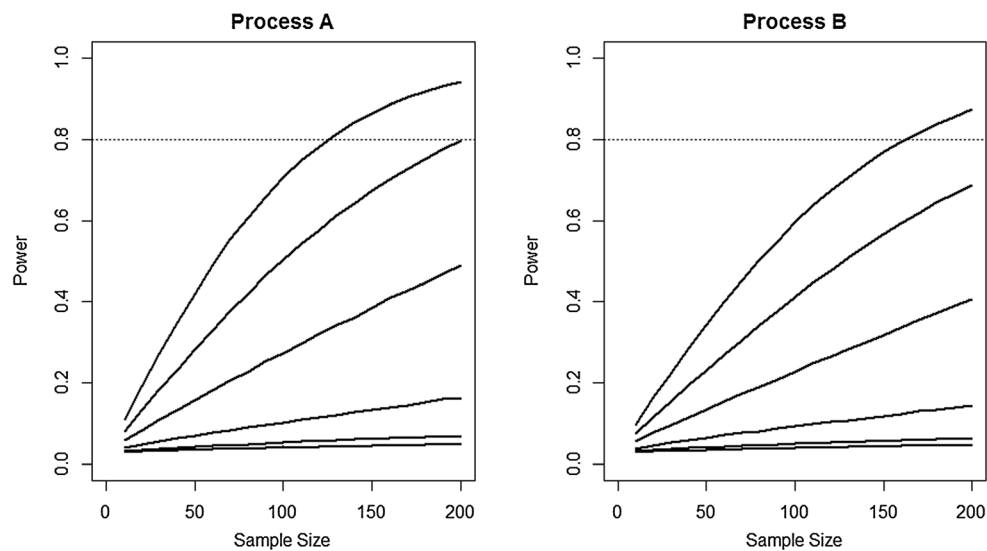


**Fig. 3** Power trajectories for Processes C–D using Wilcox et al.’s (2001) fertility overlay for all participants across sample size. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

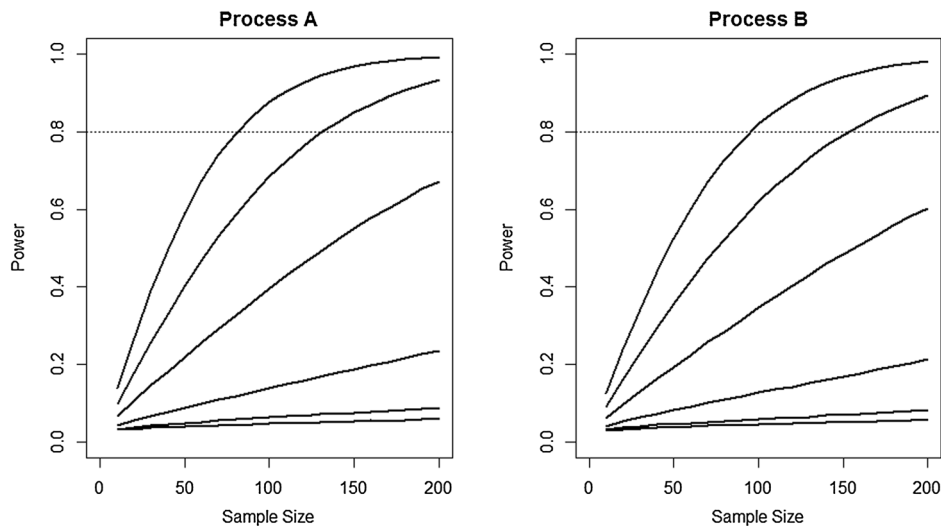
**Table 7** Average within-cycle correlation values and power to detect effects using fertility overlays

	Wilcox fertility overlays									
	All				Regular			Irregular		
Process	$M_{max}$	$M_{cor}$	$SD$	$Power$	$M_{cor}$	$SD$	$Power$	$M_{cor}$	$SD$	$Power$
A	0.15	0.03	0.19	0.03	0.03	0.19	0.03	0.02	0.19	0.03
	0.25	0.05	0.19	0.04	0.05	0.19	0.04	0.03	0.19	0.04
	0.50	0.10	0.19	0.08	0.10	0.20	0.08	0.06	0.20	0.07
	1.00	0.19	0.20	0.19	0.18	0.21	0.19	0.13	0.23	0.16
	1.50	0.26	0.21	0.33	0.25	0.22	0.33	0.17	0.27	0.29
	2.00	0.31	0.23	0.45	0.31	0.23	0.44	0.21	0.30	0.41
B	0.15	0.03	0.19	0.03	0.03	0.19	0.03	0.03	0.19	0.03
	0.25	0.04	0.19	0.04	0.04	0.19	0.04	0.04	0.19	0.04
	0.50	0.09	0.19	0.07	0.08	0.19	0.07	0.09	0.19	0.07
	1.00	0.16	0.20	0.16	0.16	0.21	0.16	0.16	0.21	0.17
	1.50	0.23	0.21	0.27	0.22	0.22	0.26	0.22	0.23	0.30
	2.00	0.28	0.23	0.38	0.26	0.24	0.36	0.27	0.25	0.42
C	0.15	0.02	0.19	0.03	0.02	0.19	0.03	0.02	0.19	0.03
	0.25	0.03	0.19	0.04	0.03	0.19	0.03	0.02	0.19	0.03
	0.50	0.06	0.19	0.05	0.05	0.19	0.05	0.05	0.19	0.05
	1.00	0.11	0.20	0.09	0.11	0.20	0.09	0.09	0.20	0.08
	1.50	0.16	0.20	0.15	0.15	0.21	0.15	0.13	0.21	0.13
	2.00	0.20	0.21	0.22	0.19	0.22	0.22	0.17	0.22	0.20
D	0.15	0.02	0.19	0.03	0.01	0.19	0.03	0.01	0.19	0.03
	0.25	0.03	0.19	0.04	0.03	0.19	0.04	0.03	0.19	0.04
	0.50	0.06	0.19	0.05	0.05	0.19	0.05	0.05	0.19	0.05
	1.00	0.11	0.19	0.09	0.11	0.19	0.08	0.09	0.20	0.09
	1.50	0.16	0.20	0.14	0.15	0.20	0.14	0.14	0.22	0.15
	2.00	0.21	0.20	0.22	0.20	0.20	0.21	0.17	0.23	0.23
Null	0.00	0.00	0.19	0.05	0.00	0.19	0.05	0.00	0.19	0.05

