Alexander Strobel
Faculty of Psychology

Workshop Open Science Practices

Open Science and Preregistration

Meta-Analytic Tools

Introduction

Can psychologists make better predictions than meteorologists?

Accuracy of whether forecasts

Whether forecasts are increasingly accurate, e.g., in predicting tomorrow's temperature (see figure)

That means that the respective hypotheses on how cold/warm it would be the next day could be confirmed

— 1987: in 75-80% of the forecasts
— 2010: in somewhat above 90% of the forecasts

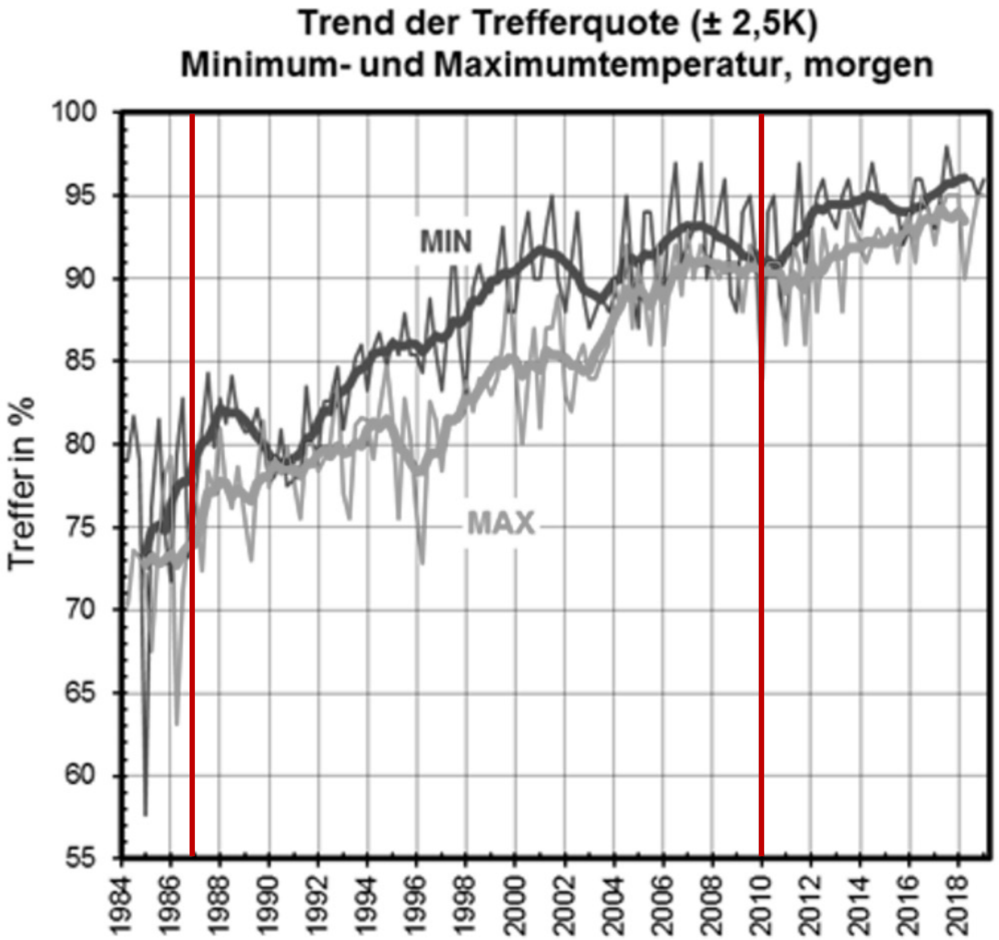


Figure taken from Deutscher Wetterdienst, see article here

Introduction

Can psychologists make better predictions than meteorologists?

