Adjusting for Publication Bias in Meta-Analysis

Perspectives on Psychological Science 2016, Vol. 11(5) 730–749

Blakeley B. McShane
Ulf Bockenholt
Karsten T. Hansen

A potential problem for meta-analysis is that the set of studies examined may not be representative

SOURCES & CONSEQUENCES OF PUBLICATION BIAS

- *Source 1*: Selective reporting based on the size, direction, and statistical significance of study results.
- *Source 2*: Availability and accessibility of studies, for example due to cost, language, familiarity, etc.

- *Consequence*: Distorted meta-analytic estimates:
  - Estimates of population average effect sizes typically biased upwards.
  - Estimates of heterogeneity (between-study variation) typically biased downwards.

- *Note*: "Publication" is this literature is a technical term with a meaning quite different than its ordinary meaning. For all intents and purposes, it means "included in the meta-analysis."

Selection methods are a prominent class of techniques that attempt to adjust for publication bias in meta-analysis

OVERVIEW

- ***Explicitly-specified statistical model*** with two components:
  - *Data model*: Describes how the data are generated in the absence of any publication bias. Generally chosen to be equivalent to the data models typically employed in behavioral research.
  - *Selection model*: Describes the "publication" process, for example:
    - Only studies with results that are statistically significant are published.
    - Only studies with results that are statistically significant and directionally consistent are published.
    - Studies with results that are not statistically significant (or directionally consistent) are relatively less likely to be published than studies with results that are statistically significant (and directionally consistent).
- Estimation of the model is typically via the principled **maximum likelihood estimation** (MLE) procedure.
- *Note*: Selection methods (and alternative adjustment techniques) deal only with the first source of bias (i.e., selective reporting based on the size, direction, and statistical significance of study results). They also do not deal with other so-called questionable research practices.

What's so great about an explicitly-specified statistical model and maximum likelihood estimation?

PRINCIPLED AND ADVANTAGEOUS

- ***Model***:
  - Fully open and transparent about the assumptions made about the data ("data generating process").
  - Easily generalizable to new settings.
  - Allows for theoretical (i.e., mathematical) investigation of model properties such as when it may and may not perform well.
  - Allows for use of MLE which has desirable theoretical properties.
- ***Maximum Likelihood Estimation (MLE)***:
  - Asymptotically minimum variance unbiased estimates (i.e., they are optimal).
  - (Asymptotic) standard errors and confidence intervals.
  - Likelihood values that can be compared across model variants.
  - Mainstay estimation technique in meta-analytic research and statistical research more broadly.

We focus only the very first and most basic selection method

HEDGES (1984) & TWO "NEW" METHODS

- Hedges (1984):
  - *Data model*: Effect sizes are modeled as *homogeneous* across studies. Effect size estimates are modeled as normally distributed with unknown variance.
  - *Selection model*: Only studies with results that are statistically significant are published.

- *p*-curve and *p*-uniform (Simonsohn et al., 2014; van Assen et al., 2015):
  - Same model for the observed data--i.e., same *data model* and *selection model*--as Hedges (1984).
  - Different estimation strategy: Instead of using MLE as in Hedges (1984), these techniques use *ad hoc* improvised strategies that result in inferior estimates.

- *Note*: Selection methods have been an active research area since 1984 and have been extended to incorporate heterogeneous effect sizes, study-level moderators, and other features as well as to allow the likelihood of publication to be a very flexible function of the individual study *p*-value (for an overview, see Hedges & Vevea, 2005; Chapter 13 of Schmidt & Hunter, 2014; and Jin et al., 2015).

The three identical approaches make two key assumptions and perform poorly when the assumptions fail to hold

ASSUMPTIONS AND CONSEQUENCES

- *Assumptions*:
  - Only studies with results that are statistically significant are published.
  - Effect sizes are homogeneous across studies.

- When one or both assumptions fail to hold, all three perform ***poorly***:
  - When Assumption 1 fails, they are inefficient and yield noisy, inaccurate estimates of the population average effect size.
  - When Assumption 2 fails, they yield upwardly-biased and highly inaccurate estimates of the population average effect size.

- ***Both assumptions are nearly always false in practice!***

- *Note*: More sophisticated selection methods dating from as early as ~1988-1992 relax both of these assumptions and perform just fine when they are false.

Nonetheless, SNS continue to defend use of *p*-curve--even when effects are heterogeneous

ISSUE #1: MATHEMATICAL

- Some straightforward mathematical facts:
  - The *p*-curve employs the same model for the observed data as Hedges (1984).
  - The *p*-curve does not use MLE and thus yields estimates inferior to Hedges (1984).
  - When effect sizes are heterogeneous, both of these procedures yield upwardly-biased and highly inaccurate estimates of the population average effect size.