Note:  $M_{cor}$  is the average within-person Pearson correlation between overlay fertility and simulated behavioral data across all 20,000 simulated cycles.  $SD$  describes between-cycle variability in correlation values. Power is the proportion of within-cycle correlations (out of 20,000) that detected the true effect



**Fig. 4** Power trajectories for Processes A–B using Wilcox et al.'s (2001) fertility overlay for all participants centered using forward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0



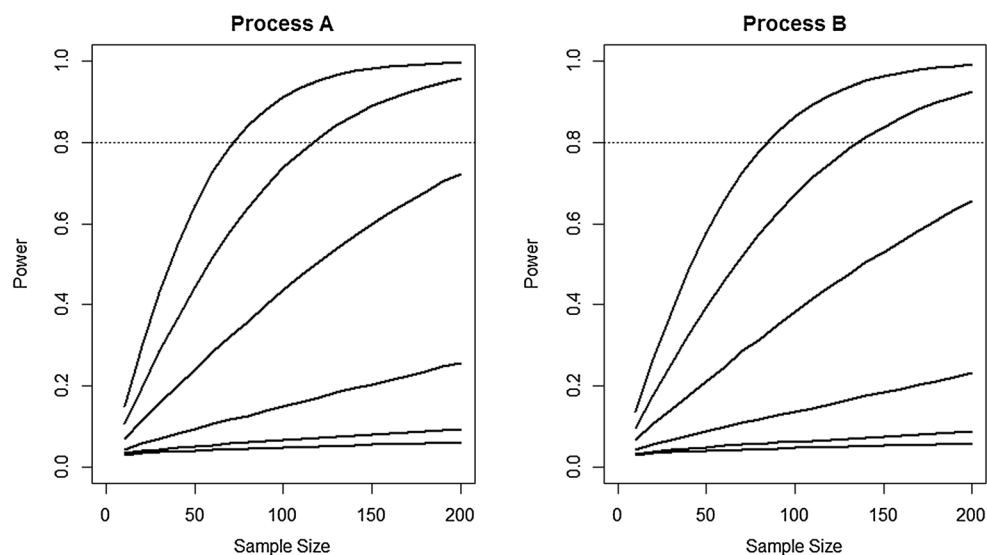
**Fig. 5** Power trajectories for Processes A–B using Wilcox et al.'s (2001) fertility overlay for all participants centered using pseudo backward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

## Discussion

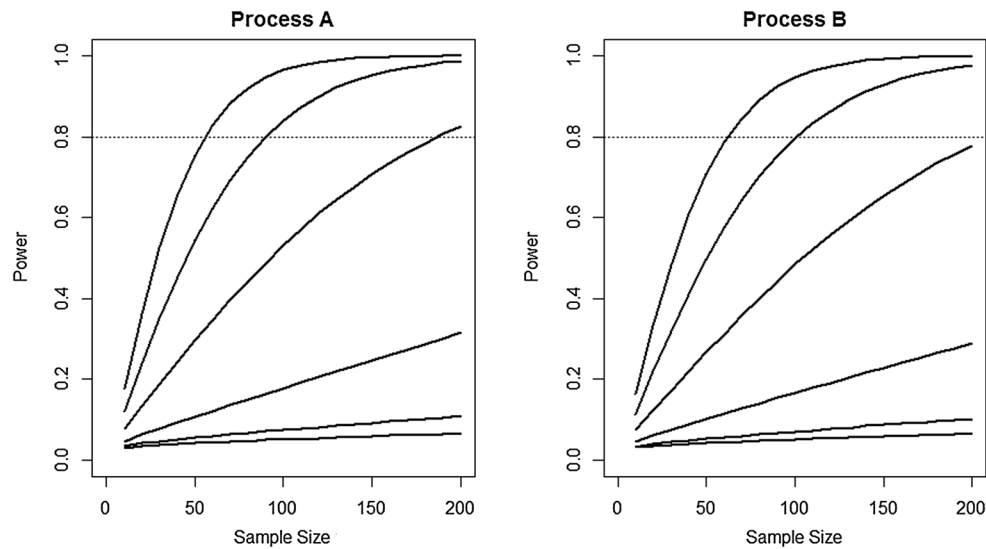
### Summary

In the present study we used a simulated population of ovulatory cycles and an empirical data set to demonstrate significant discrepancies between fertility overlays and counting methods when compared with more accurate hormonal measures to detect ovulation (e.g., Bullivant et al., 2004; Fehring et al., 2013; Lloyd & Coulam, 1989; Tanabe et al., 2001; Trussel, 2008). Furthermore, we demonstrated that between-subject methods (i.e., independent samples *t*-tests and fertility overlays) showed significantly less power to detect effects present in the population when compared to within-subjects approaches.

Power disparities between counting methods and true ovulation, with respect to detection of ovulatory effects, are likely due to differences in days identified as fertile. As error rates in fertile day identification increased, estimates of mean differences tended to be less than those observed at the population level. This was particularly true for curvilinear processes C and D because their apex corresponds with the population level maximum change in simulated behavior. In contrast, because processes A and B had stable differences across all fertile days, errors in fertile day identification did not diminish estimates of population mean differences to the same degree. As a result, while pseudo backwards and backwards estimation was underpowered compared to true ovulation, these differences were not as large across most sample sizes and



**Fig. 6** Power trajectories for Processes A–B using Wilcox et al.'s (2001) fertility overlay for all participants centered using backward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0