Accuracy of whether forecasts

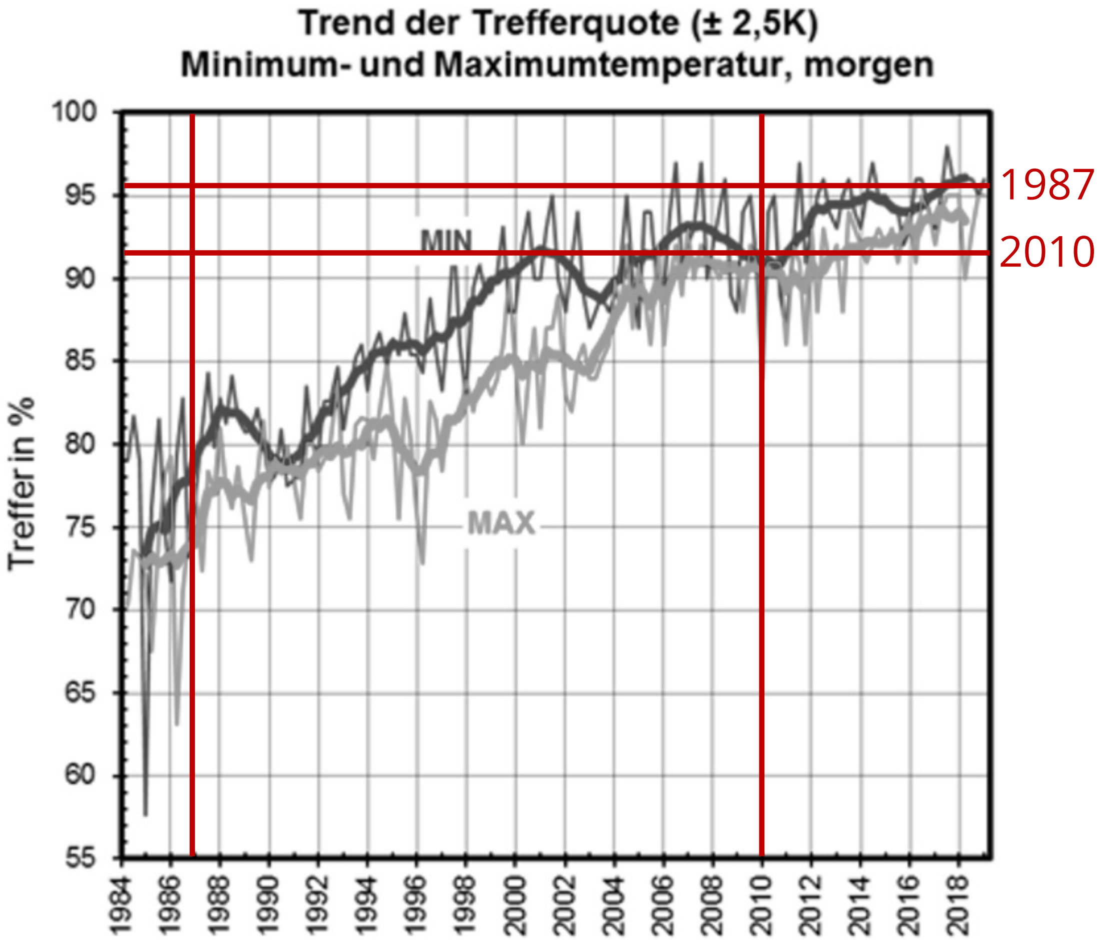
Whether forecasts are increasingly accurate, e.g., in predicting tomorrow's temperature (see figure)

That means that the respective hypotheses on how cold/warm it would be the next day could be confirmed

— 1987: in 75-80% of the forecasts
— 2010: in somewhat above 90% of the forecasts

At the same time, psychologists reported a much higher or in recent years similarly high percentage of confirmed hypotheses (i.e., significant findings):

— 1986-1987: 95.6% (Sterling et al., 1995)
— 2010: 91.5% (Fanelli, 2010)



Sterling et al. (1995). *Am Stat, 49*(1), 108-112. | Fanelli (2010). *PLoS ONE, 5*(4), e10068.

