Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications

Steven W. Gangestad [a,*], Martie G. Haselton [b,c], Lisa L.M. Welling [d], Kelly Gildersleeve [b,c], Elizabeth G. Pillsworth [e], Robert P. Burriss [f], Christina M. Larson [b,c], David A. Puts [g]

[a] *Department of Psychology, University of New Mexico, Albuquerque, NM, USA*
[b] *Department of Communication Studies, University of California, Los Angeles, CA, USA*
[c] *Department of Psychology, University of California, Los Angeles, CA, USA*
[d] *Department of Psychology, Oakland University, Oakland, MI, USA*
[e] *Department of Anthropology, California State University, Fullerton, CA, USA*
[f] *Psychology Department, University of Stirling, Stirling, UK*
[g] *Department of Anthropology, The Pennsylvania State University, University Park, PA, USA*

ARTICLE INFO

Article history:
Initial receipt 19 December 2014
28 August 2015
Final revision received 8 September 2015
Available online xxxx

Keywords:
Menstrual cycle
Evolutionary psychology
Fertility
Fertile phase
Ovarian cycle

ABSTRACT

Over the past two decades, a large literature examining psychological changes across women's ovulatory cycles has accumulated, emphasizing comparisons between fertile and non-fertile phases of the cycle. While some studies have verified ovulation using luteinizing hormone (LH) tests, counting methods – assessments of conception probability based on counting forward from actual or retrospectively recalled onset of last menses, or backward from actual or anticipated onset of next menses – are more common. The validity of these methods remains largely unexplored. Based on published data on the distributions of the lengths of follicular and luteal phases, we created a sample of 58,000+ simulated cycles. We used the sample to assess the validity of counting methods. Aside from methods that count backward from a confirmed onset of next menses, validities are modest, generally ranging from about .40–.55. We offer power estimates and make recommendations for future work. We also discuss implications for interpreting past research.

1. Introduction

Two studies in the late 1990s triggered a rapid expansion of interest in psychological changes related to the ovulatory cycle (Gangestad & Thornhill, 1998; Penton-Voak et al., 1999; see also Grammer, 1993). Both documented increases in women's preferences for purported indicators of heritable fitness at high fertility relative to low fertility within the cycle. A decade and a half later, there are dozens of studies of cycle shifts in women's mate preferences and dozens more of cycle shifts in women's attractiveness, including changes in women's body odors, voices, facial appearance, and proceptive behavior (reviewed in Gildersleeve, Haselton, & Fales, 2014a; Gangestad, Thornhill, & Garver-Apgar, 2015; Haselton & Gildersleeve, 2011; Thornhill & Gangestad, 2008).

Cycle shift effects have attracted attention and intense research effort for at least two reasons. First, cycle shifts are non-intuitive and difficult to explain without an explicit evolutionary account. Therefore, these findings have been viewed as powerful evidence of the utility of an evolutionary approach for understanding human behavior (e.g., Neuberg, Kenrick, & Schaller, 2010). Second, these findings have challenged the widespread prior conclusion that human sexuality – unlike that of many non-human species, including most other primates – is independent of hormonal control (e.g., Symons, 1979). Thus, the discovery of cycle shifts in women's mate preferences and attractiveness has heralded a potentially radical revision to understandings of human sexuality and its evolutionary and hormonal underpinnings.

In a meta-analysis of studies examining cycle shifts in women's mate preferences, Gildersleeve et al. (2014a) found robust but modest effects. In a subsample of studies examining targeted cycle shifts in attraction to hypothesized male fitness indicators (e.g., facial, body, and vocal masculinity; facial symmetry and scents associated with symmetry; and behavioral dominance), weighted mean effect sizes in a short-term mating context and unspecified context were .26 and .20, respectively (Hedge's *g*, comparable to Cohen's *d*). In a more recent meta-analysis of studies examining detectable changes accompanying women's fertile phase, Gildersleeve and Haselton (2014) found robust effects of comparable magnitude. Subsamples of studies examining fertility cues that are relatively likely or unlikely to be under women's volitional control (e.g., proceptive behavior vs. natural body odor attractiveness) yielded mean effect sizes of .20 and .28, respectively.

At the same time, many studies have yielded null findings. Indeed, of 42 published and unpublished studies in Gildersleeve et al.'s (2014a) subsample of targeted cycle shifts, 17 (40%) produced a statistically significant finding, whereas 60% did not. This variability in outcomes has

* Corresponding author. Department of Psychology, University of New Mexico, Albuquerque, NM 87131.
E-mail address: sgangest@unm.edu (S.W. Gangestad).

http://dx.doi.org/10.1016/j.evolhumbehav.2015.09.001
1090-5138/© 2015 Published by Elsevier Inc.

led some to argue that previous findings were false positives, and support for effects was largely due to publication bias (Wood, Kressel, Joshi, & Louie, 2014). Others have noted that wide variation in methods used to assess women's fertility within the cycle permits considerable analytic flexibility. As a result, researchers may well have tried multiple analyses (e.g., with different high- and low-fertility windows) and reported only favorable results (e.g., Harris, Pashler, & Mickes, 2014). In other words, positive findings might have been “*p*-hacked” (Simonsohn, Nelson, & Simmons, 2014).

To empirically examine publication bias and *p*-hacking, Gildersleeve, Haselton, and Fales (2014b) constructed *p*-curves of significant findings in the meta-analysis sample. Consistent with the existence of real cycle shifts, these curves were robustly right-skewed, with a disproportionately large number of *p*-values < .01.The estimated mean effect size was .30, slightly greater than meta-analytic estimates. As well, *p*-curves were consistent with statistical power of only about 33%. One possible explanation for variability in the significance of cycle shift effects, then, is relatively weak power.

Most studies examining cycle shifts have assessed conception probability using a counting method – either counting forward from last menstrual onset or backward from next menstrual onset to the current day to assess whether a woman is presently in her “fertile window.” Yet the validities of these methods have never been thoroughly evaluated, let alone quantified (but see our discussion of Gonzales & Ferrer, 2015, below). An evaluation of these methods is timely for two reasons. First, such an evaluation can make clear which methods have greatest validity and thereby encourage more uniform and accurate procedures moving forward. Second, extant data suggest that effect sizes are robust but modest; and the typical study, underpowered. However, there remains the question of *why*. Effect sizes detected in studies are a function of the “true” effect of conception probability and the validity with which fertility status is measured. One possibility is that the effect of conception probability truly is small (e.g, Harris et al., 2014). However, an alternative possibility is that effect sizes merely *appear* small because measurement is poor. For example, if the correlation between estimated and true conception probability is only .5, the study will produce an effect size 50% of the true effect size. Because we do not know the validities of methods used to assess conception probability, we cannot yet draw confident conclusions based on the extant literature.

In this paper, we evaluate the validity of these methods. We aim to contribute to *methodological standards* for the future, but our results can also contribute to a proper *theoretical interpretation* of findings to date.

1.1. Methods used in studies of shifts across the ovulatory cycle

A woman has a non-zero conception probability – probability of conceiving following unprotected sex – on the day of ovulation and up to 5 days prior (e.g., Baird et al., 1995). All days outside of this “fertile window” are non-fertile. The follicular phase extends from the onset of menses until ovulation. The luteal phase extends from ovulation until next menstrual onset. The fertile window, then, is the latter part of the follicular phase. Aside from a few hours following ovulation, the luteal phase is non-fertile. See Fig. 1.

Researchers have typically used one of two methods to assess where women fall within the ovulatory cycle: Detection of an LH surge and day-of-cycle counting.

1.1.1. LH detection

Luteinizing hormone (LH), released by the pituitary gland, characteristically surges 24–36 hours prior to ovulation (e.g., Guermandi et al., 2001). Typically marketed to women actively trying to conceive, test sticks that detect an *LH surge* are commercially available (e.g., Clearblue©, OvuSign©). Kits typically consist of plastic-encased strips that contain an immunoassay sensitive to LH in urine.

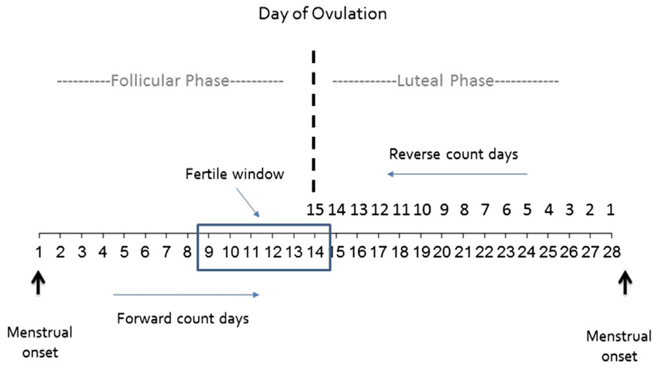

Fig. 1. Hypothetical cycle of a woman whose cycle length was 28 and day of ovulation was day 14. The “fertile window” in this cycle extends from forward count day 9 to forward count day 14. By the reverse count, her day of ovulation was day 15, and her fertile phase was reverse count day 20 to reverse count day 15. The follicular phase ends at ovulation. The luteal phase begins at ovulation.

When correctly used, LH detection tests are very accurate. In one study, Clearblue© found that over 99% of LH surges were detected by their tests (see http://www.clearblueeasy.com/healthcare/clearblue-digital-ovulation-test.php). As LH surges vary in their duration and intensity (Direito, Bailly, Mariani, & Ecochard, 2013; Park, Goldsmith, Skurnick, Wojtczuk, & Weiss, 2007), however, accurate detection is enhanced when LH tests are administered daily until the onset of the surge. Some studies (e.g., Fales, Gildersleeve, & Haselton, 2014) have followed up positive results by verifying the date of next menstrual onset, which usually (~80% of the time) occurs 14 ± 2 days after ovulation (e.g., Baird et al., 1995).