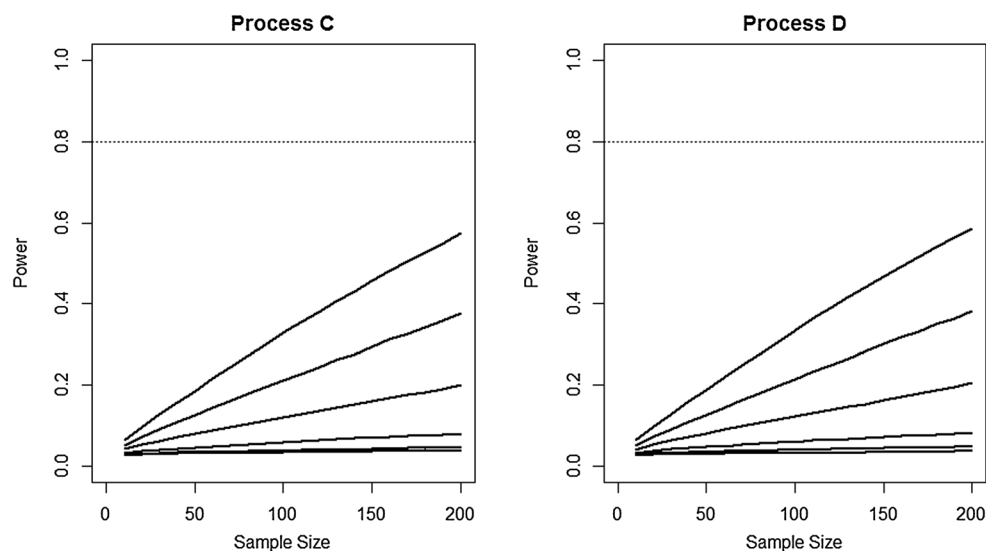
**Fig. 7** Power trajectories for Processes A–B using Wilcox et al.’s (2001) fertility overlay for all participants centered using true peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

maximum difference values when compared with discrepancies observed when using forward counting.

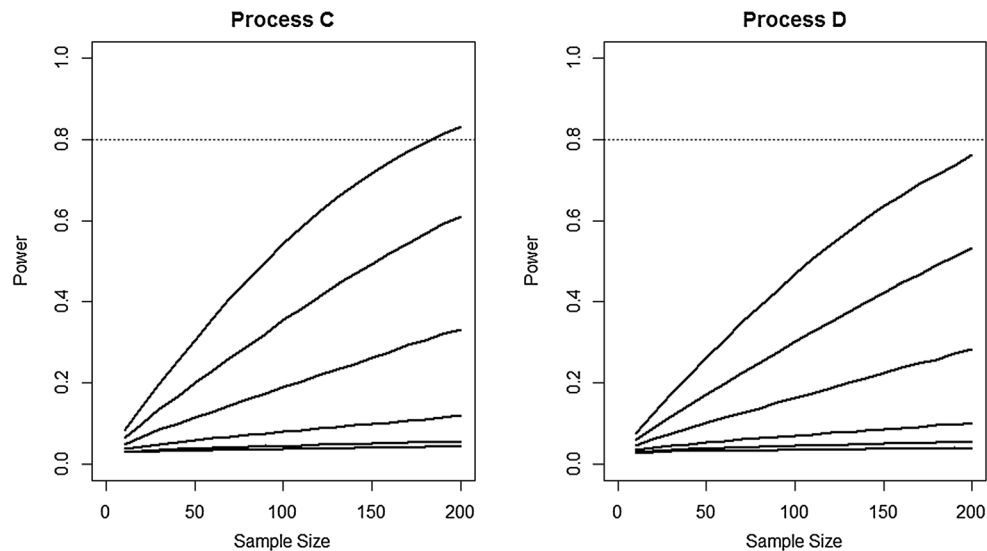
Differences in fertile day estimation using forward and backward counting methods compared with hormonal detection were replicated in our empirical example. We observed that both forward and backward counting methods differed in the days identified as fertile when compared to hormonal detection. These differences were consistent with those from our simulated data set, which was derived from data describing a different empirical sample (i.e., Fehring et al., 2006). Taken together, this lends support to the veracity of our simulated data, and conclusions drawn from it with respect to power estimates.

### Implications

Several implications can be drawn from these findings; one of the most evident is that within-subjects  $t$ -tests are preferable when studying ovulatory effects compared with between-subjects  $t$ -tests. This is likely due to the inability to separate between-subject variability in scores from between-phase (high fertility vs. low fertility) differences when using between-subjects designs. While these conclusions do not contradict any typically expressed expectations, what is interesting about them is that with an average predicted increase in power of .1007 when using a within-subjects design, and a predicted increase in power of .0019 for each participant



**Fig. 8** Power trajectories for Processes C–D using Wilcox et al.’s (2001) fertility overlay for all participants centered using forward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

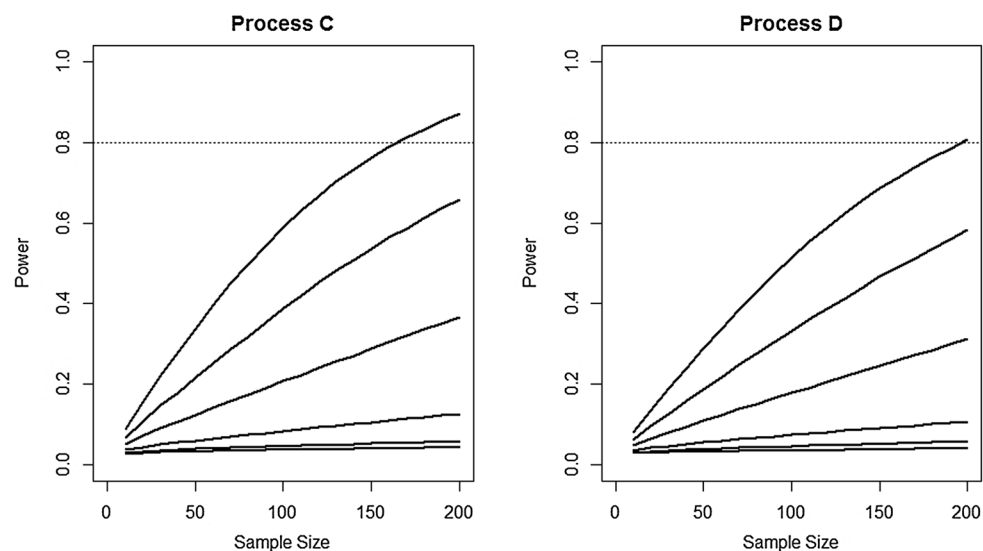


**Fig. 9** Power trajectories for Processes C–D using Wilcox et al.’s (2001) fertility overlay for all participants centered using pseudo backward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

