Interpreting Past Research When Aware of QRPs and QIPs

Hal Pashler

UCSD

Progress on Improving Research Methods Going Forward has been Dramatic!

1. Pre-Reviewed Paper Models
   (new weapons against HARKing & CARKing).
2. Upsurge in people doing replications.
3. Upsurge in willingness of journals to publish replications.
4. Many Labs & Reproducibility Projects.
5. Awareness of and stigmatization of p-hacking.

etc.

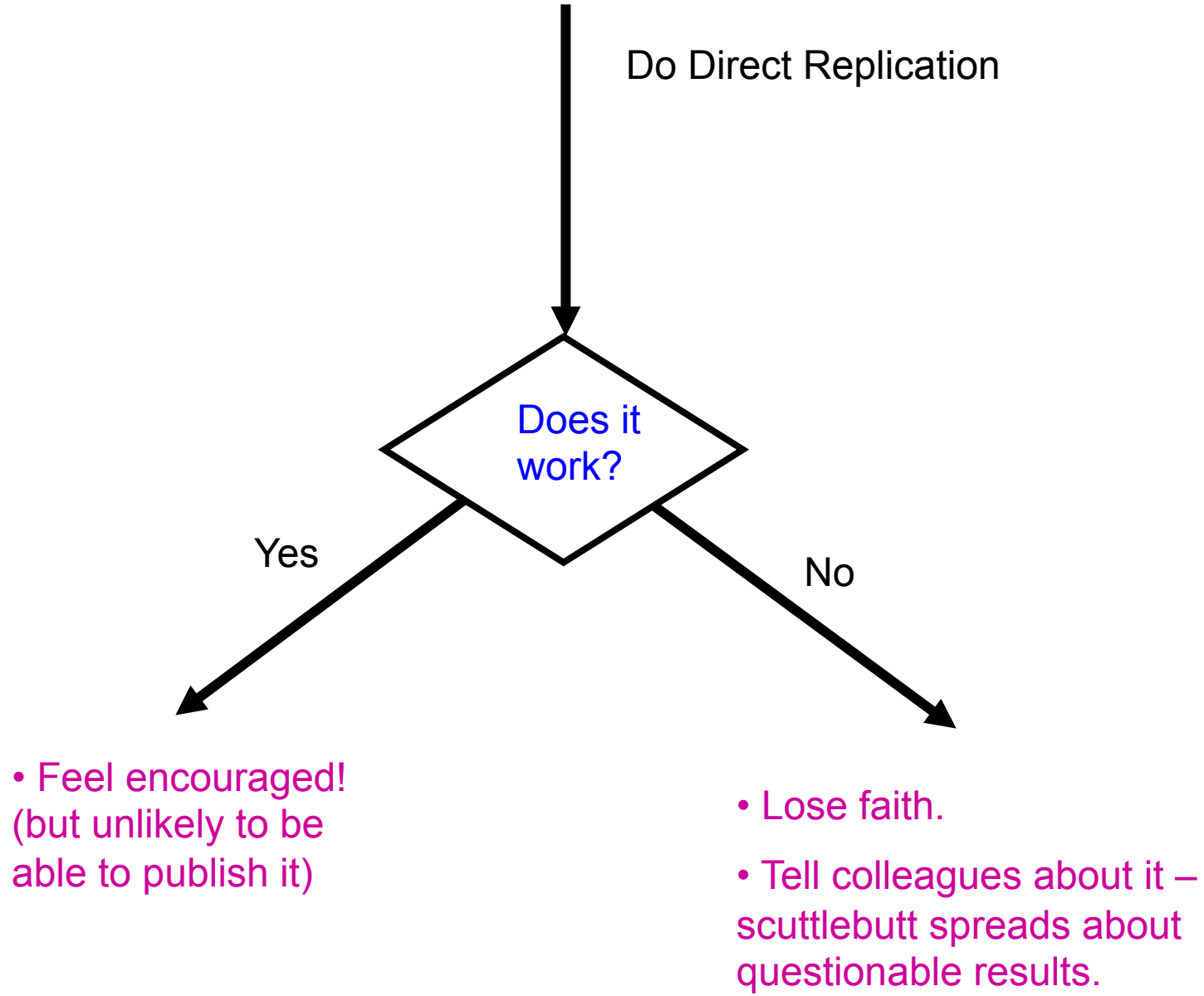
What about Looking Backward?