- These facts in tandem with the fact that moderate to large effect size heterogeneity is the norm in psychological research (see Error #3), means the *p*-curve cannot be relied upon to provide valid or definitive adjustments for publication bias.

Methodological *p*-hacking

ISSUE #2: METHODOLOGICAL

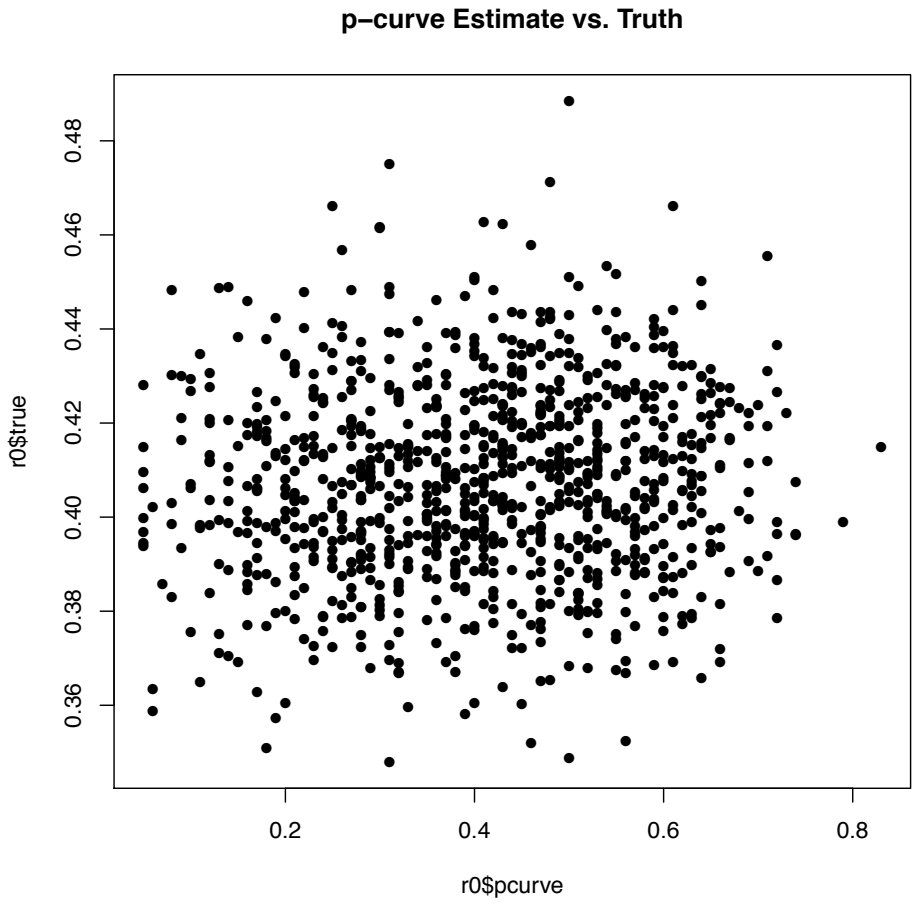
- Meta-analytic research has for decades focused on estimating the ***population average effect size*** and estimating it ***accurately*** as measured by mean square error (MSE) or similar quantities.

- In contrast, SNS s*electively report*:
  - Various novel estimands in lieu of the population average effect size.
  - Various novel model evaluation metrics in lieu of accuracy.

- **Example**: "*P*-curve Handles Heterogeneity Just Fine" (Datacolada #67; Jan 8, 2018):
  - Estimand: average historical power.
  - Evaluation metric: bias.

The *p*-curve is Not "Just Fine"

RESULTS FROM SNS'S OWN SIMULATION STUDY / BLOG POST



Heterogeneity is the norm in psychological research

ISSUE #3: EMPIRICAL

- Heterogeneity is rife and large in comprehensive meta-analyses of psychological studies (Stanley et al., 2017; van Erp et al., 2017).
- More interesting, it persists—and to a reasonable degree—even in large-scale replication projects where rigid, vetted protocols with identical study materials are followed across labs in a deliberate attempt to eliminate it (see Table 3 of Klein et al., 2014; McShane, Böckenholt, & Hansen, 2016; Hagger et al., 2016; Eerland et al., 2016).
- Uri downplays heterogeneity in a recent blog post (Datacolada #63) but misses the point:
  - Regardless of the degree of heterogeneity found in large-scale replication projects, that there is any whatsoever in a setting where every effort is taken to eliminate it is both substantively interesting and strong evidence that it simply cannot be eliminated.
  - To argue for his point that homogeneity is the norm in psychological research, it is not sufficient to argue or provide evidence that there is low or even no heterogeneity where one *would not* expect to find it.
  - Instead, one must show evidence that there is low or no heterogeneity precisely where one *would* expect to find it.
  - The post *selectively reports* one particular effect (anchoring) rather than examining all thirteen effects studied in the data.

