Selection Models

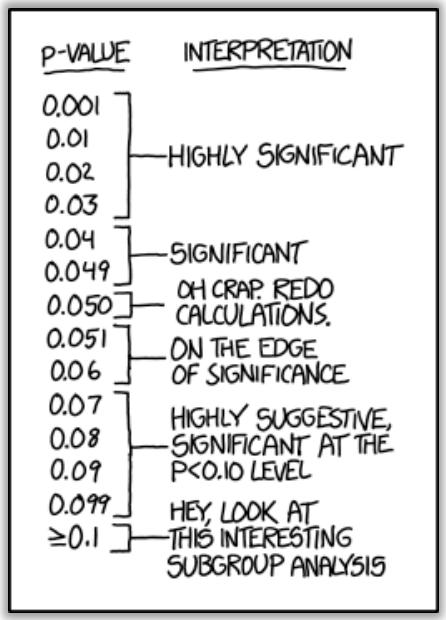
Previously, we covered methods to adjust for small-study effects. All these methods are based on the idea that selective reporting causes a study's **effect size** to **depend on its sample size**.

A more realistic stance may be to say that publication bias **operates through *P*-values**, since, in practice, research findings may only be considered worth publishing when the results are $P < 0.05$:

"Significant P, or no PhD"

Selection models are a method to use when we assume that P-value thresholds may cause our results to be distorted.

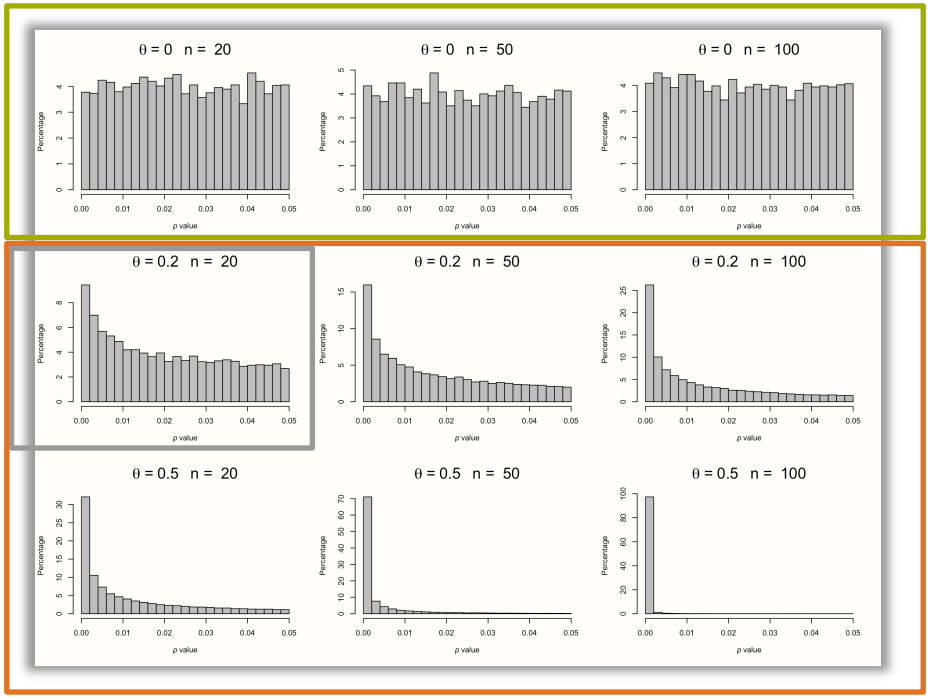
→ This can also include the presence of ***P*-hacking** (which SSE methods do not explicitly account for).



xkcd.com/1478/

P-Curve Simonsohn, Nelson & Simmons (2014)

- **P-Curve** is a type of selection model that only focuses on the **distribution of significant *P*-values** in our data
- Basic idea: if there is **no effect**, *P*-values of studies would follow a **uniform distribution**
- As soon as there is even a minimal effect, we would expect the distribution of effect sizes to become **right-skewed**
- Once we have many studies, this pattern would be perceptible even if the studies themselves are **underpowered**.