Outline

Sources of bias in the literature

- — Publication bias
- — Questionable research practices

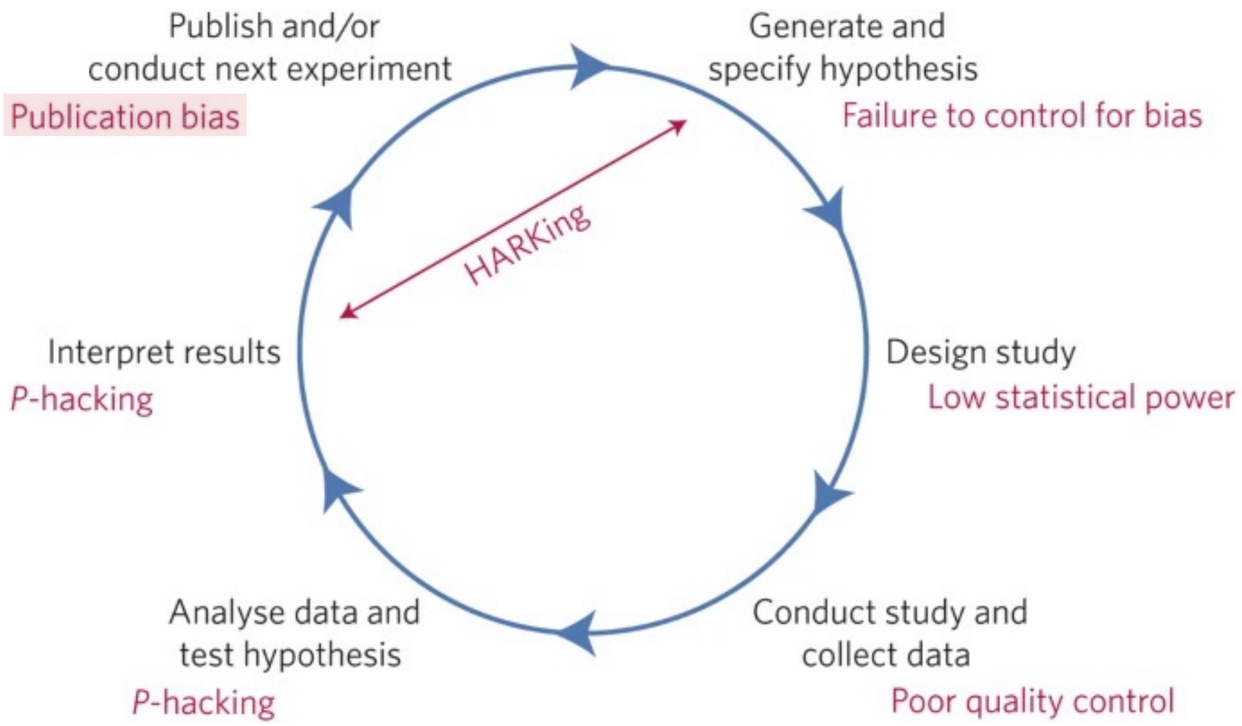
Detecting bias

- — Critical thinking
- — Meta-analytic tools

Avoiding bias

- — Self-reflection and honesty
- — Preregistration
- — Forking path/multiverse analyses

Summary


Publish and/or conduct next experiment
Publication bias
Generate and specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and test hypothesis
P-hacking
Conduct study and collect data
Poor quality control

Sources of bias in the literature

Sources of bias in the literature
Publication bias

Publication bias

manuscripts reporting evidence in favor of their hypotheses have a higher chance of being published due to

— **Reviewer bias** (editors and reviewers selectively reject manuscripts with negative results)

Reviewer bias

“101 consulting editors of the *Journal of Counseling Psychology* and the *Journal of Consulting and Clinical Psychology* were asked to evaluate 3 versions of a research manuscript, differing only with regard to level of statistical significance. The statistically nonsignificant and approach-significance versions were more than 3 times as likely to be recommended for rejection than the statistically significant version.“

Atkinson et al. (1986). *J Counsel Psychol, 29*(2), 189-194.

Scheel et al. (2021). *Adv Meth Pract Psychol Sci, 4*(2), 1-12.

Sources of bias in the literature
Publication bias

Publication bias

manuscripts reporting evidence in favor of their hypotheses have a higher chance of being published due to

- **Reviewer bias** (editors and reviewers selectively reject manuscripts with negative results)
- **File drawer bias** (researchers do not submit studies with negative results for publication, i.e., put it in the *file drawer*)

File drawer bias

“Franco et al. use a Time-sharing Experiments in the Social Sciences archive of nearly 250 peer-reviewed proposals of social science experiments conducted on nationally representative samples. They find that only 10 out of 48 null results were published, whereas 56 out of 91 studies with strongly significant results made it into a journal.“

Franco et al. (2014). *Science, 345*(6203), 1502-1505.

Scheel et al. (2021). *Adv Meth Pract Psychol Sci, 4*(2), 1-12.