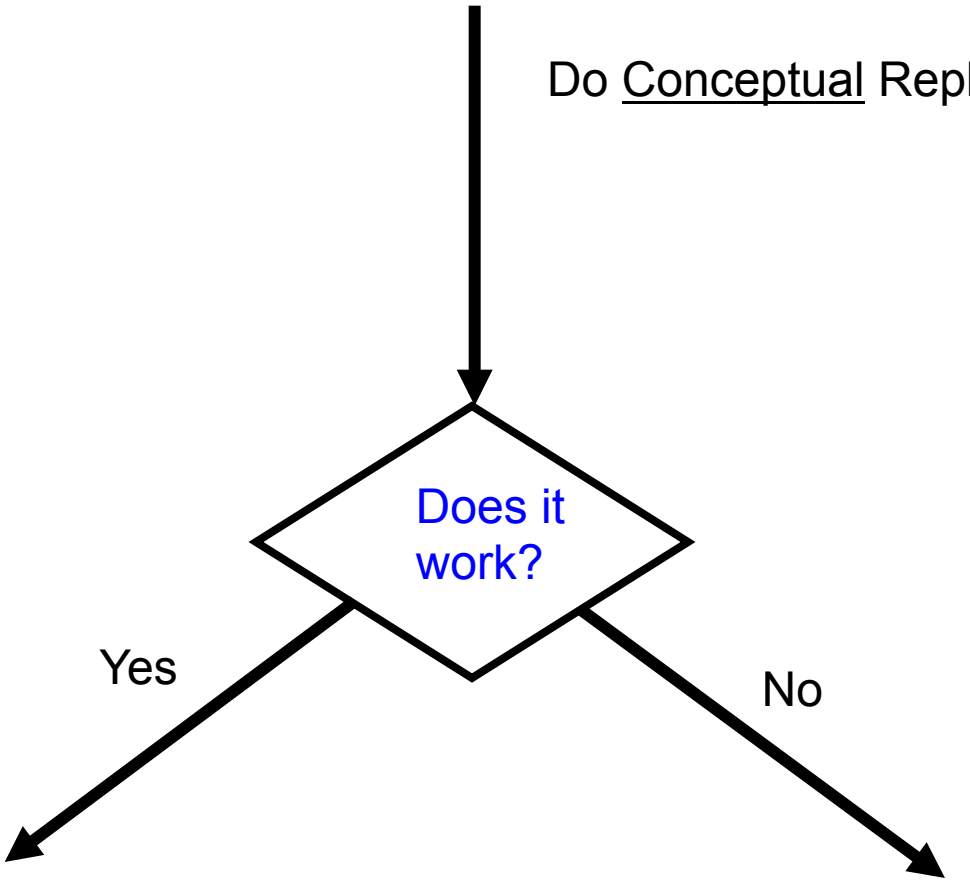
Attempting to extract reliable meaning from the literatures we've got in many cases leads us into dark labyrinths from which neither old advice nor new-fangled techniques offer any hope of escape.

In writing a review article, textbook, popular overview, what are best practices for interpreting experimental literatures?

Old Rule #1. Trust findings for which there are numerous (and ideally, diverse) “conceptual replications”: multiple experiments which involve same basic finding with different populations, manipulations, and measures.

When someone does a Direct or Conceptual replication, and they either succeed or fail, what happens then?


Do Direct Replication
Does it work?
Yes
No
• Feel encouraged! (but unlikely to be able to publish it)
• Lose faith.
• Tell colleagues about it – scuttlebutt spreads about questionable results.



- Publish it!

- No loss in faith in original result.
- "I shouldn't have changed so many things…"

For further discussion see Pashler & Harris,

Perspectives on Psychological Science November 2012 *7: 531-536,*

Consequence:

Any field in which people do

no direct replications

and

many conceptual replications

(particularly if underpowered and/or boosted up with p-hacking) will naturally spawn large literatures in which even non-effects appear to enjoy diverse empirical support. The file-drawer of failed attempts is not only invisible to outsiders—it also doesn't even shake the faith of the investigators whose attempts have failed.