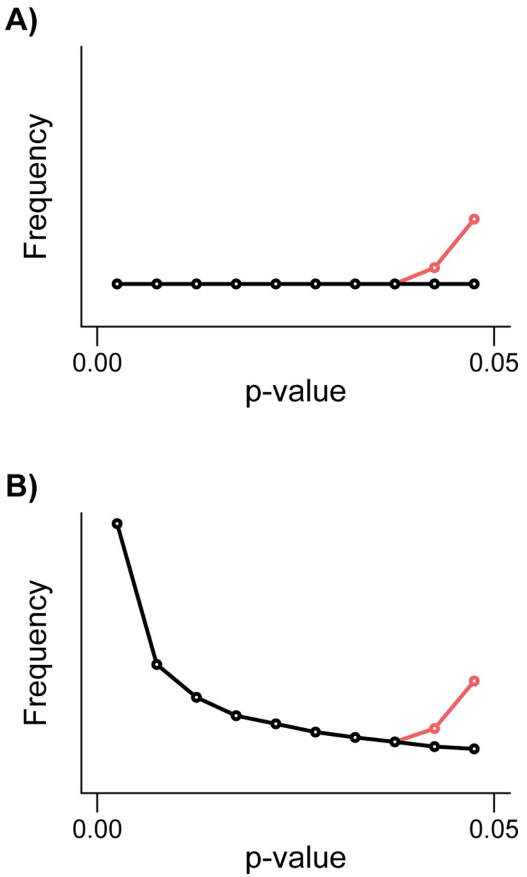


P-Curve Simonsohn, Nelson & Simmons (2014)

If researchers start ***P*-hacking** (i.e., play with their data until results are below 0.05), we expect the P-curve to become **left-skewed.**

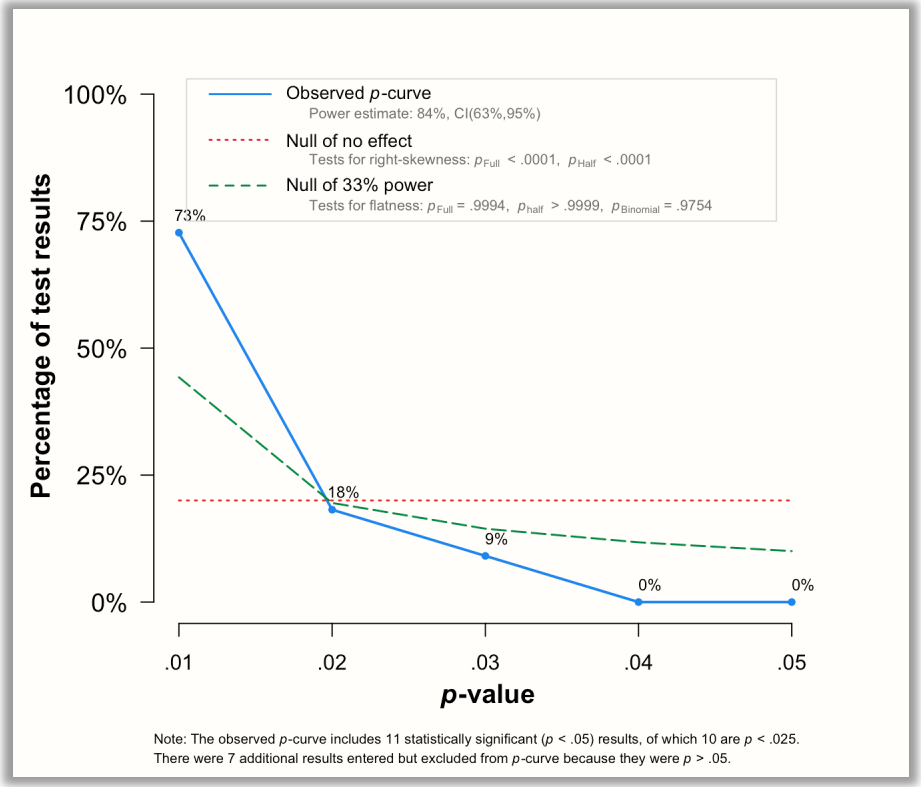
P-curve employs **tests of right-skewness and left-skewness** to determine if there is **evidential value in the data**