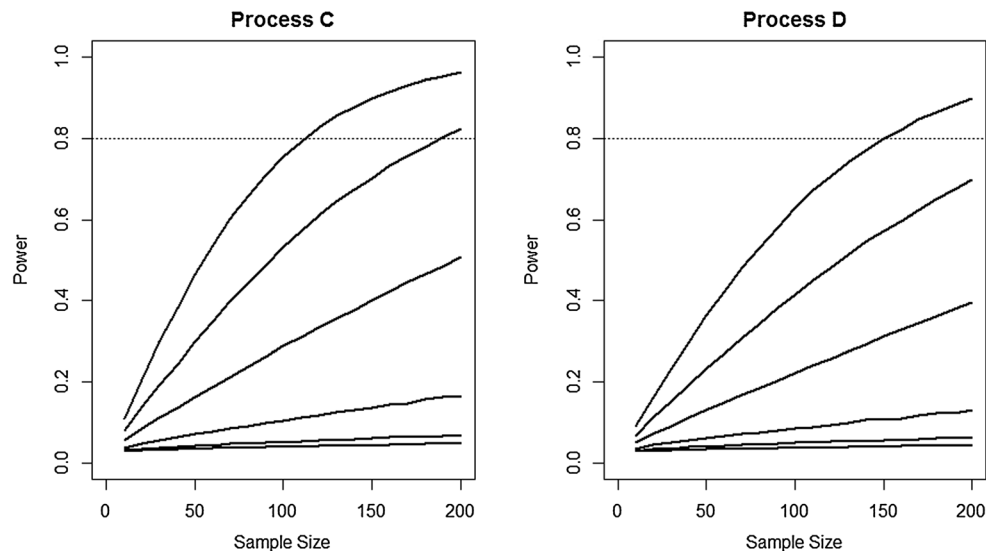
sampled, we would predict needing approximately 53 additional participants to achieve comparable power using a between-subjects design using hormonal, backward counting, or pseudo backward counting estimation—assuming self-report error in pseudo backward counting was not worse than that used in our simulation study. It should be noted that due to the significant interaction between fertility estimation methods and sample size it would take an additional 30 participants (in addition to the initial 53) to achieve equal power for a between-subjects design if using forward counting estimation.

Another implication is that, whereas backwards and pseudo backwards counting estimation outperformed forward

counting estimation of ovulation, they appear to be inferior methods compared to hormonal assessments. While it may be argued that the inferiority of backwards estimation compared with other assessment methods is already known within the field (e.g., Bullivant et al., 2004; Debruine et al., 2010), the present state of the supporting literature suggests it is very popular (particularly pseudo backward counting; e.g., Gildersleeve et al., 2014a, b; Harris, 2013). Although it is true that larger and larger sample sizes would diminish power differences due to estimation method, the average sample size employed in a sample of the extant literature is a heavily skewed 95 participants ( $Mdn = 50$ ,  $SE = 37.4$ ; i.e., Gildersleeve et al., 2014a, b). Therefore, while larger sample



**Fig. 10** Power trajectories for Processes C–D using Wilcox et al.’s (2001) fertility overlay for all participants centered using backward counting estimated peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0



**Fig. 11** Power trajectories for Processes C–D using Wilcox et al.’s (2001) fertility overlay for all participants centered using true peak fertility. Power trajectories from the bottom to the top correspond with  $M_{\max}$  values of 0.15, 0.25, 0.5, 1.0, 1.5, and 2.0

sizes could diminish concerns over estimation effects on power, those studies employing more typical sampling procedures should take these effects into account during study design.

An important issue that has been raised is whether fertility window size may have an influence on the likelihood of detecting effects. Our simulated results suggest that fertility window size does have several effects. Most consistently, that using the single day of estimated peak fertility (i.e., cycle day  $-1$ ) resulted in greater power to detect effects than using days sampled from either a six- or nine-day window, regardless of estimation method. Also, while differences between six- and nine-day windows were not detected using any of the counting methods, a difference was detected when using true ovulation. This is likely due to the ratio of target days identified compared to the number of non-target days identified. That is, while the six-day window identifies slightly fewer target days than the nine-day window, it also identifies fewer non-target days than the nine-day window. These ratios appear to average out, suggesting that both six-day and nine-day windows are, generally, going to yield similar results when using counting methods. This, however, has less to do with increased accuracy, and more to do with averaged error to accuracy ratios in identifying fertile days.

Finally, our simulated results suggest that fertility overlays, though popular, perform very poorly with respect to both effect size and power. Based on these results, we have to disagree with recommendations that researchers use fertility overlays as a method for estimating fertility (Gildersleeve et al., 2014a). While using overlays may help alleviate concerns of researcher degrees of freedom abuse due to inconsistent definitions of fertility and fertility window size (e.g., Harris, 2013), such an approach may substantially impair power to detect effects. However, if research design imposes

limitations preventing other methods of assessment (e.g. LH testing), it is strongly recommended that researchers maximize potential power by at least following up with participants regarding onset of menstruation so that backwards estimation can be used to center overlay fertility values with participants’ observed data, or so that researchers can forego the use of fertility overlays in favor of more accurate identification of high and low fertility days.

### Limitations and concerns

One possible limitation of our study is that we simulated ovulatory cycles using descriptive statistics and covariances from a study that used hormonal assessment methods to determine occurrence of ovulation. As a result, any error inherent in that estimation method may have been passed on to our simulated data. However, there is substantial evidence suggesting that hormonal estimation of ovulation is highly accurate (Behre et al., 2000; Lloyd & Coulam, 1989; Tanabe et al., 2001). This suggests that estimation inaccuracy was likely low in the empirical data informing this study (Fehring et al., 2006; Fehring et al., 2013), and that hormone-based estimates of fertility are a reasonable proxy of true ovulation for comparison with counting methods.

Another possible limitation is the process trajectories considered in the present study. These were selected for two principal reasons. One, they demonstrate processes that are either solely dependent on fertility (Processes A and C), or that are associated with co-occurring processes across the ovulatory cycle such as hormone level fluctuations (Processes B and D; e.g., Lukaszewski & Roney, 2009; Roney & Simmons, 2008). Two, they demonstrate varying degrees of complexity ranging from a simple, constant, step-wise effect during

periods where change occurs (Processes A and B) to a curvilinear trajectory (Processes C and D) similar to those in reports of ovulatory effects on behavior (e.g., Miller et al., 2007; Roney & Simmons, 2008). While we cannot simply assume that any of these simulated trajectories are perfectly representative of specific behavioral processes, they do demonstrate how representative processes varying in complexity affect power to detect ovulatory effects, and can help inform researchers as they develop future research designs.