Studies that use LH tests are typically within-subject designs, with individual women assessed twice during a cycle: once when fertile, as verified by LH tests, and once during the mid-luteal phase (e.g., Gangestad, Thornhill, & Garver, 2002, 2014; Gangestad, Garver-Apgar, Cousins, & Thornhill, 2014; Gangestad, Thornhill, & Garver-Apgar, 2005; Pillsworth & Haselton, 2006; Durante, Griskevicius, Hill, Perilloux, & Li, 2011), though some studies have assessed women 3 + times (e.g., Burriss et al., 2015). When fertile phase assessments precede luteal phase assessments, researchers typically schedule luteal phase sessions to follow fertile phase sessions by a week or more. When luteal phase assessments precede fertile phase assessments, researchers typically ask women to report their menstrual onset between sessions, thereby verifying that the luteal phase session did in fact occur during the luteal phase (e.g., Larson, Pillsworth, & Haselton, 2012, Larson, Haselton, Gildersleeve, & Pillsworth, 2013).

A few studies have scheduled women's high-fertility session only after detecting an LH surge (e.g, Cantú et al., 2014). However, most have scheduled women's high-fertility session on a day when they were expected to be fertile but only counted that session as fertile if women experienced an LH surge no more than 2 days before or 4 days after it (e.g., Gildersleeve, Haselton, Larson, & Pillsworth, 2012). Although one can assign specific continuous conception probabilities depending on timing of a session relative to the LH surge (e.g., Burriss et al., 2015), most studies to date have simply categorized sessions as being in or outside of the fertile window.

1.1.2. Day-of-the-cycle counting methods

The most widely used methodology involves counting days from menstrual onset to assess cycle position. Within this approach, multiple methods have been used.

The *forward* method counts days from last menstrual onset forward to the day of assessment. For instance, if a woman was assessed on the 15th of the month, and her last menstrual period (“day 1” of her cycle) began on the 5th of the month, then her session was on “day 11” of her cycle.

The *backward* method counts days from next menstrual onset backward to the day of assessment (e.g., Puts, 2005). If a woman was assessed on the 15th of the month, and her next menstrual period began on the 25th of the month, then her session was on "reverse count day 10" of her cycle (10 days prior to the end of the cycle, as marked by the onset of her next menstrual period). The rationale for backward counting is that luteal phase lengths are less variable than follicular phase lengths (e.g., Baird et al., 1995; Fehring, Schneider, & Raviele, 2006). Hence, referencing a woman's day of the cycle relative to the end of it, rather than its onset, should be a more accurate way of assessing her fertile window. The method, however, requires an assessment of onset of next menses. Some researchers have followed up with women to verify their date of next menstrual onset. For practical reasons, however, many researchers have estimated when women will start their next period based on women's self-reports of their date of last menstrual onset and typical cycle length or women's own prediction of when their next period will begin.

Counting methods must convert day of the cycle into a conception probability. The most common approach to date defines *discrete windows*. By the forward approach, researchers define a particular range of days as high fertility, and some or all of the remaining days as low fertility. Different studies have designated different windows. Hence, Penton-Voak et al. (1999) and others (e.g., Little, Jones, & Burriss, 2007), following conception probabilities published by Jöchle (1973), defined days 6–14 as high fertility. Others have defined the high fertility window differently (see Gildersleeve et al., 2014a). As noted by Gildersleeve et al. (2014a) and Wood et al. (2014), window sizes have varied; 6–9 days are most common.

Women ovulate on different days of the cycle, as both follicular and luteal phase lengths vary. Therefore, some researchers have represented conception probability as a *continuous measure*. Conception probabilities are actual probabilities of conception among women having unprotected sex on different days of the cycle, the most widely used of which were developed by Wilcox, Duncan, Weinberg, Trussell, and Baird (2001). For instance, day 7, 12, 19, and 27 have conception probabilities of .017, .084, .032, and .007, respectively. See Fig. 2. With a forward procedure, women are assigned a conception probability corresponding to their current day in cycle. With a backward procedure, women can be placed on a standard 28- or 29-day cycle through calculations of how far they are from the end of their cycles and assumptions that, on a standard 29-day cycle, women ovulate on day 15. Women are then assigned conception probabilities based on the published estimates (e.g., Gangestad, Garver-Apgar, Simpson, & Cousins, 2007; Puts, 2006). Finally, some studies have averaged conception probabilities based on forward and backward procedures, as each could have unique error (e.g., Gangestad et al., 2007).

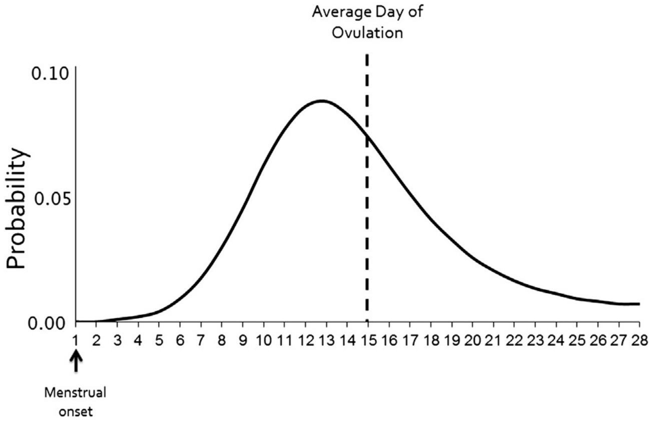

Fig. 2. Probabilities of conception resulting from a single act of intercourse by day of the cycle estimated by Wilcox et al. (2001). For any given cycle, the fertile window lasts up to 6 days, with varying probabilities of conception resulting from a single act of intercourse (maximal 1–2 days prior to day of ovulation). Mean day of ovulation (length of the follicular phase) is 15. But day of ovulation varies across cycles, both within and between women. Probabilities for conception, then, are a function of (a) the distribution of follicular phase lengths, and (b) variation in probabilities of conception across days within the fertile window.

1.2. The validity of conception probability measures

How valid are measures of conception probability? That is, how highly do they correlate with women's *true* conception probabilities?

1.2.1. LH detection

Aside from more expensive procedures (e.g., measurement of daily estradiol and progesterone levels, ultrasound; e.g., Cobey, Buunk, Pollet, Kippling, & Roberts, 2013; Roney & Simmons, 2013), LH detection sticks offer the greatest accuracy. If, as may be the case, inaccurate classification of women as being in the fertile phase vs. the luteal phase is no more than ~5%, the validity of conception probability, expressed as a Pearson *r* (a phi coefficient), equals or exceeds .9. Even with LH detection, however, some women will be assessed during their fertile phase but not necessarily on their most fertile days. Validity is .8 if one assumes a misclassification rate of 10% - likely an upper-bound estimate, and hence one that yields a lower-bound estimate of accuracy.

1.2.2. Counting methods

Counting methods, naturally, are less accurate. But how much so? And which methods outperform others? These questions remain unaddressed.

1.3. A methodology to accurately estimate the validity of counting methods

The accuracy of counting methods depends upon a number of parametric features of the distributions of women's cycles. Forward methods depend on the distribution of follicular phase lengths – that is, when ovulation occurs. Backward methods depend crucially on the distribution of luteal phase lengths, as well as the precision with which one knows the length of the current cycle (e.g., based on confirmed onset of next menses or self-reported typical cycle length).

Much is now known about these distributions. Recently, Stirnemann, Samson, Bernard, and Thalabard (2013)estimated the day of conception, via ultrasound fetal biometry (conducted 11–14 weeks post-conception, with statistical adjustments to reduce error) in nearly 6000 women. As women typically conceive within hours after ovulation (e.g., Harper, 1994), the distribution of days of conception effectively matches the distribution of days of ovulation. Results, then, should closely match the true distribution of women's follicular phase lengths (see also Fehring et al., 2006; Wilcox et al., 2001; Cole, Ladner, & Byrn, 2009, who also estimated distribution of follicular phase length, but with smaller samples and/or less accurate methodology).

Multiple studies have estimated the distribution of luteal phase lengths (Baird et al., 1995; Cole et al., 2009; Fehring et al., 2006; Lenton, Landgren, & Sexton, 1984). Additionally, Fehring et al. (2006) report a modest negative correlation between follicular phase and luteal phase lengths in a sample of over 1000 women, a value consistent with data presented by Cole et al. (2009). Cole et al. (2009) partitioned variation across cycles into between-women and within-woman components. Finally, at least two studies have examined the accuracy of women's self-reported average cycle length (Jukic et al., 2008; Small, Manatunga, & Marcus, 2007).

Though this information cannot estimate the accuracy of conception probability measures algorithmically, a way forward is possible: One can use this knowledge to simulate a sample that mimics real cycles. Within a large, representative sample in which one knows for each simulated cycle precisely when ovulation occurred and, hence, which days are fertile, one can compute and evaluate the accuracy of conception probability estimates given by various counting methods. We used these procedures to estimate the validity coefficients of specific counting methods.