Sources of bias in the literature
Questionable research practices

Questionable research practices

research behaviors that make evidence in favor of a certain hypothesis look stronger than it is, e.g.,

— **HARKing** (hypothesizing after results are known; presenting unexpected results as having been predicted a priori)

— ***p*-hacking** (exploiting flexibility in data analysis to obtain statistically significant results)

These QRPs are not necessarily implemented intentionally, but may also simply arise from

— a strong prior belief in some effect

— the belief that significant results are more valuable

— wishful thinking and

— good command, but incoherent use of statistical computing

A hypothetical example

I generated a population of N = 200.000 with five random normally distributed variables

I randomly(!) assigned five meaningful variable names to these data:

— Intelligence

— Motivation

— Socioeconomic status

— Gender (dichotomized)

— School grades

The first four were considered predictors for the last variable.

For R code for this simulation see the *Exercises* folder!

Sources of bias in the literature
Questionable research practices

Questionable research practices

research behaviors that make evidence in favor of a certain hypothesis look stronger than it is, e.g.,

— **HARKing** (hypothesizing after results are known; presenting unexpected results as having been predicted a priori)

— ***p*-hacking** (exploiting flexibility in data analysis to obtain statistically significant results)

These QRPs are not necessarily implemented intentionally, but may also simply arise from

— a strong prior belief in some effect

— the belief that significant results are more valuable

— wishful thinking and

— good command, but incoherent use of statistical computing

A hypothetical example

I then drew a small sample of $N = 50$ from this population. It turned out that "Motivation" was somehow, but not significantly ($p = .165$) associated with "School Grades"

Yet, when I removed three multivariate outliers from the sample, the effect was significant ($p = .016$)

I drew another sample of $N = 50$. No "Motivation" effect here, but without multivariate outliers and with "Intelligence" and "Gender" as moderators and "Socioeconomic status" as covariate, it at least "approached significance" ($p = .076$)

For R code for this simulation see the *Exercises* folder!

Detecting bias

Detecting bias
Critical thinking

What effect size one would expect?

— A small to medium true effect is more likely than a large one

— Psychological phenomena are usually determined by a multitude of factors, anyone of which will most likely not explain much of a phenomenon's variance

What are benchmarks in a given discipline?

— In personnel psychology, among the best predictors for job performance are work samples (r = .54) or structured interviews (r = .51), while self-reported Conscientiousness reaches r = .30 (Schmidt & Hunter, 1998), which is quite large given what one could achieve at all

How large is the sample?

— Does it include only a handful of individuals or several dozens or hundreds?

How plausible is an effect in the first place?

— Does it run counter your expectations or can it be predicted by a sound theory?

Do the authors make the impression that they want to sell their results as somehow "sexy"?

Schmidt & Hunter (1998). *Psychol Bull, 124*(2), 262-274.

Detecting bias
Critical thinking

An example: Bem (2011)

Nine experiments on precognition and premonition were performed. All yielded significant results, with an average effect size of d = 0.22. Mean sample size was N = 122 (range 100 to 200)

- **What effect size should one expect?** Given the nature of the study, an at best small effect would be assumed, so a d = 0.22 is about the size or even smaller than one could expect
- **What are benchmarks in a given discipline?** A d = 0.22 translates into r = .10, so this effect size is smaller than the average effect size in social psychology (r = .20) and thus not really unlikely
- **How large is the sample?** With an average of N = 122, the sample sizes do not seem to be particularly small, but to have a power to detect an effect of d = 0.22, you would need a total sample size of N = 199 to have a power of 80% (calculated using G*Power)
- **How plausible is an effect in the first place?** So, this depends on whether you think that extrasensory perception, precognition and premonition are somehow real phenomena
- **Were the results sold as somehow "sexy"?** Well, having some fancy phrase like "Feeling the Future" in the title, one should already get suspicious … and if the paper is on *psi*, the more so

Bem (2011). *J Pers Soc Psychol, 100*(3), 407-425.