One of the estimation methods was pseudo backward counting. Pseudo backward counting is commonly used in lieu of true backward estimation. Rather than confirming menstrual onset and counting backwards, researchers use a participant's estimate of their next menstrual onset and count back from that date. The problem, as previously discussed, is that many factors may cause a participant to be inaccurate in their estimation of menstrual onset—even if they consider themselves to be regularly cycling (e.g., Wilcox et al., 2000). However, this problem is not studied as well as some others. As a result, we erred on the side of caution in developing our probability of participant error in predicting their next menstrual onset, such that error rates were approximately normally distributed around the correct date, with error being no more extreme than  $\pm 2$  days. The rationale behind this was to be as consistent as possible in defining the cycles as regular for participants (though this is rarely confirmed). Using this relatively small rate of error, pseudo backward and backward estimation resulted in similar efficacy.

However, if the error rate was increased in relation to frequency of error and degree of error then it is easy to demonstrate how pseudo backward estimation could be as ineffective, if not worse, than forward counting estimation. For example, if we modify the simulation procedure used to generate the previously reported pseudo backward days to have a 30 % chance of correct prediction, 22.2 % chance of  $\pm 1$  day, 17.8 % chance of  $\pm 2$  days, 13.3 % chance of  $\pm 3$  days, 8.8 % chance of  $\pm 4$  days, and 4.4 % chance of  $\pm 5$  days then rates of correct

identification and incorrect identification are comparable, if not a little worse, than forward counting (see Table 8). Similarly, when using empirically demonstrated prediction error proportions based on average cycle length (i.e., Creinin et al., 2004) pseudo backward counting performs on par with forward counting. Clearly, pseudo backward estimation is heavily dependent on accuracy of menstrual onset prediction. As a consequence, pseudo backward estimation should not be assumed to be as effective as backward counting given that without menstrual onset confirmation there is no way to be certain that the correct date to count back from was identified. Further, while backward counting appears to be more accurate than forward counting because the luteal phase is more stable than the follicular phase, it is important to recall that it is still variable.

Another potential limitation is that for *t*-tests, while we focused on different fertile windows as tenable for sampling high fertility days, we only sampled from the same luteal phase day (ovulatory calendar day 7 or day 21), or its closest neighbor. However, like high fertility day sampling, low fertility day sampling methodology is quite variable and ranges from a target day in the luteal phase (e.g., Cárdenas & Harris, 2007), a range of days or all days in the luteal phase (e.g., Caryl et al., 2009; Moore et al., 2011; Morrison, Clark, Gralewski, Campbell, & Penton-Voak, 2010; Pawlowski & Jasienska, 2005), and all days outside the high fertility window (e.g., Little, Jones, & Burriss, 2007; Little, Jones, & DeBruine, 2008; Penton-Voak & Perrett, 2000; Puts, 2005; Vaughn, Bradley, Byrd-Craven, & Kennison, 2010). In terms of error in identifying high fertility days as low fertility days, these methods did not differ substantially in the present study (see Table 9), and we have focused on reporting results when using the luteal sampling method that yielded the lowest overall error rate.

It is worth noting, though perhaps self-evident from the trajectories that were considered, that selecting different low fertility sampling methods would have little impact on effect

**Table 8** Accuracy of pseudo backward estimation using alternative error probabilities in simulation

Condition	Increased prediction error		Creinin et al. (2004) error	
	Correct %	Incorrect %	Correct %	Incorrect %
Pseudo Back: 1-day in 1-day	12.56	87.45	10.59	89.41
Pseudo Back: 1-day in 6-day	45.68	92.39	44.93	92.51
Pseudo Back: 6-day in 6-day	45.69	54.32	38.39	61.61
Pseudo Back: 6-day in 9-day	38.83	61.17	36.33	63.67

*Note:* The Correct % 1-day in 1-day indicates the percent of true days of peak fertility that fell within the estimation method's predicted day of peak fertility; 1-day in 6-day indicates the percent of true days of peak fertility fell within the 6-day window of the estimation method; 6-day in 6-day indicates the percent of true fertile days (six per cycle) that fell within the 6-day window of the estimation method; and 6-day in 9-day indicates the percent of true fertile days (six per cycle) that fell within the 6-day window of the estimation method. The Incorrect % corresponds with the Correct %; however, these scores represent the percent of days incorrectly selected as corresponding with fertility



**Table 9** Percent of errors made using different low fertility sampling methods

Low fertility sampling strategy	Forward %	Pseudo backward %	Backward %
Day 7 (or closest neighbor) of luteal phase	1.7	0	0
All luteal days	1.6	0.9	0.8
All outside 6-day fertile window	1.4	0.9	0.8
All outside 9-day fertile window	1.2	0.7	0.6

*Note:* All values are percents. Forward, Pseudo backward, and Backward indicate the ovulation estimation method being used in connotation with the low fertility sampling strategy. Low fertility sampling strategies consisted of sampling from day 7 of the ovulatory calendar (i.e., -13 through 14), all luteal days (i.e., days 1–14), all days outside of a 6-day fertile window (i.e., days -13 through -6 and 1 through 14), and all days outside of a 9-day fertile window (i.e., days -13 through -9 and 1 through 14). Errors occur when a counting method and low fertility sampling strategy select a high fertility day as being low fertility. These error rates are based on 200,000 samplings

sizes when sampling from processes that only generated a behavioral shift during high fertility (i.e., Processes A and C). However, when the process has a secondary shift during the luteal phase (i.e., Processes B and D) then only sampling low fertility days from the luteal phase would reduce the

average difference in the target behavior between high and low fertility phases (see Table 10). If a process had a consistent relation to hormonal levels (i.e., Process D), then regardless of where low fertility days are sampled from, the average difference between low fertility and high fertility days may be