One other recent study simulated cycles using similar methodology. Gonzales and Ferrer (2015) used data on parameters of cycle characteristics reported by Fehring et al. (2006) to create a simulation sample. Our simulation procedures and those of Gonzales and Ferrer (2015) differ in a number of ways: First, we used Stirnemann et al. (2013) distribution of day of ovulation, estimated from fetal bimetry, to create a sample of cycles, whereas Gonzales and Ferrer (2015) used mean day of ovulation, along with an estimate of its standard deviation, and assumed normality to create their sample. Second, though Gonzales and Ferrer (2015) attempted to account for inaccuracy of women's estimates of next menstrual onset, our estimates, based on pertinent literature, assumed less accuracy. Third, we examined the effect of an ovulation, whereas Gonzales and Ferrer (2015) did not. The most important difference between our simulations and those of Gonzales and Ferrer (2015), however, is that the two simulations had fundamentally different aims. Gonzales and Ferrer (2015) sought to estimate the power of studies of particular designs. Specifically, in between-subject studies evaluated in these simulations women were either sampled from a phase estimated to consist of fertile days or from a phase estimated to consist of non-fertile days; in within-subject studies women were assessed during both. The authors specifically sought to estimate the statistical power of such studies. By contrast, we sought to estimate the validity of methods that assign conception probabilities to women (either by classifying women into fertile and non-fertile groups, or by assigning quantitative values of conception probability) when women are sampled randomly. Our simulations not only speak to validity. They also speak to statistical power of studies that sample women randomly from across the cycle and assign conception probability based on counting methods, which has been the most common method used in cycle studies to date (Gildersleeve et al., 2014a). Therefore, our simulations, and their implications, apply to a broader range of study designs and considerably more of the extant literature.

2. Methods

2.1. Generating a representative sample of cycle days

We created our simulation sample in 6 steps. See Supplemental Online Materials (SOM) for an expanded description.

1. *A sample with day of ovulation and current day of the cycle*. First, we created a sample with a representative distribution of days of ovulation within a cycle (or, equivalently, follicular phase lengths, as the follicular phase begins the first day of the cycle and ends on the day of ovulation). Stirnemann et al. (2013) used ultrasound fetal biometry to estimate the day of conception (number of days following beginning of last menses to day of conception) on a sample of nearly 6000 women. As they noted, conception typically occurs within 12 hours of ovulation; hence, the distribution of days of conception should closely match that of days of ovulation. We used an online graphical data extractor (http://arohatgi.info/WebPlotDigitizer/) on Stirnemann et al.'s Fig. 1 to obtain proportions of cycles in which conception occurred on a given day. We then created 1000 cases that matched these proportions. This sample of 1000 was multiplied 35-fold, with each of the 35 sets given a current day of the cycle ranging from 1 to 35. Hence, our sample of 35,000 cycles had a distribution of days of ovulation matching that of Stirnemann et al. and a uniform distribution of current day of the cycle, representing days 1 to 35.
2. *Assumed distribution of luteal phase lengths*. Several large-sample studies have estimated the length of the luteal phase (days from ovulation to beginning of next menses) to average 13–14 days (Baird et al., 1995: 13.1; Cole et al., 2009: 13.2 days; Lenton et al., 1984: 14.1 days; Fehring et al., 2006: 12.4 days; sample sizes range from 327–1060), with standard deviations of approximately 2.0 days (Baird et al.: 2.2; Fehring et al.: 2.0; Cole et al.: 2.0; Lenton et al.: 1.4 with outlying values excluded). We created a sample of luteal phase lengths approximating a normal distribution with mean 13.5 days and standard deviation of 2.0.
3. *Assumed correlation between follicular and luteal phase lengths*. Across over 1000 cycles, Fehring et al. (2006) estimated a correlation of − .323 between follicular and luteal phase lengths, consistent with an estimate by Cole et al. (2009). We modeled a correlation of − .3 by generating a normally distributed standard random variable and creating a weighted sum of that variable and *z*-scored follicular phase lengths, weights being $\sqrt{(1\text{-}.3^2)}$ and .3, respectively. We then transformed this sum to a variable with mean 13.5 and standard deviation of 2.0 (above), and rounded values to the nearest integer; this variable is luteal phase length.
4. *Elimination of cases with cycle day exceeding cycle length*. As current day of the cycle cannot possibly exceed length of the current cycle (i.e., follicular + luteal phase lengths), we eliminated from our data base all such impossible cases (19.5% of all simulated cycles).
5. *A second sample*. To assess the impact of random variation in our simulated sample, we created a second sample using precisely the same procedures but generating a new random variable to compute luteal phase lengths. Results for the two samples were nearly identical: mean absolute difference in estimated validity coefficients (expressed as *r*; see below) was .006. We report results for the two samples combined.
6. *The fertile phase and estimated probabilistic conception probability*. Following Stirnemann et al. (2013), we defined the true fertile window as beginning mid-way through the day five days prior to the day of ovulation and ending mid-way through the day of ovulation. Hence, we coded this variable for the 5 days prior to the day of ovulation and the day of ovulation itself as .5, 1, 1, 1, 1, .5, respectively, and all other days as 0. Following Wilcox, Weinberg, and Baird (1995), we also assigned continuous probabilities of conception resulting from unprotected sex. See SOM.

Data were created and analyzed using SPSS 22.0.We generated 28,197 and 28,148 cases in the two samples, for a total combined sample of 56,345 cases. Validities of all methods were assessed based on this total combined sample. Data files are freely downloadable from http://psych.unm.edu/people/directory-profiles/steven-gangestad.html.

Table 1
Characteristics of the simulated representative sample

	Observed		Target	
	Mean	SD	Mean	SD
Cycle length	28.56	3.39	28.56[1]	3.34
Length of the follicular phase	15.00	3.44	14.96[2]	3.40
Length of the luteal phase	13.56	2.02	13.50[3]	2.00

	Pearson product–moment correlation coefficients	
	Observed	Target
Lengths of the follicular and luteal phases	−.317	−.300[4]
Lengths of cycle and follicular phase	.825	.829[1]
Lengths of cycle and luteal phase	.276	.255[1]

Notes. *N* = 56,345.

[1] Based on Fehring et al. (2006).

[2] Based on Stirnemann et al. (2013).

[3] Based on Lenton et al. (1984), Baird et al. (1995), Fehring et al. (2006), Cole et al. (2009).

[4] Based on Fehring et al. (2006) and Cole et al. (2009); value reported in Fehring et al.: −.323.

2.2. Characteristics of the representative sample

To assess how well our sample represents real cycles, we examined its features. As Table 1 shows, mean length of the follicular phase is 15.00 ($SD = 3.44$). As outlined in Table 1, the mean lengths of the follicular phase, estimated from Stirnemann et al. (2013) (eliminating lengths > 29 days), and the luteal phase in our target distribution are nearly identical to our observed distribution. The correlation between lengths of the follicular and luteal phases is − .317, close to our target of − .3, and even closer to the value of − .323 reported by Fehring et al. (2006). Cycle length, then, has a mean of 15.00 + 13.56, or 28.56, with a standard deviation of 3.39. Fehring et al.'s (2006) unweighted mean of five large studies is also 28.56 ($s = 3.34$). Fehring et al. reported correlations between the length of the entire cycle and length of the follicular and luteal phases as .829 and .255, respectively. In our sample, these correlations are .825 and .276. In sum, our simulated sample almost perfectly matches real samples of cycles with respect to the distribution of cycle lengths, follicular phase lengths, luteal phase lengths, and their associations – precisely the features that affect the distribution of fertile days within the cycle.

Table S1 (see SOM) compares the percentages of cases possessing particular days of ovulation we targeted based on Stirnemann et al. and the percentages of cases in our sample with those days of ovulation. Again, our sample closely matches our target. We therefore feel confident that our simulated cycles can accurately evaluate the validity of conception probability measures.

2.3. The methods of estimated conception probability evaluated

We evaluated two sets of methods for estimating fertility status: those based on (1) Wilcox et al. (2001) and on (2) Stirnemann et al. (2013). Wilcox et al. (2001) estimated probabilities of conception following unprotected sex for each day of the cycle, using a sample of 696 cycles from 213 women. These values were based on days of ovulation estimated from daily patterns of urinary metabolites of estradiol and progesterone, in conjunction with estimated probabilities of conception for days relative to ovulation (Wilcox et al., 1995). Based on their estimated distribution of days of ovulation and a presumed 5-day fertile window (extending over 6 days, with half of the first day and last day being within the window), Stirnemann et al. (2013) estimated probability of being within the fertile window for each day of the cycle. As the distribution of the fertile window in our sample is based on Stirnemann et al.'s work, estimates of validity might be slightly biased in favor of Stirnemann et al.'s methods.

Within each of these two sets, we evaluated specific methods of estimating fertility within the cycle.

2.3.1. Continuous forward estimate of fertility

Wilcox et al. (2001; Table 1) and Stirnemann et al. (2013; Fig. 5) give conception probability or probability of being within the fertile window for each day of the cycle. We extracted values from Stirnemann et al.'s figure using http://arohatgi.info/WebPlotDigitizer/. See SOM for all values and SPSS syntax.