Conclusion: Highly compelling-looking literatures fully consistent with complete absence of real effects.

Several parameters likely to govern magnitude of distortion:

- ** attractiveness of topic
  (exciting, easy)
- ** prevalence of QRPs—culture of the field.
- ** cost of obtaining good power.

Can Meta-Analysis Save Us?

Common corrections for publication bias (Trim and Fill) shown inadequate in simulations.

P-Curve?

** Assumes very specific form of p-hacking.

** Assumes no fraud.

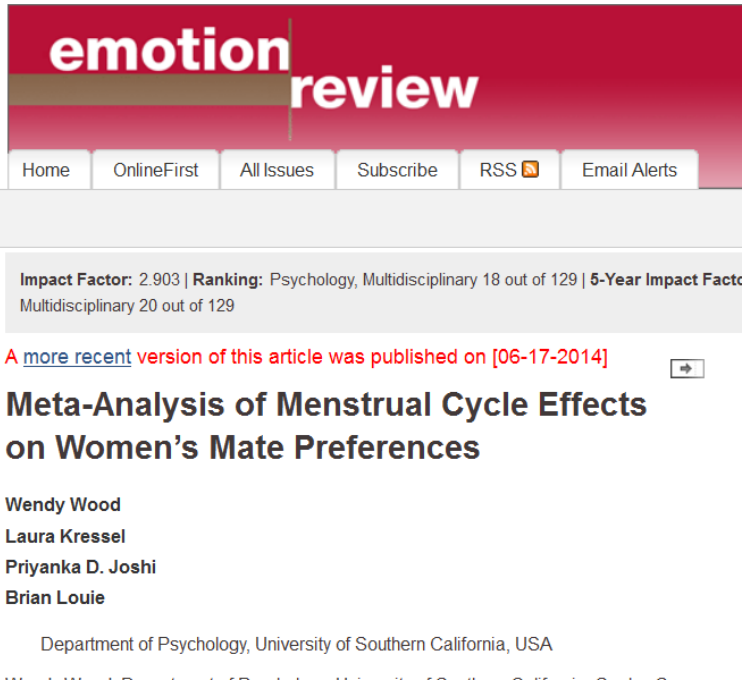
Psychological Bulletin

0033-2909/14/$12.00 DOI: 10.1037/a0035438

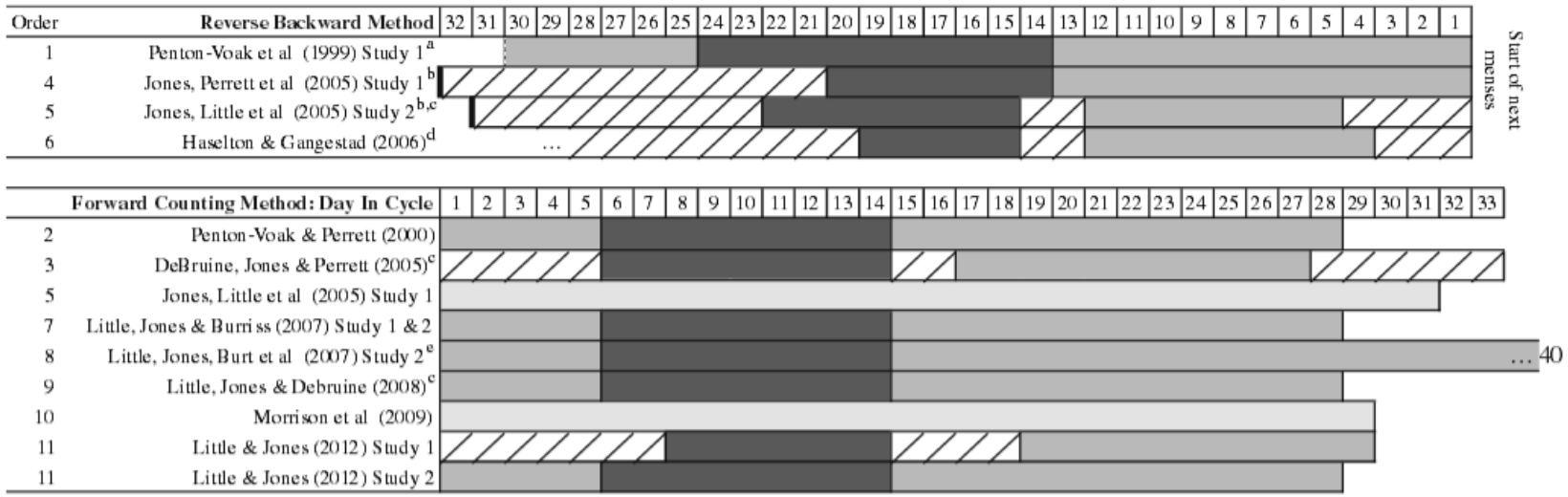
Do Women's Mate Preferences Change Across the Ovulatory Cycle? A Meta-Analytic Review