**Table 10** Average behavioral scores by ovulatory cycle day

Ovulatory day	Null	A 0.50	A 2.00	B 0.50	B 2.00	C 0.50	C 2.00	D 0.50	D 2.00
-13	0.00	0.00	0.01	0.00	0.00	-0.01	-0.01	0.08	0.29
-12	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.08	0.30
-11	0.00	0.01	0.00	0.00	0.00	-0.01	0.01	0.10	0.35
-10	0.00	-0.01	0.00	-0.01	0.01	0.01	0.00	0.09	0.34
-9	-0.01	-0.01	0.01	0.00	0.00	0.01	0.00	0.09	0.43
-8	0.00	-0.01	0.02	0.00	0.00	0.00	0.00	0.12	0.47
-7	0.00	0.00	0.01	-0.01	0.01	0.00	0.01	0.14	0.51
-6	0.00	0.01	0.00	0.00	0.02	0.00	0.01	0.19	0.75
-5	-0.01	0.50	2.00	0.51	2.01	0.07	0.22	0.21	0.83
-4	-0.01	0.50	2.00	0.49	2.01	0.11	0.45	0.26	1.01
-3	0.00	0.50	2.00	0.51	2.00	0.22	0.85	0.39	1.56
-2	-0.02	0.49	2.00	0.50	2.00	0.42	1.68	0.49	1.96
-1	0.00	0.50	2.00	0.51	2.00	0.51	2.00	0.50	1.99
0	0.01	0.50	2.01	0.50	1.99	0.31	1.24	0.29	1.18
1	0.00	0.00	0.00	0.00	0.00	0.06	0.23	0.11	0.42
2	0.00	0.00	-0.01	0.04	0.11	-0.01	0.00	0.08	0.31
3	0.01	0.00	0.00	0.11	0.40	0.01	0.00	0.10	0.38
4	0.01	0.00	0.02	0.16	0.60	0.00	0.00	0.13	0.50
5	-0.01	0.00	0.01	0.14	0.60	0.01	0.01	0.13	0.57
6	-0.01	0.00	0.00	0.17	0.62	0.01	0.00	0.13	0.55
7	0.00	0.00	-0.01	0.13	0.48	0.00	-0.01	0.14	0.58
8	0.00	0.00	0.01	0.11	0.49	0.01	0.01	0.15	0.56
9	0.00	0.00	0.01	0.12	0.49	0.00	0.01	0.15	0.55
10	0.02	-0.01	0.00	0.06	0.23	0.00	0.00	0.17	0.61
11	0.00	0.00	-0.01	0.07	0.26	-0.01	0.01	0.09	0.39
12	0.01	0.00	0.00	0.08	0.32	0.01	0.01	0.10	0.38
13	0.00	-0.01	0.01	0.01	-0.01	0.01	-0.01	0.07	0.27
14	-0.02	-0.01	-0.01	0.00	0.01	0.01	0.00	0.07	0.31

*Note:* Null and Max Mean values of 0.50 and 2.00 are shown for processes A–D. From top to bottom, black boxes encircle selection of follicular low fertility, all follicular high fertility, and selection of luteal low fertility days

smaller. Therefore, how researchers choose to sample both high and low fertility days can impact their ability to detect behavioral shifts between these phases. Problematically, little is known about behavioral trajectories in relation to phases over the cycle, so it may be difficult for researchers to determine a priori what the best sampling approach may be.

## Future directions

Within-subjects *t*-tests performed better than between-subject *t*-tests and overlay correlations, but this represents a very minimal approach to studying ovulatory effects. An advantage of sampling more frequently from each participant's ovulatory cycle is that it would allow researchers to assess between-phase effects using multilevel modeling (e.g., Miller et al., 2007; Prokosch et al., 2009; Roney & Simmons, 2013). This modeling approach has several advantages, such as allowing for the evaluation of multiple time scales concurrently (e.g., participant age, ovulatory calendar day, and day of the week) while controlling for within-subject variability. Additionally, with more observations across the cycle, estimates of average participant-specific scores and fluctuations would be more accurate, and would allow for more complex partitioning of score variance.

Moreover, increased frequency in sampling would help researchers to actually describe the trajectories of target behaviors across the cycle—better informing theory about ovulatory shifts of target behaviors and best methodological practices to sample them. As sampling frequency from each participant increases, the likelihood of sampling from distinct ovulatory phases for each participant also increases. Consequently, researchers could describe these processes across the ovulatory cycle (e.g., using linear growth models; McArdle, 2009), and investigate how distinct ovulatory phases (e.g., fertile days) or hormonal correlates (e.g., estrogen and progesterone) are associated with these behavioral trajectories. In this way, the study of ovulatory effects could transition from an investigation of one to two days sampled from a 22- to 36-day cycle to the study of the cycle itself.

Finally, an inherent limitation in one- to two-day sampling from the cycle is the possibility of under-representing periods of interest from the cycle. While counting estimation methods can reduce the likelihood of this, they are not full-proof. In fact, we found that the easiest and most common counting method, forward, is less than half as effective at sampling from fertile days as taking a random sample, and both backward methods, while more effective than forward, were still only about three-quarters as effective as taking a random sample (i.e., for a one-day peak). Further, true backwards counting can only truly be applied after sampling has occurred. The more typically used pseudo-backward counting method relies on an estimate of next menstrual onset that necessitates an

assumption of cycle length regularity and participant accuracy. This is an assumption that has been demonstrated to be untenable for several reasons (e.g., Creinin et al., 2004; Fehring et al., 2006; Small et al., 2007; Wilcox et al., 2000). By increasing sampling frequency, at a minimum, true backwards estimation or LH testing could be incorporated, and the risk of failing to sample from critical phases of the ovulatory cycle would be reduced.

## Conclusion

The results of our study cannot inform the validity of the ovulatory shift hypothesis, or the broader field of research on ovulatory effects in humans. Nor does our study discuss a novel concern over the potential inadequacy of methodology popularly employed in this area of research. What our study does is make explicit the degree of the problem that popular methods can create. With these methodological limitations made clear, debate over best practices on these topics can reside on more than conjecture, and the field can adopt some minimal methodological guidelines to maximize power while mitigating the risk of spurious findings as it moves forward.