2.3.2. Continuous backward estimate of fertility, actual cycle length

Backward estimates are based on day of the cycle in conjunction with cycle length, assuming luteal phase length is nearly invariant. To derive these estimates using continuous values, we first placed a woman's current day of the cycle on a standard 29-day cycle based on cycle length, then assigned probability values from Wilcox et al. (2001) for regular cycles and Stirnemann et al. (2013) for all cycles. Based on the modal luteal phase length reported by Baird et al., 1995), we assumed a luteal phase length of 14 days. If a woman was within 14 days of the end of the cycle, then, her place on a standard 29-day cycle would be 29 minus the number of days from the end of her cycle. (For example, if she were on day 20 of a 31-day cycle, she would be placed at day 18 of a 29-day cycle – the same 11 days from the end of her cycle.) If she were more than 14 days from the end of her cycle, we estimated her comparable day of the cycle to be the day proportionately at the same place within the follicular phase, rounded to the nearest integer day. (For example, if she were on day 11 of a 31-day cycle – 65% of the way into her presumed follicular phase of 17 days – she would be placed at day 10 within a 29-day cycle, the value closest to 65% of the way into a follicular phase of 15 days.) When using Wilcox et al.'s (2001) estimates, we assigned a conception probability using the values they reported for *regular* cycles (Table 1). Stirnemann et al. (2013) did not report probabilities of being within the fertile window for regular cycles; rather, they reported probabilities for all cycles. We estimated probabilities for regular cycles from data they present and found that they do not increase validity. Therefore, we used the values they reported for all cycles. (See Gangestad et al., 2007; Garver-Apgar, Gangestad, & Thornhill, 2008; Puts, 2006, for similar procedures.) This backward estimate assumes that the length of the cycle is known precisely and therefore can be used only when researchers obtain a follow-up report from participants of onset of next menses after assessment. (See SOM for SPSS syntax that computes these continuous estimates).

2.3.3. Continuous backward estimate of fertility, reported typical cycle length

Some researchers have computed backward estimates using women's self-reported typical cycle length. These self-reports will not perfectly match actual length of the current cycle because (1) women's cycle lengths vary, and (2) women might not accurately report their typical cycle length. Literatures speak to the impact of each source of error. Cole et al. (2009) estimated that the between-woman individual variance in cycle length is 5.2, while the within-woman variance is 2.8. By these estimates, 35% of the variance in cycle length is within-woman. Jukic et al. (2008) and Small et al. (2007) examined associations between self-reported typical cycle length and observed mean cycle length, and found correlations of only .45 and .50, respectively. As imperfect measurement of actual mean cycle length may partly explain poor matching, we modeled two different scenarios: an optimistic one, which presumed a correlation of true mean cycle length and reported mean cycle length of .7, and a pessimistic one, which presumed this correlation to be .5.

To create measures of self-reported typical cycle length in our data, then, we (a) constructed a measure of true mean cycle length by computing a variable based on current cycle length but with 35% error added in and with 35% less variance than current cycle length ($SD = 2.73$); (b) used this measure to create two self-reported typical cycle length measures, one possessing a Pearson r of .7 with true mean cycle length, the other possessing an r of .5 with true mean cycle length. Mean typical cycle length was set at 28.5, closely reflecting the actual mean in the sample. Standard deviations in the reports matched that in estimated true mean cycle length. Values were truncated at 22 and 35 (affecting <1% of all cases).

To estimate the probability of being within the fertile window using a backward procedure based on self-reported typical cycle length, then, we substituted self-reported cycle length for current cycle length in the procedures we describe above.

2.3.4. Averages of forward and backward estimates

Both forward and backward estimates are subject to error, with error in forward estimates resulting from variation in day of ovulation and error in backward estimates resulting from departures from the assumed 14-day luteal phase length and, for reported typical cycle lengths, the imperfect association between reported typical cycle length and current cycle length. Because errors are imperfectly correlated, an average of the forward and backward estimates could in theory yield a measure more valid than either one alone. Hence, we also evaluated three such averages: the average of the forward estimate with, separately, the backward estimates based on (1) current cycle length, (2) self-reported typical cycle length with validity coefficient (r) of .7, and (3) self-reported typical cycle length with validity coefficient of .5.

2.3.5. Discrete high-and low-fertility windows

Most studies to date have defined a discrete high-fertility window based on either forward or backward procedures (see Gildersleeve et al., 2014a). Size of the high-fertility window has typically varied from 6 to 9 days, with the remaining days (or a subset of them, as we discuss later) designated as low fertility. We constructed high-fertility windows of all sizes within this range based on both forward and backward estimates. Optimal windows were identified by defining the high-fertility window as the 6, 7, 8, or 9 days associated with the highest conception probability or probability of falling within the fertile window. For forward procedures, we identified optimal windows using both Wilcox et al.'s (2001) and Stirnemann et al.'s (2013) conception probability estimates. For backward procedures, we confirmed that our windows were optimal for each window size. Table 2 lists all high-fertility windows we examined. In total, we examined 20 different measures based on discrete windows: 2 (Wilcox et al. vs. Stirnemann et al.) × 4 (6, 7, 8, 9-day windows) = 8 forward windows; 4 (6, 7, 8, 9-day windows) × 3 (known onset of next menses plus two methods based on reported typical cycle length, differing in assumed validity of reports) = 12 backward estimates.

2.3.6. Elimination of cases

With all methods, every case was given a conception probability value, whether continuously or discretely distributed. Of course, researchers can eliminate cases falling on days of the cycle on the fringes of the fertile window, thereby more cleanly distinguishing high from low fertility cases. With continuous measures, such a procedure is akin to a post hoc extreme groups design. As Preacher, Rucker, MacCallum, and Nicewander (2005) note, such procedures rarely increase power, as loss of power due to reduced N offsets increases due to enhanced validity of measurement. Moreover, elimination of data raises suspicion, warranted or not, of post hoc data analytic decisions. Accordingly, they firmly advise against this practice. In the discussion, we address the question of whether elimination of days is advisable.

2.4. Anovulatory cycles

A small proportion of cycles are anovulatory, even among healthy women who regularly cycle. One recent study found that, in 509 cycles of healthy, premenopausal women (mean age = 27) without diagnosed menstrual or ovulatory disorders, approximately 8% were anovulatory (as determined by daily assays of reproductive hormones; Ahrens et al., 2014). Naturally, the presence of anovulatory cycles decreases the validity of counting methods, as such a method will mistakenly judge certain days within anovulatory cycles to be fertile. To assess the impact of anovulatory cycles, we randomly selected 8% of our sample and assigned these cases conception probability values of zero, no matter what the day of the cycle.

We note that undergraduate populations may have higher rates of anovulatory cycles. Using a criterion recommended by Ellison, Lager, and Caffee (1987); failure to reach 300 pmol/L of progesterone during the luteal phase, Roney and Simmons (2013) classified 31% of cycles as anovulatory. They noted that their criterion may be overly conservative and misclassify some ovulatory cycles. Nonetheless, our simulations may under represent anovulatory cycles in young college women.

Of course, hormonal variation, rather than ovulation per se, probably gives rise to cycle shifts (e.g., Roney & Simmons, 2013). Anovulatory cycles are characterized by substantially smaller changes in hormone levels across the cycle, but not an absence of changes altogether (e.g., Ellison et al., 1987). We present findings both taking into account and not accounting for anovulatory cycles.

Table 2
Discrete high fertility windows examined in our analyses

	Forward		Backward
Length (days)	Wilcox et al.	Stirnemann et al.	
6	11–16	10–15	14–19
7	10–16	9–15	14–20
8	10–17	9–16	13–20
9	9–17	8–16	13–21

Notes. Forward windows are ranges of forward counting days. Backward windows are ranges of reverse counting days. Ranges under "Wilcox et al." maximize conception probability using daily values reported by Wilcox et al. (2001). Ranges under "Stirnemann et al." maximize conception probability using daily values reported by Stirnemann et al. (2013).

2.5. Reporting error

Forward estimation procedures assume that first day of last menstruation is reported accurately. Wegienka and Baird (2005) assessed the accuracy of women's retrospective reports. Most women (57%) recalled a day that matched a prospective assessment. But nearly 20% of women's reports were off by more than 3 days. As some error could have been due to unclear instructions, our modeling of error was more optimistic: 66% of women were assumed to report a day that exactly agreed, with error of 1, 2, 3, 4, and 5 days constituting 16%, 8%, 5%, 3%, and 3% of the sample, respectively. Consistent with Wegienka and Baird's (2005) findings, error in the simulated sample increased with day of the cycle (> day 22: 50% spot-on vs. 39% in Wegienka & Baird; < day 7: 85% spot-on vs. 68% in Wegienka & Baird), and underestimation of time passed exceeded overestimation (21% vs. 14%, compared to 25% and 19% in Wegienka & Baird.)

3. Results

3.1. Estimated validities of the measures

Analyses examining validity of our measures were straightforward: Within the total sample of 56,345, we computed Pearson product–moment correlations between estimated conception probability based on possible methods with two sets of values for "true" conception probability (continuous estimates of conception probability; estimates of being in the high- vs. low-fertility window). For measures based on discrete high fertility windows, these values are point-biserial correlations. In short, we refer to these correlation coefficients as "validities." In our initial analyses, we assumed that every cycle is ovulatory, and reports of cycle length are fully accurate; subsequently, we assessed the impact of anovulatory cycles and reporting error.

Assigned or "true" conception probability, based on values from Wilcox et al. (1995), covaried very highly with assigned state of falling within the 5-day fertile window, $r = .956$. Validities estimated using these two criteria, then, are very similar.

Full results are presented in Table 3. Validity coefficients average a very modest .54 but vary substantially (.41–.70).

3.1.1. Estimates based on Wilcox et al. (2001) vs. Stirnemann et al. (2013)

Validities of estimates based on Wilcox et al. and Stirnemann et al. are very similar, mean absolute difference in validity = .017. We see no substantive or practical differences between these estimates of fertility status.