Detecting bias
Critical thinking

An example: Bem (2011)

Overall evaluation

— The effect size reported was rather small, or small to medium at best, so nothing suspicious here

— Benchmark in social psychology is r = .20 (Lovakov & Agadullina, 2021), so with r = .10, it does not look like an inflated effect size

— Sample size is larger than average (N ~ 100, Fraley & Vazire, 2014), but too small to have adequate power to detect the effect in question (N ~200)

— The effect examined is extremely unplausible for a natural sciences researcher's mind

— The whole paper, although well written and often self-reflective, seems somehow odd …

Overall recommendation

— I would not rest my own studies on this paper (and in the meantime, it has indeed faced non-replication, see Wagenmakers et al., 2011)

Wagenmakers et al. (2011). *J Pers Soc Psychol, 100*(3), 426-432.

Detecting bias
Critical thinking

Another example: Westbrook et al. (2019)

One experiment on cognitive effort discounting and its relation to the habitual tendency to invest mental effort (i.e., the trait Need for Cognition) was performed. It yielded a significant result with an effect size of *r* = 0.32. Sample size was *N* = 50

- **What effect size should one expect?** An *r* =.32 would be considered a large one given what can be found in the literature on individual differences research
- **What are benchmarks in a given discipline?** Typical effect sizes in individual differences research lie between *r* = .20-.30, so an *r* = .32 may raise the concern that it is not very likely
- **How large is the sample?** With *N* = 50, the sample had only about 65% power to detect an effect of *r* = .32, it should have been at least *N* = 71 (calculated using G*Power)
- **How plausible is an effect in the first place?** It seems quite likely that individuals who rate themselves as prone to invest mental effort will show less cognitive effort discounting
- **Were the results sold as somehow "sexy"?** The title of the paper is rather cumbersome, and while the writing appears to be quite self-confident, that is nothing to worry about

Detecting bias
Critical thinking

Another example: Westbrook et al. (2013)

Overall evaluation

- The effect size reported is rather large
- Benchmark in individual differences research is *r* = .20-.30 (Gignac & Szodorai, 2016), so with *r* = .32, it could be that the effect size is inflated
- Sample size is smaller than average and does not ensure adequate power to detect the effect in question
- The effect examined is rather plausible
- The whole paper seems very convincing

Overall recommendation

- I would be a bit cautious in following up on this paper, trying to replicate it in a larger sample (and in the meanwhile, it has indeed been replicated, but using a *considerably* larger sample, i.e., *N* = 294, see Kramer et al., 2021)

Kramer et al. (2021). *Cogn Devel, 57*, 100978.

Detecting bias
Meta-analytic tools

Requirements for meta-analysis

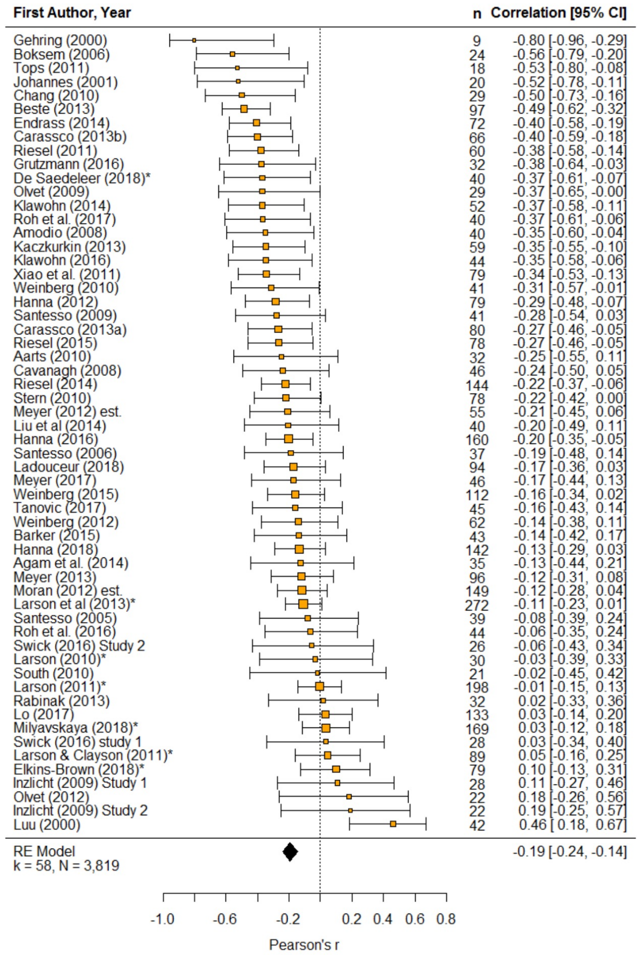
≥ 2 sufficiently similar studies that report effect sizes and sample sizes (the more, the better)