## References

- Beall, A. T., & Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological Science*, *24*, 1837–1841. doi:10.1177/0956707613476045
- Behre, H. M., Kuhlage, J., Gabner, C., Sonntag, B., Schem, C., Schneider, H. P. G., & Nieschlag, E. (2000). Prediction of ovulation by urinary hormone measurements with the home use ClearPlan® Fertility Monitor: Comparison with transvaginal ultrasound scans and serum hormone measurements. *Human Reproduction*, *15*, 2478–2482.
- Bullivant, S. B., Sellergren, S. A., Stern, K., Spencer, N. A., Jacob, S., Mennella, J. A., & McClintock, M. K. (2004). Women's sexual experience during the menstrual cycle: Identification of the sexual phase by noninvasive measurement of luteinizing hormone. *The Journal of Sex Research*, *41*, 82–93. doi:10.1080/00224490409552216
- Cárdenas, R. A., & Harris, L. J. (2007). Do women's preferences for symmetry change across the menstrual cycle? *Evolution and Human Behavior*, *28*, 96–105. doi:10.1016/j.evolhumbehav.2006.08.003
- Caryl, P. G., Bean, J. E., Smallwood, E. B., Barron, J. C., Tully, L., & Allerhand, M. (2009). Women's preference for male pupil size. Effects of conception risk, sociosexuality, and relationship status. *Personality and Individual Differences*, *46*, 503–508. doi:10.1016/j.paid.2008.11.024
- Creinin, M. D., Keverline, S., & Meyn, L. A. (2004). How regular is regular? An analysis of menstrual cycle regularity. *Contraception*, *70*, 289–292. doi:10.1016/j.contraception.2004.04.012
- DeBruine, L., Jones, B. C., Frederick, D. A., Haselton, M. G., Penton-Voak, I. S., & Perrett, D. I. (2010). Evidence for menstrual cycle shifts in women's preferences for masculinity: A response to Harris

- (in press) “Menstrual cycle and facial preferences reconsidered”. *Evolutionary Psychology*, 8, 768–775.
- Dunson, D. B., Baird, D. D., Wilcox, A. J., & Weinberg, C. R. (1999). Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction*, 14, 1835–1839.
- Dunson, D. B., Colombo, B., & Baird, D. D. (2002). Changes with age in the level and duration of fertility in the menstrual cycle. *Human Reproduction*, 17, 1399–1403.
- e-Publications@Marquette. (2012). *Randomized comparison of two internet-supported methods of natural family planning [Data file and description]*. Retrieved 6 April 2014 from [http://epublications.marquette.edu/data\\_nfp/7/](http://epublications.marquette.edu/data_nfp/7/)
- Fehring, R. J., Schneider, M., & Raviele, K. (2006). Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecological, & Neonatal Nursing*, 35, 376–384. doi:10.1111/j.1552-6909.2006.00051.x
- Fehring, R. J., Schneider, M., Raviele, K., Rodriguez, D., & Pruszyński, J. (2013). Randomized comparison of two internet-supported fertility-awareness-based methods of family planning. *Contraception*, 88, 24–30.
- Ferguson, C. J. (2014). Comment: Why meta-analyses rarely resolve ideological debates. *Emotion Review*, 6, 251–252. doi:10.1177/1754073914523046
- Gangestad, S. W., Thornhill, R., & Garver, C. E. (2002). Changes in women’s sexual interests and their partners’ mate-retention tactics across the menstrual cycle: Evidence for shifting conflicts of interest. *Proceedings of the Royal Society of London B*, 269, 975–982. doi:10.1098/rspb.2001.1952
- Garver-Apgar, C. E., Gangestad, S. W., & Thornhill, R. (2008). Hormonal correlates of women’s midcycle preference for the scent of symmetry. *Evolution and Human Behavior*, 29, 223–232. doi:10.1016/j.evolhumbehav.2007.12.007
- Gildersleeve, K., DeBruine, L., Haselton, M. G., Frederick, D. A., Penton-Voak, I. S., Jones, B. C., & Perrett, D. I. (2013). Shifts in women’s mate preferences across the ovulatory 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. (2014a). Do women’s mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*. doi:10.1037/a0035438
- Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014b). Meta-analyses and p-curves support robust cycle shifts in women’s mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin*, 140, 1272–1280.
- 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., 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, 1260–1264. doi:10.1037/a0036478
- Haselton, M. G., Mortezaie, M., Pillsworth, E. G., Bleske-Rechek, A., & Frederick, D. A. (2007). Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress. *Hormones and Behavior*, 51, 40–45. doi:10.1016/j.yhbeh.2006.07.007
- Howards, P. P., Schisterman, E. F., Wactawski-Wende, J., Reschke, J. E., Frazer, A. A., & Hovey, K. M. (2008). Timing Clinic Visits to Phases of the Menstrual Cycle by Using a Fertility Monitor: The BioCycle Study. *American Journal of Epidemiology*, 169, 105–112. doi:10.1093/aje/kwn287
- Larson, C. M., Haselton, M. G., Gildersleeve, K. A., & Pillsworth, E. G. (2013). Changes in women’s feelings about their romantic relationships across the ovulatory cycle. *Hormones and Behavior*, 63, 128–135. doi:10.1016/j.yhbeh.2012.10.005
- Little, A. C., Jones, B. C., & Burriss, R. P. (2007). Preferences for masculinity in male bodies change across the ovulatory cycle. *Hormones and Behavior*, 51, 633–639. doi:10.1016/j.yhbeh.2007.03.006
- Little, A. C., Jones, B. C., & DeBruine, L. M. (2008). Preferences for variation in masculinity in real male faces change across the menstrual cycle: Women prefer more masculine faces when they are more fertile. *Personality and Individual Differences*, 45, 478–482. doi:10.1016/j.paid.2008.05.024
- Lloyd, R., & Coulam, C. B. (1989). The accuracy of urinary luteinizing hormone testing in predicting ovulation. *American Journal of Obstetrics and Gynecology*, 160, 1370–1375.
- Lukaszewski, A. W., & Roney, J. R. (2009). Estimated hormones predict women’s mate preferences for dominant personality traits. *Personality and Individual Differences*, 47, 191–196.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605. doi:10.1146/annurev.psych.60.110707.163612
- Miller, S. L., & Maner, J. K. (2010). Scent of a woman: Testosterone responses to olfactory ovulation cues. *Psychological Science*, 21, 276–283. doi:10.1177/0956797609357733
- Miller, G., Tybur, J. M., & Jordan, B. D. (2007). Ovulatory cycle effects on tip earnings by lap dancers: Economic evidence for human estrus? *Evolution and Human Behavior*, 28, 375–381. doi:10.1016/j.evolhumbehav.2007.06.002
- Moore, F. R., Cornwell, R. E., Law Smith, M. J., Al Dujaili, E. A. S., Sharp, M., & Perrett, D. I. (2011). Evidence for the stress-linked immunocompetence handicap hypothesis in human male faces. *Proceedings of the Royal Society B: Biological Sciences*, 278, 774–780. doi:10.1098/rspb.2010.1678
- 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
- Oinonen, K. A., & Mazmanian, D. (2007). Facial symmetry detection ability changes across the ovulatory cycle. *Biological Psychology*, 75, 126–145. doi:10.1016/j.biopsycho.2007.01.003
- Pawlowski, B., & Jasienska, G. (2005). Women’s preferences for sexual dimorphism in height depend on menstrual cycle phase and expected duration of relationship. *Biological Psychology*, 70, 38–43. doi:10.1016/j.biopsycho.2005.02.002
- 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
- Peters, M., Simmons, L. W., & Rhodes, G. (2009). Preferences across the menstrual cycle for masculinity and symmetry in photographs of male faces and bodies. *PLoS ONE*, 4(1), e4138. doi:10.1371/journal.pone.0004138
- 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
- Prokosch, M. D., Coss, R. G., Scheib, J. F., & Blozis, S. A. (2009). Intelligence and mate choice: Intelligent men are always appealing. *Evolution and Human Behavior*, 30, 11–20. doi:10.1016/j.evolhumbehav.2008.07.004
- Puts, D. A. (2005). Mating context and menstrual phase affect women’s preferences for male voice pitch. *Evolution and Human Behavior*, 26, 388–397. doi:10.1016/j.evolhumbehav.2005.03.001
- Roberts, S. C., Havlicek, J., Flegr, J., Hruskova, M., Little, A. C., Jones, B. C., ... Petrie, M. (2004). Female facial attractiveness increases during the fertile phase of the menstrual cycle. *Proceedings of the*