3.1.2. Continuous vs. discrete measures

Uniformly, continuous measures outperform comparable discrete measures. For forward estimates and backward estimates based on reported typical cycle length, the validities of continuous measures average .058 greater than discrete measures. Even when the precise cycle length is known, continuous measures perform better (mean difference = .042). Because luteal phase length is not a constant

Please cite this article as: Gangestad, S.W., et al., How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical imp..., *Evolution and Human Behavior* (2015), http://dx.doi.org/10.1016/j.evolhumbehav.2015.09.001

Table 3
Validities of measures of conception risk/fertility status

	Continuous		Discrete windows			
	Single	Average	6	7	8	9
Conception probability						
Forward	**.521/.555**		.432/.480	.480/.502	.465/.505	.493/.509
Backward – known	**.700/.678**	.654/.655	.659	.671	.661	.650
backward – report .7	.567/.560	**.570/.582**	.472	.495	.502	.513
Backward – report .5	.531/.528	**.550/.565**	.424	.449	.457	.470
In fertile window						
Forward	**.510/.551**		.416/.473	.469/.500	.452/.499	.484/.506
Backward – known	**.704/.690**	.651/.660	.657	.675	.658	.649
Backward – report .7	**.564**/.563	.562/**.582**	.463	.492	.494	.509
Backward – report .5	.527/.529	**.546/.564**	.417	.445	.450	.465

Notes. $N = 56{,}345$. Conception probability, in fertile window: two criterion measures. Forward: a forward estimate of conception risk; backward – known: a backward estimate based on confirmed first day of next menses; backward – report .7: a backward estimate based on self-reported typical cycle length, with validity .7; backward – report .5: a backward estimate based on self-reported typical cycle length, with validity .5. For continuous measures and forward windows, two estimates are reported, separated by a slash: ones based on Wilcox et al. (2001) (before slash); ones based on Stirnemann et al. (2013) (after slash). 6, 7, 8, 9: Number of days in the high fertility window. **Bold** values: highest validities on row, within Wilcox et al./Stirnemann et al. sets.

14 days, a continuous measure captures the effects of luteal phase variation better than discrete windows.

3.1.3. Discrete high-and low-fertility windows

Though use of any discrete high-fertility window is a mistake, given the better continuous measures available, one can nonetheless compare performance across window sizes. Although one review of the literature (Wood et al., 2014; see also Harris et al., 2014) claimed 6-day windows to be more valid than longer windows, this claim is wrong: 6-day windows perform poorly. Longer windows (8–9 days) outperform shorter windows (6–7 days), likely because shorter ones often designate actual high-fertility days as low fertility. An exception is a backward estimate based on a confirmed length of the current cycle. This is understandable: Longer windows hedge bets on true day of ovulation, but a backward count based on a known cycle length narrows the range of likely true days of ovulation. Even in these instances, however, a 7–8 day window is best.

One review of the literature (Wood et al., 2014; see also Harris et al., 2014) claimed 6-day windows to be more valid than longer windows. This claim is wrong: 6-day windows perform poorly.

In addition to the windows listed in Table 2, we examined the validity of several other discrete windows represented in the literature. See SOM, Table S2.

3.1.4. Averages vs. individual continuous measures

When length of the current cycle is precisely known, backward estimates clearly outperform either forward estimates or averages of forward and backward estimates. Indeed, in that circumstance, the validity coefficient is ~ .70. But when reported typical cycle length is used, the validity of forward-backward averages matches or exceeds both forward and backward estimates. Precise validity of the self-report (.7 vs. .5) has little impact: Validities for averages based on such self-reports are ~ .57 and .55, respectively.

3.2. The impact of anovulatory cycles and reporting error

The impact of 8% anovulatory cycles can be summarized simply: No matter what the counting method, validities are reduced by .02–.03, approximately 5%. For instance, the validity of a continuous backward method based on Wilcox et al.'s estimates falls from .70 to .67. Validities of averages of continuous forward and backward estimates fall from roughly .56 to .54, on average. See SOM, Table S3.

The reporting error we modeled reduced validities by 5–6% on average. Backward estimates based on confirmed first day of next menses are not affected by reporting error and hence are not included in this average. See SOM, Table S4.

Jointly, then, anovulatory cycles and reporting error reduce validities of any measure based on a forward estimate by ~10%, on average.

3.3. Power calculations

Choice of a measure has two major implications. First, relatively valid measures lead one to be able to more accurately estimate the true size of a correlation between conception probability and a variable of interest. Second, relatively valid measures increase power to detect, through traditional hypothesis testing, a true effect of conception probability on a psychological variable of interest (or, in Bayesian analysis, find support for a model claiming a true effect exists relative to one positing no effect). Suppose that there exists a true effect of falling within the fertile window on a psychological state equal to Cohen's *d* of .5. Unless fertility status is perfectly measured, measurement error will attenuate that true effect; hence, one is tasked with detecting a manifest effect smaller than .5. Naturally, greater measurement error attenuates the manifest effect size more dramatically. Power is a function of the effect size to be detected and sample size. To achieve the same power to detect an effect of interest using a measure of poorer validity compared to one of superior validity, one must boost sample size.

We estimated power to detect true effect sizes (Cohen's *d*) ranging from .4 (a low-medium effect) to .8 (large effect; Cohen, 1988) for measures with specific, targeted validities:

A validity of 1.0. Though not practically achievable, this value is an ideal comparison.
A validity of .85. Daily LH tests in the presumed high-fertility phase may yield validity of measurement of .8–.9 (see introduction). We assume here a validity of .85.
A validity of .7. This is the highest estimated validity of any counting method we examined: a continuous backward estimate with follow-up confirmation of the accurate next menstrual onset.
A validity of .55. This is the approximate validity of an average of continuous forward and backward estimates based on a self-report of typical cycle length.
A validity of .43. This is a slightly optimistic estimate of the validity (estimated at .41–.43) of a discrete forward estimate with a 6-day window based on Wilcox et al. (2001).

3.3.1. Power: between-subject designs, representative sampling

We first consider statistical power in between-subject designs in which women are representatively sampled from the population of normally ovulating women. In such samples, we expect that approximately 5 women for every 28.5 sampled are in the fertile phase (as 5 days are fertile out of an average cycle of 28.5 days – 17.5%). Unequal representation of fertile and infertile women affects power; e.g., *d* of .5 and equal numbers of fertile and infertile women translate to *r* of .243 (roughly half of *d*), but representative sampling translates to *r* of .187 – nearly 25% smaller.

Table 4 reports sample sizes needed to achieve 80% power (desired) and 70% power (moderate) for true *d* of .4, .5, .6, .7, and .8. Cohen's *ds* of .5 and .8 are considered medium and large effect sizes, respectively. As can be seen, even when highly valid measures of conception probability are used, sample sizes exceeding 200 are needed to achieve 80% power to detect a medium effect size. For example, a backward estimate based on a precisely known date of next menstrual onset has validity of .7 but requires a sample size of ~450 to achieve 80% power to detect a medium effect. A forward estimate based on a discrete 6-day high-fertility window has validity of .43 but requires a sample size of 1200.

Table S5 lists power for *d* of .5 and .8 and sample sizes of 50, 100, and 200. Within this range, only sample sizes nearing 200 and using conception probability measures with validity of at least .7 have reasonable power, but only to detect large effect sizes. Even if validity is .7, power to detect a true effect size of .5 is <50%. When typical day-of-cycle based measures are used, power to detect medium effect sizes with sample size 200 falls below 30%. Table S6 gives power estimates when 8% anovulatory cycles and reporting error are assumed.

Though most between-subject studies conducted to date have sampled women representatively across the cycle, some researchers have attempted to oversample women in the fertile phase by pre-screening women. Naturally, for any given sample size power is improved in such cases, though at the considerable cost of pre-screening many women who do not end up in the sample. Improvement of power in such instances depends on the extent to which the fertile phase is oversampled, and the precise days of the cycle researchers target for oversampling. Given that this particular kind of between-subject studies are relatively rare (Gildersleeve et al., 2014a), we do not estimate power for such instances here. Using similar simulation procedures, Gonzales and Ferrer (2015) provide power estimates for certain specific instances and, like us, estimate power to be relatively low. (For example, they report that a between-subjects study with N = 200 has approximately 20% power to detect a medium effect size.) By downloading and running analyses on our simulation sample tailored to a particular set of targeted days, researchers can estimate power for any specific design.

3.4. Power: Within-subject designs

In within-subject designs, women are typically assessed twice in their cycles: Once in the fertile phase and once in the nonfertile, luteal phase. Power of a within-subject design is enhanced when individual differences across women produce a positive correlation in responses assessed across the fertile and luteal phases. We assessed power of within-subject designs under assumptions of this correlation being .3, .5, and .7.

Table 5 presents results. Naturally, one expects that within-subject designs achieve adequate power with smaller *N* than a between-subjects design. The differences here, however, are dramatic: Depending on the *r* across phases, 10% to 21% the sample size is needed relative to a comparably powered between-subjects design. In our experience, a moderate correlation across phases of .5 or more can typically be expected. In that case, required sample sizes to achieve power with the two designs differ by a factor of ~6.

For instance, a within-subject design that assesses fertility status with LH tests (assumed validity of .85) requires a sample size of just under 50 to achieve 80% power to detect a *d* of .5. A comparable between-subject design using LH sticks would require over 300 participants. Comparable power in between-subjects designs using a backward method with confirmed day of next menses requires close to 500 participants. A between-subjects design using a forward-backward average based on self-reported typical cycle length requires close to 750 participants.