Example

What follows is based on data and code provided with a study by Saunders and Inzlicht (2020) on the correlation between the error-related negativity and anxiety

Tools used

- Funnel plot
- Egger's test for funnel plot asymmetry
- Trim and Fill method
- Peter's test
- PET & PEESE



First Author, Year	n	Correlation [95% CI]
Gehring (2000)	9	-0.80 [-0.96, -0.29]
Boksem (2006)	24	-0.56 [-0.79, -0.20]
Tops (2011)	18	-0.53 [-0.80, -0.08]
Johannes (2001)	20	-0.52 [-0.78, -0.11]
Chang (2010)	29	-0.50 [-0.73, -0.16]
Beste (2013)	97	-0.49 [-0.62, -0.32]
Endrass (2014)	72	-0.40 [-0.58, -0.19]
Carassco (2013b)	66	-0.40 [-0.59, -0.18]
Riesel (2011)	60	-0.38 [-0.58, -0.14]
Grutzmann (2016)	32	-0.38 [-0.64, -0.03]
De Saedeleer (2018)*	40	-0.37 [-0.61, -0.07]
Olvet (2009)	29	-0.37 [-0.65, -0.00]
Klawohn (2014)	52	-0.37 [-0.58, -0.11]
Roh et al. (2017)	40	-0.37 [-0.61, -0.06]
Amodio (2008)	40	-0.35 [-0.60, -0.04]
Kaczkurkin (2013)	59	-0.35 [-0.55, -0.10]
Klawohn (2016)	44	-0.35 [-0.58, -0.06]
Xiao et al. (2011)	79	-0.34 [-0.53, -0.13]
Weinberg (2010)	41	-0.31 [-0.57, -0.01]
Hanna (2012)	79	-0.29 [-0.48, -0.07]
Santesso (2009)	41	-0.28 [-0.54, 0.03]
Carassco (2013a)	80	-0.27 [-0.46, -0.05]
Riesel (2015)	78	-0.27 [-0.46, -0.05]
Aarts (2010)	32	-0.25 [-0.55, 0.11]
Cavanagh (2008)	46	-0.24 [-0.50, 0.05]
Riesel (2014)	144	-0.22 [-0.37, -0.06]
Stern (2010)	78	-0.22 [-0.42, 0.00]
Meyer (2012) est.	55	-0.21 [-0.45, 0.06]
Liu et al (2014)	40	-0.20 [-0.49, 0.11]
Hanna (2016)	160	-0.20 [-0.35, -0.05]
Santesso (2006)	37	-0.19 [-0.48, 0.14]
Ladouceur (2018)	94	-0.17 [-0.36, 0.03]
Meyer (2017)	46	-0.17 [-0.44, 0.13]
Weinberg (2015)	112	-0.16 [-0.34, 0.02]
Tanovic (2017)	45	-0.16 [-0.43, 0.14]
Weinberg (2012)	62	-0.14 [-0.38, 0.11]
Barker (2015)	43	-0.14 [-0.42, 0.17]
Hanna (2018)	142	-0.13 [-0.29, 0.03]
Agam et al. (2014)	35	-0.13 [-0.44, 0.21]
Meyer (2013)	96	-0.12 [-0.31, 0.08]
Moran (2012) est.	149	-0.12 [-0.28, 0.04]
Larson et al (2013)*	272	-0.11 [-0.23, 0.01]
Santesso (2005)	39	-0.08 [-0.39, 0.24]
Roh et al. (2016)	44	-0.06 [-0.35, 0.24]
Swick (2016) Study 2	26	-0.06 [-0.43, 0.34]
Larson (2010)*	30	-0.03 [-0.39, 0.33]
South (2010)	21	-0.02 [-0.45, 0.42]
Larson (2011)*	198	-0.01 [-0.15, 0.13]
Rabinak (2013)	32	0.02 [-0.33, 0.36]
Lo (2017)	133	0.03 [-0.14, 0.20]
Milyavskaya (2018)*	169	0.03 [-0.12, 0.18]
Swick (2016) study 1	28	0.03 [-0.34, 0.40]
Larson & Clayson (2011)*	89	0.05 [-0.16, 0.25]
Elkins-Brown (2018)*	79	0.10 [-0.13, 0.31]
Inzlicht (2009) Study 1	28	0.11 [-0.27, 0.46]
Olvet (2012)	22	0.18 [-0.26, 0.56]
Inzlicht (2009) Study 2	22	0.19 [-0.25, 0.57]
Luu (2000)	42	0.46 [0.18, 0.67]
RE Model k = 58, N = 3,819		-0.19 [-0.24, -0.14]