- Royal Society of London: *B*, 271(Suppl 5), S270–S272. doi:10.1090/rsbl.2004.0174
- Roney, J. R., & Simmons, Z. L. (2008). Women's estradiol predicts preference for facial cues of men's testosterone. *Hormones and Behavior*, 53, 14–19.
- Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of sexual motivation in natural menstrual cycles. *Hormones and Behavior*, 63, 636–645.
- Roney, J. R., Simmons, Z. L., & Gray, P. B. (2011). Changes in estradiol predict within-women shifts in attraction to facial cues of men's testosterone. *Psychoneuroendocrinology*, 36, 742–749. doi:10.1016/j.psycneuen.2010.10.010
- Rosen, M. L., & López, H. H. (2009). Menstrual cycle shifts in attentional bias for courtship language. *Evolution and Human Behavior*, 30, 131–140. doi:10.1016/j.evolhumbehav.2008.09.007
- Schwarz, S., & Hassebrauck, M. (2008). Self-perceived and observed variations in women's attractiveness throughout the menstrual cycle—a diary study. *Evolution and Human Behavior*, 29, 282–288. doi:10.1016/j.evolhumbehav.2008.02.003
- 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
- Small, C. M., Manatunga, A. K., & Marcus, M. (2007). Validity of self-reported menstrual cycle length. *Annals of Epidemiology*, 17, 163–170. doi:10.1016/j.annepidem.2006.05.005
- Stricker, R., Eberhart, R., Chevailler, M., Quinn, F. A., Bischof, P., & Stricker, R. (2006). Establishment of detailed reference values for luteinizing hormone, follicle stimulating hormone, estradiol, and progesterone during different phases of the menstrual cycle on the Abbott ARCHITECT analyzer. *Clinical Chemistry and Laboratory Medicine*, 44, 883–887. doi:10.1515/CCLM.2006.160
- Tanabe, K., Susumu, N., Hand, K., Nishii, K., Ishikawa, I., & Nozawa, S. (2001). Prediction of the potentially fertile period by urinary hormone measurements using a new home-use monitor: Comparison with laboratory hormone analyses. *Human Reproduction*, 16, 1619–1624.
- Trussel, J. (2008). Contraceptive efficacy. In R. A. Hatcher, J. Trussel, & A. L. Nelson (Eds.), *Contraceptive Technology* (19th ed.). New York: Ardent Media.
- Vaughn, J. E., Bradley, K. I., Byrd-Craven, J., & Kennison, S. M. (2010). The effect of mortality salience on women's judgments of male faces. *Evolutionary Psychology*, 8, 477–491.
- Wilcox, A. J., Weinberg, C. R., & Baird, D. D. (1995). Timing of sexual intercourse in relation to ovulation. *The New England Journal of Medicine*, 333, 1517–1521.
- Wilcox, A. J., Weinberg, C. R., & Baird, D. D. (1998). Post-ovulatory ageing of the human oocyte and embryo failure. *Human Reproduction*, 13, 394–397.
- Wilcox, A. J., Dunson, D., & Baird, D. D. (2000). The timing of the “fertile window” in the menstrual cycle: Day specific estimates from a prospective study. *British Medical Journal*, 321, 1259–1262.
- 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.
- Wilcox, A. J., Baird, D. D., Dunson, D. B., McConaughey, D. R., Kesner, J. S., & Weinberg, C. R. (2004). On the frequency of intercourse around ovulation: Evidence for biological influences. *Human Reproduction*, 19, 1539–1543. doi:10.1090/humrep/deh305
- Wood, W., & Carden, L. (2014). Elusiveness of menstrual cycle effects on mate preferences: Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, 140, 1265–1271.
- Wood, W., & Joshi, P. D. (2011). *A meta-analysis of women's mate preferences across the menstrual cycle*. Paper presented at the Annual Conference of the Society of Experimental Social Psychology, Washington, DC.
- Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2012a). *Empirical findings of menstrual cycle effects on mate preferences: A meta-analytic review*. Paper presented at the Annual Conference of the Society for Personality and Social Psychology, San Diego, California.
- Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2012b). *Meta-analytic review of menstrual cycle effects on mate preferences*. Paper presented at the Annual Conference of the Association for Psychological Science, Chicago, Illinois.
- Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review*. Advance online publication. <http://emr.sagepub.com/content/early/2014/03/24/1754073914523073>