Validities listed – e.g., .7, .55, .43 – correspond to validities for designs that sample women twice during the cycle on representative days, not targeted days, using backward estimate with next menstruation confirmed, average of forward and backward using typical cycle length, and forward estimate using discrete windows, respectively. Many within-subject studies, however, target specific high fertility and low fertility dates. Such designs afford greater power, though how much depends on precisely how days were targeted. We estimate that designs that assess women twice during a cycle during targeted high and low fertility windows, based on counting methods, typically achieve validity ~ .1 greater than that achieved with representative sampling. Hence, for instance, a study that used a backward design with onset of next menstruation confirmed would typically have a validity of

Table 4
Sample size necessary to achieve 80% and 70% power: between-subjects studies with representative sampling

	Cohen's *d*				
	.4	.5	.6	.7	.8
Equivalent *r*	.150	.187	.222	.257	.291
Validity of conception risk measure					
1.0	344	**222**	156	116	90
	271	*175*	*123*	*92*	*71*
.85	477	**309**	217	162	126
	375	*243*	*171*	*127*	*99*
.70	705	**456**	321	239	187
	554	*359*	*252*	*189*	*147*
.55	1143	**740**	521	389	302
	898	*581*	*410*	*306*	*239*
.43	1872	**1213**	854	638	498
	1469	*952*	*671*	*501*	*391*

Notes. *Ns* needed for 80% power given in the top row; *Ns* needed for 70% power given in the bottom row (italicized). Cohen's *d*: true standardized difference between high fertility and low fertility means. Equivalent *r*: Value of *r* Cohen's *d* translates to with representative sampling (5 of every 28.5 women being in the fertile phase). **Bolded** values: Recommended sample size to achieve adequate power. Two-tailed tests assumed.

Table 5
Sample size necessary to achieve 80% and 70% power: within-subjects studies

	Cohen's *d*					
	.5			.8		
r across phases:	.3	.5	.7	.3	.5	.7
1.0	47	**34**	22	20	15	12
	37	*27*	*17*	*16*	*12*	*8*
.85	65	**48**	30	28	21	14
	51	*38*	*24*	*22*	*17*	*12*
.70	96	**71**	45	42	32	22
	76	*56*	*36*	*33*	*25*	*17*
.55	157	**116**	74	69	52	36
	123	*91*	*58*	*54*	*41*	*29*
.43	258	**190**	122	113	86	60
	202	*149*	*96*	*89*	*68*	*47*

Notes. Left-hand column: validity of measurement of conception risk. *Ns* needed for 80% power given in the top row; *Ns* needed for 70% power given in the bottom row (italicized). Cohen's *d*: true standardized difference between high fertility and low fertility means. **Bolded** values: Recommended sample size to achieve adequate power. Two-tailed tests assumed.
Validities listed – e.g., .7, .55, .43 – correspond to validities for designs that sample women twice during the cycle on representative days, not targeted days, using backward estimate with next menstruation confirmed, average of forward and backward using typical cycle length, and forward estimate using discrete windows. Many within-subject studies, however, target specific high fertility and low fertility dates. Such designs afford greater power, though how much depends on precisely how days were targeted. We estimate that designs that assess women twice during a cycle during targeted high and low fertility windows, based on counting methods, typically achieve validity ~ .1 greater than that achieved with representative sampling. Hence, for instance, a study that used a backward design with onset of next menstruation confirmed would typically have a validity of measurement of .7. Yet if high fertility and low fertility days are targeted, validity might increase to .8. 80% power to detect an effect size of .5 then might be achieved with a sample size of about 55, as opposed to 71. Readers should treat these values as guidelines.

measurement of .7. Yet if high fertility and low fertility days are targeted, validity might increase to .8. In such a case, 80% power to detect an effect size of .5 would be achieved with a sample size of about 55, as opposed to 71. Readers should treat values listed in Table 5 as guidelines.

Table S7 presents power of within-subject design studies using measures of varying validity, with sample sizes of 25, 50, and 100 and assuming a correlation of .5 across phases for the dependent measure. Table S8 presents power of within-subject design studies assuming 8% ovulatory cycles and reporting error.

4. Discussion

We aimed to evaluate the validities of various counting methods to assess conception probability. To do so, we simulated a large sample of cycles that demonstrably possess the distributional characteristics of real cycles. Our evaluations yield several broad observations. First, the validities of counting methods are modest, overall. Their median is approximately .5. Virtually all fall short of .6, even ones that account for between-women variation in typical cycle length. Second, to achieve reasonable statistical power, studies using most counting methods require extremely large sample sizes (see also Gonzales & Ferrer, 2015). To be able to detect a *medium* effect size of .5 with 80% power, a between-subjects study that uses a measure with validity near the median (.5) demands a sample size of about 900 or over 1000 when anovulatory cycles and reporting errors are present.

These findings have important implications. First, they inspire a set of recommendations going forward. Second, they inform interpretations of the current literature.

4.1. Recommendations for future research

Our findings yield clear recommendations for future research examining shifts across the cycle.

4.1.1. Recommendation 1: a within-subject design should generally be the design of choice

Between-subject studies of cycle effects require very large sample sizes to achieve acceptable levels of statistical power. Even if the most valid counting method is used to measure conception probability, 80% power to detect a medium effect size requires a sample size nearing 500 – and that method requires a follow-up confirmation of first day of next menstruation. The best method based on a single session – an average of continuous forward and backward estimates – demands a sample size exceeding 700. By contrast, a within-subject design can achieve comparable power with a sample size of 50–80. We suspect that researchers will typically find within-subject studies to be more efficient. Based on evaluations of particular kinds of between-subject designs, Gonzales and Ferrer (2015) offered a similar recommendation.

One notable exception may be when women are recruited to complete questionnaires online. Particularly when researchers wish to access a non-college population fairly cheaply or avoid the potential problem of sensitizing women to researchers' interest in cycle shifts, between-subject studies with a large *N* might be reasonable. However, between-subject studies with an *N* less than 500 cannot be recommended; most methods call for *N* > 700. Researchers opting for these designs should interpret manifest effect sizes with caution. Relatively low validities of these methods for determining effects of fertility will reduce estimates of effect sizes, on average, by about half relative to the ideal of a validity of 1.0.

Naturally, the potential weaknesses of within-subject designs should also be kept in mind. If presented with a stimulus two or more times, individuals may recall their previous responses and, in an effort to appear consistent, may give the same response. Accordingly, in one study researchers instructed participants to "answer as you feel *now*, which could be different than how you *usually* feel" (Gangestad, Thornhill, & Garver-Apgar, 2010). Cantú et al. (2014) created two different stimulus sets, and each woman responded to each one just once – one during her high fertility session and the other during her low fertility session (with high and low fertility stimuli counterbalanced across women). Researchers should implement procedures that minimize carry-over effects in within-subject designs when possible.

4.1.2. Recommendation 2: in within-subject studies, detection of LH surges and a continuous backward estimate with confirmed onset of next menses are methods of choice

Within-subject designs are best if they incorporate a highly valid means of assessing the fertile phase. Two methods yield 80% power to detect medium effect sizes with $N < 100$: detection of LH surge and a backward estimate with confirmed onset of next menses. We recommend sample sizes of 50 + and 80 +, respectively (see Tables 5, S7, and S8). Stunningly, within-subject studies with these sample sizes can be as informative as between-subject studies of 1000 + (specifically, if the correlation between the dependent variable assessed at low fertility and high fertility is .5 or greater).

Researchers will not detect LH surges for some women recruited for participation – typically, up to one third of the recruited sample (e.g., Gangestad et al., 2005; Larson et al., 2013). Some of these women are precisely those one wishes to exclude, as their current cycle is anovulatory or irregular, with day of ovulation not well predicted by a counting method. Nonetheless, if researchers target a final sample size of 50 women, they might need to recruit 75 women, close to the same number required for a backward count method with confirmed onset of next menses (with 10–15% attrition due to lack of follow-up; e.g., Larson et al., 2013). Perhaps the primary trade-off dictating choice between these two methods, then, is the benefit of increased certainty of confirmed ovulation with LH sticks and the cost in money, time, and effort that LH surge detection entails. If taken for 5 consecutive days in the presumed high-fertility window, LH sticks typically run < $10 per participant.

4.1.3. Recommendation 3: assay reproductive hormones instead

As ovulatory cycle shifts likely arise as a function of changing hormonal levels, studies that examine covariation of estradiol, progesterone, and testosterone levels with variables of interest across time are desired (e.g., Grillot, Simmons, Lukaszewski, & Roney, 2014; Puts et al., 2013; Welling et al., 2007). They do not require researchers to assess timing of ovulation per se. We recommend this form of study independent of the validity of conception probability measures (as specific hormones may have different effects; e.g., DeBruine, Jones, & Perrett, 2005; Jones et al., 2005) but recognize that they are costly. Given low validity of most counting methods, however, some researchers might find these costs worth their expense. Once again, within-subject studies are most powerful. Because hormonal effects are likely a function of both hormone levels and tissue-specific receptor density (for which there could be meaningful individual differences), researchers may be interested in examining both within-woman and between-woman hormonal correlates (see Roney & Simmons, 2013, on potential time-lagged correlation). If researchers sample hormones every other day or more frequently, they can also identify the timing of ovulation within the study (see, e.g., Roney & Simmons, 2013; see also Puts et al., 2013).

4.1.4. Exclusion of days and loss of participants

One research strategy is to collect data on a large sample of women and then remove participants whose sessions do not, with at least modest probability, fall into the fertile and nonfertile phases. Classification is relatively accurate in the resulting sample. Hence, for instance, one can classify the 5 days running from 10–14 (those with conception probability > .07; Wilcox et al., 2001) as "fertile" and the days 1–7 and 21 and greater (those with conception probability < .02; Wilcox et al., 2001) as "non-fertile," leaving the 8 days from 8–10 and 16–20 unclassified and therefore unanalyzed. Whereas a forward 6-day window based on Wilcox et al. has a validity of .43, this classification variable has a validity of .60.