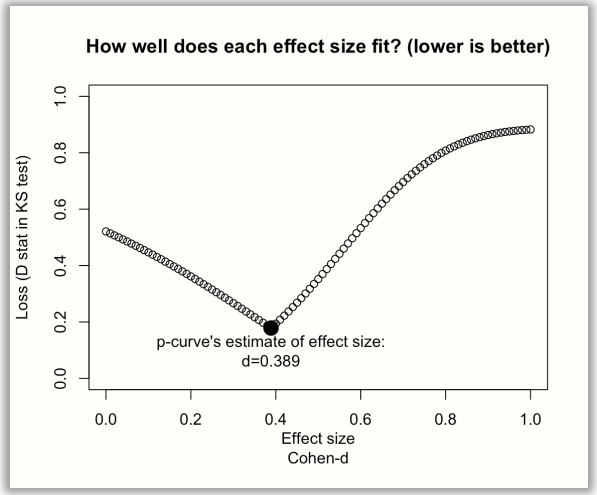
It can also be used to obtain an **adjusted effect estimate**.



Mead et al., 2015

P-Curve Simonsohn, Nelson & Simmons (2014)





		Full Curve		Half Curve		Evidential Value		
	$p_{Binomial}$	z_{Full}	p_{Full}	z_{Half}	p_{Half}	present	absent	$\hat{d}$
Right-Skewness Test	0.020	-3.80		-2.74	0.003	yes	no	0.39
Flatness Test	0.952	1.54	0.938	3.42	>0.999	yes	no	0.39

Which Method Should We Use?

- There is evidence that no publication bias method **consistently outperforms all the others** (Carter et al. 2019).
- There is currently no method providing acceptable results when the **between-study heterogeneity is high** (Aert, Wicherts, and Assen 2016; i.e. $I^2 \approx 75\%$).
- Many methods are based on distinct assumptions that are difficult to test.

	Advantages	Disadvantages
Duval & Tweedie Trim-and-Fill	Very heavily used in practice. Can be interpreted by many researchers.	Often fails to correct the effect size enough, for example when the true effect is zero. Not robust when the heterogeneity is very large; often outperformed by other methods.
PET-PEESE	Based on a simple and intuitive model. Easy to implement and interpret.	Sometimes massively over- or underestimates the effect. Weak performance for meta-analyses with few studies, low sample sizes, and high heterogeneity.
Limit Meta-Analysis	Similar approach as PET-PEESE, but explicitly models between-study heterogeneity.	Performance is less well studied than the one of other methods. May fail when the number of studies is very low (
P-Curve	Has been shown to outperform other methods (particularly trim-and-fill) when its assumptions are met.	Works under the assumption of no heterogeneity, which is unlikely in practice. Requires a minimum number of significant effect sizes. Less easy to interpret and communicate.
Selection Models	Can potentially model any kind of assumed selection process. The three-parameter selection model has shown good performance in simulation studies.	Only valid when the selection model describes the publication bias process adequately. Assume that other small-study effects are not relevant. Can be difficult to interpret and requires background knowledge.

Which Method Should We Use?

- → Conduct several methods as sensitivity analysis.
- → Avoid **“weaponizing” publication bias methods**, given these known limitations.

	Advantages	Disadvantages
Duval & Tweedie Trim-and-Fill	Very heavily used in practice. Can be interpreted by many researchers.	Often fails to correct the effect size enough, for example when the true effect is zero. Not robust when the heterogeneity is very large; often outperformed by other methods.
PET-PEESE	Based on a simple and intuitive model. Easy to implement and interpret.	Sometimes massively over- or underestimates the effect. Weak performance for meta-analyses with few studies, low sample sizes, and high heterogeneity.
Limit Meta-Analysis	Similar approach as PET-PEESE, but explicitly models between-study heterogeneity.	Performance is less well studied than the one of other methods. May fail when the number of studies is very low (
P-Curve	Has been shown to outperform other methods (particularly trim-and-fill) when its assumptions are met.	Works under the assumption of no heterogeneity, which is unlikely in practice. Requires a minimum number of significant effect sizes. Less easy to interpret and communicate.
Selection Models	Can potentially model any kind of assumed selection process. The three-parameter selection model has shown good performance in simulation studies.	Only valid when the selection model describes the publication bias process adequately. Assume that other small-study effects are not relevant. Can be difficult to interpret and requires background knowledge.