Kelly Gildersleeve, Martie G. Haselton, and Melissa R. Fales
University of California, Los Angeles

versus



But Christine Harris et al. analysis of evidence for p-hacking in evo-psych fertility effect literature shows: fertility criteria change from study to study, even within lab!



Note Studies are numbered (see far left) in chronological order of online publication date Contrary to claims by Gildersleeve et al , differences in calculation methods do not represent "a coherent progression in methodology" (e g , backward counting methods do not appear in later studies -- see "order" column) The top panel shows variability in studies that use a backward method to determine fertility (counting backwards from the first day of a woman's next period) The lower panel shows variability in studies that determine fertility by counting forward from the first day of a woman's period There is flexibility not only in method, but in which days make up the fertile group, the not fertile group and which women are not included in either group Some researchers also use a probability risk assessment for each day (see solid grey lines)

[a] While the range of cycle lengths was not reported, mean range was (shown by dotted bar) Number of days counted for menses was also not reported Therefore, the range for this and for the fertility phase are best estimates

[b] Cycle length range (shown by thick bar) denotes maximum average cycle length of participants

[c] Study also conducted continuous conception risk analysis

[d] Collected data from participants for 35 days Onset and duration of menses were collected but cycle length data were not reported

[e] The methods are contradictory, "we used a standard 28-day model of the female menstrual to divide women into high (days 6–14) and low (days 0–5 and 15–28) conception risk based on self-report" (p 214); yet women who were on days 29-40 were included as part of the nonfertile group (Full cycle range is not visually indicated here in order to preserve legibility of figure)

Fig. 1 Demonstration of the variability of fertility classification for self-report methods by a single research Group (Jones, Little, and Colleagues)

Can Meta-Analysis Save Us?

Common corrections for publication bias (Trim and Fill) shown inadequate in simulations.

P-Curve?

** Assumes very specific form of p-hacking.

** Assumes no fraud.

My Hunch

Many behavioral science literatures contain such a generous admixture of so many kinds of junk there is no realistic hope of inverting the distortion.

But on a more positive note… are there any overlooked rays of hope for extracting reliable information from existing literatures?

Meta-analysis with heroic-level efforts to dig up the unpublished literature?

Journal of Experimental Psychology: General

0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000083

A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource

AQ: au

Evan C. Carter
University of Miami and University of Minnesota

Lilly M. Kofler
University of Miami and University of Chicago

Daniel E. Forster and Michael E. McCullough
University of Miami

Cf Ernest O'Boyle's comparison of dissertations vs. articles

Can we be on Lookout for Fields/Designs where Publication Bias is Unusually Weak or even Absent?

Jennions says Ethology: standard follow-on study asks “Does this work with species X?” Answer publishable either way.