But will this approach increase power? No. The benefits of enhanced validity of measurement are offset by the reduced sample size. For example, with validity of .60, the 5-day discrete fertile window above yields 80% power to detect an effect size of $d = .5$ with a sample of 622. However, it eliminates, on average, 29% of a sample. Therefore, to achieve a final sample size of 622, one should run 860 women. Averaging forward and backward continuous estimates based on typical cycle length yields 80% power with740 women.

Trimming a sample has other costs, too: it eliminates potentially informative data and raises suspicions of post hoc data analytic decisions (e.g., see Gelman & Loken, 2014). Just as Preacher et al. (2005) do not recommend ad hoc extreme group selection, we do not recommend trimming a sample to more cleanly define high- and low-fertility groups.

Rather than trimming a sample, researchers can oversample women in the fertile phase of the cycle by using a pre-screening instrument. Naturally, pre-screening to target specific days of the cycle has effects on power similar to the effect of trimming a sample. Yet Gonzales and Ferrer (2015) estimated that large sample size (>200) is typically necessary even when researchers sample just two sets of days in a between-subject design: A 1–6 day high fertility window and a low fertility window during the luteal phase.

4.2. Evaluations of the existing research findings: statistical power and effect size estimation

Potentially, our findings have profound implications for interpreting the extant literature on ovulatory cycle shifts (see also Gonzales & Ferrer, 2015). As noted earlier, findings are variable – e.g., 40% of Gildersleeve et al.'s (2014a) subsample of 42 studies targeting core mate preference shifts produced significant effects. As some of these studies yielded mixed findings – some significant, others not - effects of interest were detected at a rate closer to 30%. But as also noted, *p*-curves of significant effects are robustly right-skewed, with a large proportion of them being < .01. A *p*-curve's right skew is purportedly a signature of real non-zero effects (Simonsohn et al., 2014 - though we acknowledge that *p*-curve analysis is a relatively new technique and requires additional evaluation). Thus, while at least some findings may well reflect true effects, significant effects are not detected in most studies. What explains this pattern?

Statistical power plays a role. In 100 exact replicates of a study, each with precisely 30% power to detect a true effect, 30% will, on average, detect the effect. The same is true of a set of studies of varying sample sizes assessing effects of heterogeneous size, with a mean of 30% power. The *p*-curves Gildersleeve et al. (2014b) presented yield estimates of mean power in the studies entered into it. The pattern of observed *p*-values closely follows the theoretical curve expected if power is 33%. (A larger sample of effects yielded the same estimate; Gangestad, Grebe, Gildersleeve & Haselton, unpublishEd.) Hence, power roughly matches the rate of positive effects observed in Gildersleeve et al.'s (2014a) narrow sample.

As noted above, power, in turn, is a function of effect size and sample size. So why is power poor – because true effects are exceedingly weak or because *N* is insufficient to detect meaningful true effect sizes? Here, our findings are pertinent: In light of the weak validity of counting methods for the assessment of conception probability, the sample size of most studies in the literature may possess woefully inadequate power to detect *even medium to large true effect sizes*.

Though a review of past findings is beyond the scope of this paper, we use the Gildersleeve et al. (2014a) meta-analysis sample to illustrate this point. Of the 42 studies in the "narrow" sample, 24 and 18 studies implemented between-subjects and within-subject designs, respectively. Just 3 of 42 studies assessed LH, and 4 used a backward method with confirmed next menstrual onset. Of the remaining studies, about half (18) exclusively relied on a forward counting method, and the rest (17) used backward counting or a mixture of methods. At the median, then, conception probability estimations likely had validity ~ .50. The median sample sizes were 118 and 29for between- and within-subjects designs, respectively. Given validity of .50 and these sample sizes, power to detect a medium effect size of .5 would be 17% and 24%, respectively, which yields a weighted average of 20%. Power to detect a large effect size of .8 would be 35% and 41%, which yields a weighted average of 37%. It is hardly surprising, then, that fewer than half of these studies produced significant effects, even if true effects are, on average, of medium or large size.

Of course, some studies had even weaker power. For instance, Rupp et al.'s (2009) between-subject study of 13 women, using a continuous forward estimate (estimated validity = .52), had 5% power to detect a *d* of .5. More surprisingly, some large studies yielded stunningly low power too. Harris's (2011) study of 258 women's fertility status using a discrete forward estimate of 8 days (days 6–14; estimated validity = .43), for instance, had an astonishingly small 25% power to detect a *d* of .5.

Naturally, if the proportion of effects in these studies that were significant slightly exceeds their median power to detect an effect size of .5, then one might also expect an average effect size of about .5. Gildersleeve et al. (2014a, 2014b) estimated mean effect sizes of .26 and .20 for effects on preferences in short-term and unspecified mating contexts, while their *p*-curve yielded an estimate of .30. A mean observed effect size of ~ .25 may seem inconsistent with true effect size of ~ .5. Yet a true effect size of .5 means that *true* high and low fertility group means differ by .5 of a standard deviation. Once again, when measurement of conception probability is poor, manifest effect size falls well short of.5. Specifically, if validity of measurement is .5 and true effect size is .5, manifest effect size is expected to be .24– close to values Gildersleeve et al. (2014a, 2014b) report. Hence, a true mean effect size of .5 is consistent with findings to date in light of low validity of methods used to assess conception probability.

To propose that low power possibly explains, in part, the mixed nature of results in studies to date is not to argue that every preference shift examined to date is real. Indeed, some recent studies that have failed to find cycle shifts *have* had considerable power. Zietsch, Lee, Sherlock, and Jern (2015) and Munoz-Reyes et al. (2014), for instance, examined the association between fertility status and preference for facial masculinity in sample sizes close to 600 and 500, respectively. They should have had 60–70% power to detect medium effects (see Table 4), but neither study detected an effect, with mean effect size close to zero. As other recent studies examining cycle shifts in *other* preferences *have* found positive effects (e.g., Cantú et al., 2014; Giebel, Weierstall, Schauer, & Elbert, 2013), we suspect that another reason for mixed results in this area is heterogeneity of true effects across different kinds of preferences (see Gangestad et al., unpublished). Whereas some preference shifts may be robust and substantial, others may be negligible. To identify which effects are robust, additional, appropriately powered studies are needed.

4.3. Summary and conclusions

Psychological effects of the ovulatory cycle have garnered increasingly broad interest in the evolutionary and social sciences, spurring many dozens of studies and considerable controversy over their robustness. An unusual feature of this literature is the exceptionally broad diversity of methods used to assess the key variable in question – fertility within the cycle. We sought to empirically estimate validities of these methods using a large set of simulated cycles whose distributional characteristics closely match those of real cycles. Results were striking, and yield two outcomes: (1) a set of recommendations for researchers and (2) important implications for understanding the true magnitude of cycle shift phenomena.

Please cite this article as: Gangestad, S.W., et al., How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical imp..., *Evolution and Human Behavior* (2015), http://dx.doi.org/10.1016/j.evolhumbehav.2015.09.001

If researchers adopt the methodological standards we suggest, several welcome advances should follow. First, by following these standards, researchers will help to assuage concerns that methodological flexibility has produced false positives in an absence of true cycle shifts. Second, more uniform standards will allow for comparison across studies of psychological variables of interest, identifying where cycle shifts are present and absent. Third, if our analysis of the extant literature in light of low validity is correct, higher validity methods are likely to reveal cycle shifts considerably larger and more robust than previous estimates. If such findings indeed emerge as methods improve, they will shed light on a potentially important role for fertility in regulating human social behavior, paralleling widely established patterns in our nonhuman cousins.

Supplementary Materials

Supplementary methods, results, and documentation to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2015.09.001.

References

Ahrens, K. A., Vladutir, C. J., Mumford, C. L., Schliep, K. C., Perkins, N. J., Wactawski-Wende, J., & Schisterman, E. F. (2014). The effect of physical activity across the cycle on reproductive function. *Annals of Epidemiology*, *24*, 127–134.

Baird, D. D., McConnaughey, R., Weinberg, C. R., Musey, P. I., Collins, D. C., Kesner, J. S., ... Wilcox, A. J. (1995). Application of a method for estimating day of ovulation using urinary estrogen and progesterone metabolites. *Epidemiology*, *6*, 547–550.

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., ... Rowland, H. M. (2015). Women's changes in skin color across the ovulatory cycle are not detectable by the human visual system. *PLoS One*, http://dx.doi.org/10.1371/journal.pone.0130093.

Cantú, S. M., Simpson, J. A., Griskevicius, V., Weisberg, J. Y., Durante, K. M., & Beal, D. J. (2014). Fertile and selectively flirty: Women's behavior toward men changes across the ovulatory cycle. *Psychological Science*, *25*, 431–438.

Cobey, K. D., Buunk, A. P., Pollet, T. V., Kippling, C., & Roberts, S. C. (2013). Men perceive their female partners, and themselves, as more attractive around ovulation. *Biological Psychology*, *94*, 513–516.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, L. A., Ladner, D. G., & Byrn, F. W. (2009). The normal variabilities of the menstrual cycle. *Fertility and Sterility*, *91*, 522–527.

DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2005). Women's attractiveness judgments of self-resembling faces change across the menstrual cycle. *Hormones and Behavior*, *47*, 379–383.

Direito, A., Bailly, S., Mariani, A., & Ecochard, R. (2013). Relationships between the luteinizing hormone surge and other characteristics of the menstrual cycle in normally ovulating women. *Fertility and Sterility*, *99*, 279–285.

Durante, K. M., Griskevicius, V., Hill, S. E., Perilloux, C., & Li, N. P. (2011). Ovulation, female competition, and product choice: Hormonal influences on consumer behavior. *Journal of Consumer Research*, *37*, 921–934.

