Evolution and Human Behavior 37 (2016) 85–96

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
Final revision received 8 September 2015

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/© 2016 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.

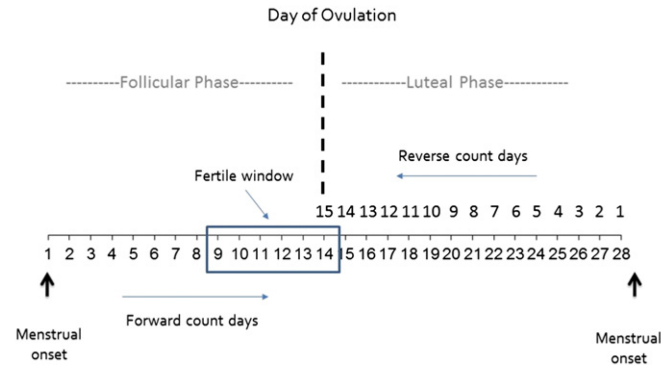

Fig. 1. Hypothetical cycle of a woman whose cycle length was 28 and day of ovulation was day 14. The "fertile window" in this cycle extends from forward count day 9 to forward count day 14. By the reverse count, her day of ovulation was day 15, and her fertile phase was reverse count day 20 to reverse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