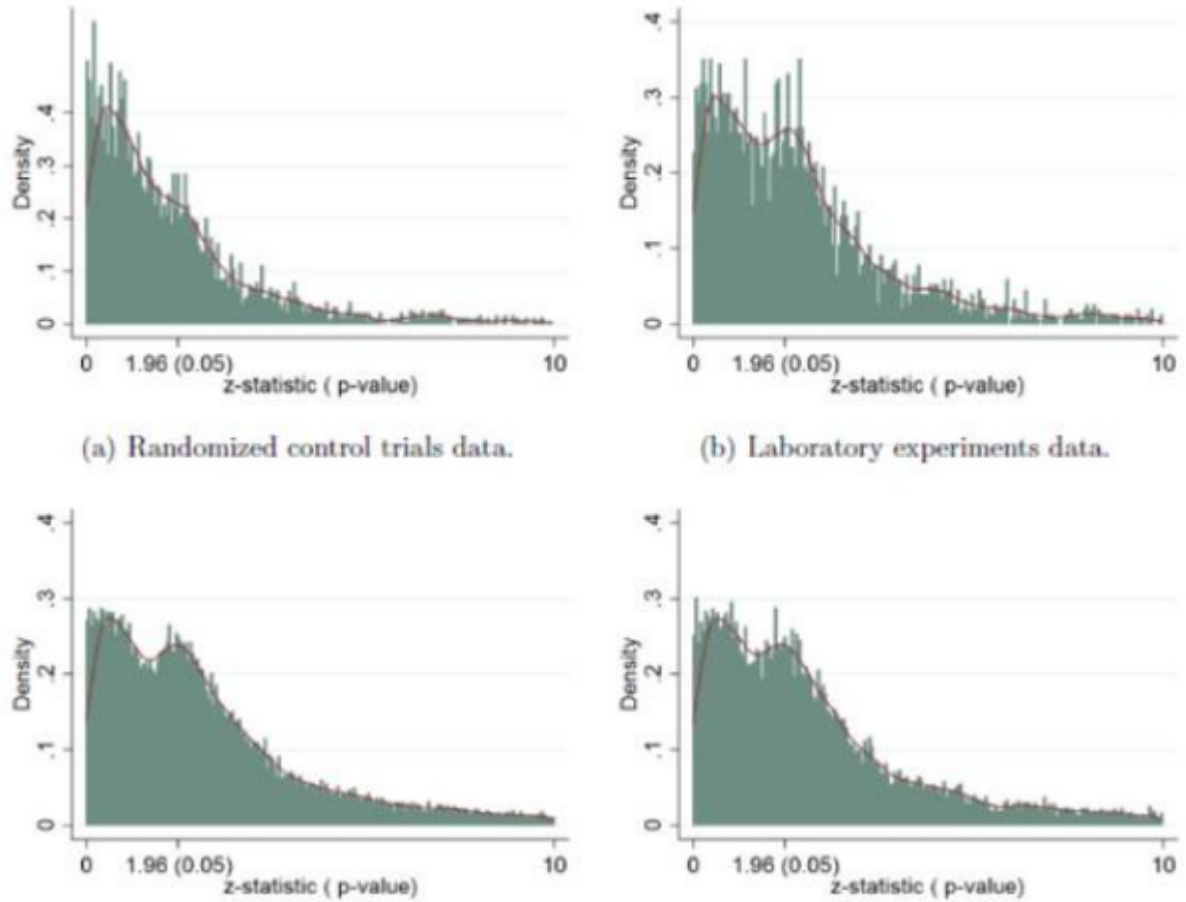
Behavior genetics/twin research. (eg Plomin)

Others? Answer probably well known to people in each field, but will they speak honestly about this?

Robert Dur
@DurRobert

@BrendanNyhan No camel shape in distribution of p-values for field-experimental studies though: pic.twitter.com/SN9uM9qIEl

Figure 9: Distributions of z-statistics for different sub-samples: type of data and journal.



(a) Randomized control trials data.

(b) Laboratory experiments data.

(c) Other sources of data.

(d) Journal 1.

There are probably many more specific fields that have abnormally low bias. Can they be identified?

Does it matter if the literatures of some fields are largely composed of nonexistent effects and the only way to find correct answer to any given question is to start again?

I am not sure.

Hunch: People will rely on simple heuristics:

** “Don’t invest time in effects where there are nonreplic’s.”

** “Don’t believe anything in Field X.”

** “Don’t believe professors.”

Ideal of Science

Objective judgment based on publically verifiable facts.

Reality

Subjective judgment heavily reliant on private observations & hunches.

Thank you.