Ellison, P. T., Lager, C., & Caffee, J. (1987). Low profiles of salivary progesterone in college undergraduate women. *Journal of Adolescent Health Care*, *8*, 204–207.

Fales, M. R., Gildersleeve, K. A., & Haselton, M. G. (2014). Exposure to perceived male rivals increases testosterone on fertile days relative to nonfertile days of their partner's ovulatory cycle. *Hormones and Behavior*, *65*, 454–460.

Fehring, R., Schneider, M., & Raviele, K. (2006). Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, *35*, 376–384.

Gangestad, S. W., Garver-Apgar, C. E., Cousins, A. J., & Thornhill, R. (2014). Intersexual conflict across the ovulatory cycle. *Evolution and Human Behavior*, *35*, 302–308.

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.

Gangestad, S. W., & Thornhill, R. (1998). Menstrual cycle variation in women's preference for the scent of symmetrical men. *Proceedings of the Royal Society of London B*, *262*, 727–733.

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.

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2005). Women's sexual interests across the ovulatory cycle depend on primary partner fluctuating asymmetry. *Proceedings of the Royal Society of London B*, *272*, 2023–2027.

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2010). Fertility in the cycle predicts women's interest in sexual opportunism. *Evolution and Human Behavior*, *31*, 400–411.

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2015). Women's sexual interests across the ovulatory cycle: Function and phylogeny. In D. M. Buss (Ed.), *Handbook of evolutionary psychology* (2nd Ed.). New York: Wiley.

Garver-Apgar, C. E., Gangestad, S. W., & Thornhill, R. (2008). Hormonal correlates of women's mid-cycle preference for the scent of symmetry. *Evolution and Human Behavior*, *49*, 223–232.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Giebel, G., Weierstall, R., Schauer, M., & Elbert, T. (2013). Female attraction to appetitive aggressive men is modulated by women's menstrual cycle and men's vulnerability to traumatic stress. *Evolutionary Psychology*, *11*, 248–262.

Gildersleeve, K., & Haselton, M. G. (2014). Are women more attractive at high fertility? A meta-analytic review. Unpublished manuscript.

Gildersleeve, K., Haselton, M. G., & Fales, M. (2014a). Do women's mate preferences change across the ovulatory cycle?: A meta-analytic review. *Psychological Bulletin*, *40*, 1205–1259.

Gildersleeve, K., Haselton, M. G., & Fales, M. (2014b). Meta-analyses and p-curves support robust cycle shifts in mate preferences: Response to Wood & Carden and Harris, Pashler & Mickes (2014). *Psychological Bulletin*, *40*, 1272–1280.

Gildersleeve, K., Haselton, M. G., Larson, C. M., & Pillsworth, E. G. (2012). Body odor attractiveness as a cue of impending ovulation in women: Evidence from a study using hormone-confirmed ovulation. *Hormones and Behavior*, *61*, 157–161.

Gonzales, J. E., & Ferrer, E. (2015). Efficacy of methods for ovulation estimation and their effect on the statistical detection of ovulation-linked behavioral fluctuations. *Behavior Research Methods*, 1–20.

Grammer, K. (1993). 5-α-androst-16en-3α-on: A male pheromone? A brief report. *Ethology and Sociobiology*, *14*, 201–208.

Grillot, R., Simmons, Z. L., Lukaszewski, A. W., & Roney, J. R. (2014). Hormonal and morphological predictors of women's body attractiveness. *Evolution and Human Behavior*, *35*, 176–183.

Guermandi, E., Vegetti, W., Bianchi, M. M., Uglietti, A., Ragni, G., & Crossignani, P. (2001). Reliability of ovulation tests in infertile women. *Obstetrics and Gynecology*, *97*, 92–96.

Harper, M. J. K. (1994). Gamete and zygote transport. In E. Nobil, & J. D. Neill (Eds.), *The Physiology of Reproduction* (pp. 123–187) (2nd ed.). New York, NY: Raven Press Ltd.

Harris, C. R. (2011). Menstrual cycle and facial preferences reconsidered. *Sex Roles*, *64*, 669–681.

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, & Fales (2014). *Psychological Bulletin*, *40*, 1260–1264.

Haselton, M. G., & Gildersleeve, K. A. (2011). Can men detect ovulation? *Current directions in psychological science*, *61*, 157–161.

Jöchle, W. (1973). Coitus induced ovulation. *Contraception*, *7*, 523–564.

Jones, B. C., Little, A. C., Boothroyd, L., DeBruine, L. M., Feinberg, D. R., Law Smith, M. J., ... Perrett, D. I. (2005). Commitment to relationships and preferences for femininity and apparent health in faces are strongest on days of the menstrual cycle when progesterone level is high. *Hormones and Behavior*, *48*, 283–290.

Jukic, A. M. Z., Weinberg, C. R., Wilcox, A. J., McConnaughey, D. R., Hornsby, P., & Baird, D. D. (2008). Accuracy of reporting of menstrual cycle length. *American Journal of Epidemiology*, *167*, 25–33.

Larson, C. M., Haselton, M. G., Gildersleeve, K. A., & Pillsworth, C. G. (2013). Changes in women's feelings about their romantic relationships across the ovulatory cycle. *Hormones and Behavior*, *63*, 128–135.

Larson, C. M., Pillsworth, C. G., & Haselton, M. G. (2012). Ovulatory shifts in women's attractions to primary partners and other men: Further evidence of the important of primary partner sexual attractiveness. *PLoS One*, *7*, e44456. , http://dx.doi.org/10.1371/journal.pone.0044456.

Lenton, E. A., Landgren, B. M., & Sexton, L. (1984). Normal variation in the length of the luteal phase of the menstrual cycle. *British Journal of Obstetrics and Gynaecology*, *91*, 685–689.

Little, A. C., Jones, B. C., & Burriss, R. P. (2007). Preferences for masculinity in male bodies change across the menstrual cycle. *Hormones and Behavior*, *31*, 633–639.

Munoz-Reyes, Iglesias-Julios, M., Martín-Elola, C., Losada-Perez, M., Monedero, I., Pita, M., & Turiegano, E. (2014). Changes in preference for male faces across the menstrual cycle in a Spanish population. *Anales de Psicología*, *30*, 667–675.

Neuberg, S. L., Kenrick, D. T., & Schaller, M. (2010). Evolutionary social psychology. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 761–796). New York: John Wiley & Sons.

Park, S. J., Goldsmith, L. T., Skurnick, J. H., Wojtczuk, A., & Weiss, G. (2007). Characteristics of the urinary luteinizing hormone surge in young ovulatory women. *Fertility and Sterility*, *88*, 684–690.

Penton-Voak, I. S., Perrett, D. I., Castles, D., Burt, M., Koyabashi, T., & Murray, L. K. (1999). Female preference for male faces changes cyclically. *Nature*, *399*, 741–742.

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.

Preacher, K. R., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192.

Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior*, *26*, 388–397.

Puts, D. A. (2006). Cyclic variation in women's preferences for masculine traits: Potential hormonal causes. *Human Nature*, *17*, 114–127.

Puts, D. A., Bailey, D. H., Cardenas, R. A., Burriss, R. P., Welling, L. L. M., Wheatley, J. R., & Dawood, K. (2013). Women's attractiveness changes with estradiol and progesterone across the ovulatory cycle. *Hormones and Behavior*, *63*, 13–19.

Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of women's sexual desire in normal menstrual cycles. *Hormones and Behavior*, *63*, 636–645.

Rupp, H. A., James, T. W., Ketterson, E. D., Sengelaub, D. R., Janssen, E., & Heiman, J. R. (2009). Neural activation in the orbitofrontal cortex in response to male faces increases during the follicular phase. *Hormones and Behavior*, *56*, 66–72.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534–537.

Small, C. M., Manatunga, A. K., & Marcus, M. (2007). Validity of self-reported menstrual cycle length. *Annals of Epidemiology*, *17*, 163–170.

Stirnemann, J. J., Samson, A., Bernard, J. -P., & Thalabard, J. -C. (2013). Day-specific probabilities of conception in fertile cycles resulting in spontaneous pregnancies. *Human Reproduction*, *28*, 1110–1116.

Symons, D. (1979). *The evolution of human sexuality*. New York: Oxford University Press.

Thornhill, R., & Gangestad, S. W. (2008). *The evolutionary biology of human female sexuality*. New York: Oxford University Press.

Wegienka, G., & Baird, D. D. (2005). A comparison of recalled date of last menstrual period with prospectively recorded dates. *Journal of Women's Health*, *14*, 248–252.

Welling, L. L. M., Jones, B. C., DeBruine, L. M., Conway, C. A., Law Smith, M. J., Little, A. C., ... Al-Dujaili, E. A. (2007). Raised salivary testosterone in women is associated with increased attraction to masculine faces. *Hormones and Behavior*, *52*, 156–161.

Wilcox, A. J., Duncan, 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., Weinberg, C. R., & Baird, B. D. (1995). Timing of sexual intercourse in relation to ovulation. *New England Journal of Medicine*, *333*, 1517–1521.

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

Zietsch, B. P., Lee, A. J., Sherlock, J. M., & Jern, P. (2015). Variation in women's facial masculinity preference is better explained by genetic differences than by previously identified context-dependent effects. *Psychological Science*, *6*, 1440–1448.

Please cite this article as: Gangestad, S.W., et al., How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical imp..., *Evolution and Human Behavior* (2015), http://dx.doi.org/10.1016/j.evolhumbehav.2015.09.001