So, what can we do?

BE SOPHISTICATED

- The three equivalent techniques (Hedges, 1984, *p*-curve, *p*-uniform) cannot be relied upon to produce valid or definitive adjustments for publication bias.
- Avoid them in favor of the more advanced selection methods that have been proposed over the past decades.
- Fortunately, these methods are easily accessible:
  - https://vevealab.shinyapps.io/WeightFunctionModel/
  - "weightr" package for R.
  - Code in our supplementary materials (although "weightr" is more comprehensive).
- But but but...

It is in vain to hope for a definitive adjustment: the best we can hope for is sensitivity analysis

BE REALISTIC

- More advanced selection methods still make rather idealistic assumptions and the adjusted estimates they yield are highly sensitive to them. Thus, they should be used less for obtaining a single estimate that purports to adjust for publication bias *ex post*.
- Instead, they are useful for **sensitivity analysis**, in particular by:
  - Fitting an "ordinary" random effects meta-analysis that does not adjust for publication bias.
  - Fitting the simple three-parameter selection method (see our paper).
  - Fitting several additional more complicated selection methods that assume different forms of and severity of publication bias.
- If the estimates from all models are relatively consistent, publication bias is unlikely to drive the unadjusted estimate.
- If the estimates vary considerably, publication bias may well drive the unadjusted estimate.

- Note: Vevea and Woods (2005) is an excellent reference on how to use selection methods for sensitivity analysis; Hedges and Vevea (2005) provides additional examples.

If we cannot rely on selection methods for *ex post* adjustment, how about alternative methods?

A FINAL CAUTIONARY NOTE & A FUTURE CHALLENGE

- Many other techniques have been proposed to assess and adjust for publication bias (e.g., funnel plot, nonparametric and regression-based tests, failsafe *N*, trim-and-fill, PET-PEESE).
- These alternative techniques do not posit a data model and a selection model—or any other statistical model.
  - Lack of a statistical model (and the concomitant advantages of one) has been a point of criticism for these alternative techniques (e.g., Becker, 2005).
- In addition, they have yet to be subject to an extensive evaluation:
  - Thus, it is not possible to know in what settings and on what dimensions these techniques perform well versus poorly or how their performance compares to that of alternative techniques such as selection methods.
  - Designing such an evaluation for these techniques is complicated by the fact that they do not posit an underlying statistical model.
- Like selection methods (or any statistical method), these methods will only perform well when their "assumptions" hold and there is sufficient data.

Table 1. Key Points and Recommendations

1. Publication bias distorts meta-analytic estimates of both the population average effect size and the degree of heterogeneity. Estimates of the former are typically biased upward, thus giving the false impression of large effect sizes, whereas estimates of the latter are typically biased downward, thus giving the false impression of homogeneity.
2. Selection methods are a prominent class of techniques that assess and adjust for publication bias in meta-analysis. They were first proposed by Hedges (1984) and have been an active research area ever since.
3. Two recent proposals, the so-called *p*-curve and *p*-uniform approaches (Simonsohn, Nelson, & Simmons, 2014; van Assen, van Aert, & Wicherts, 2015), can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies.
4. The Hedges (1984), *p*-curve, and *p*-uniform approaches are all one-parameter approaches that assume (a) that only studies with results that are statistically significant are published and (b) that effect sizes are homogeneous across studies. When these assumptions hold, the *p*-curve and *p*-uniform approaches perform reasonably well but, as a result of the alternative estimation strategies they employ, not as well as the original Hedges (1984) approach.
5. Falsely assuming that assumptions (a) and (b) hold results in a loss of efficiency (i.e., noisier estimates of the population average effect size) and bias, respectively. Consequently, when one or both assumptions fail to hold, the *p*-curve and *p*-uniform approaches perform poorly, whereas variants of the Hedges (1984) approach perform well.
6. Both assumptions are nearly always false in behavioral research. Consequently, a simple three-parameter variant of the Hedges (1984) approach that relaxes them should be the minimal model considered in applied work. More advanced selection methods may also be considered.
7. Idealistic model assumptions underlie even the most advanced selection methods, and population average effect size estimates can be highly sensitive to these assumptions. Consequently, we advocate that selection methods should be used less for obtaining a single estimate that purports to adjust for publication bias ex post and more for sensitivity analysis—that is, exploring the range of estimates that result from assuming different forms of and severity of publication bias (see Vevea & Woods, 2005, and Hedges & Vevea, 2005).