Saunders & Inzlicht (2020). *Int J Psychophysiol, 155*, 87-98. | https://osf.io/8m6a2/

Detecting bias

Meta-analytic tools

Funnel plot

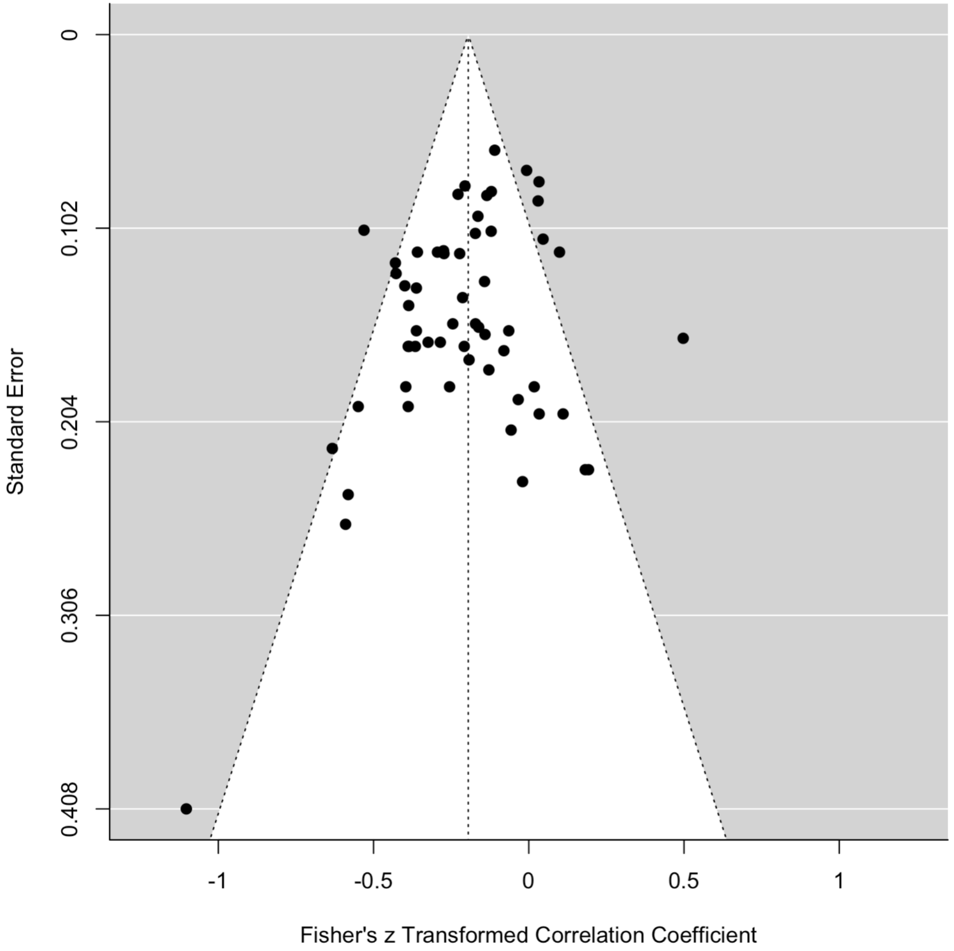
Plots standard errors (SE) against effect sizes of studies included in meta-analyses, making potential asymmetries due to small sample effects visible

Here, we see that the study with the largest SE (i.e., the smallest study, *N* = 9) reported the highest negative correlation (expressed as z-scores), while the highest positive correlation was found for an intermediate SE

Inspection for asymmetry does not suggest bias …

Egger‘s test for funnel plot asymmetry

Regressing the effect size on its SE, weighted by the effect size’s inverse variance also does not suggest bias, $Z = -1.72$, $p = .085$



Egger et al. (1997). *BMJ, 315*, 629.

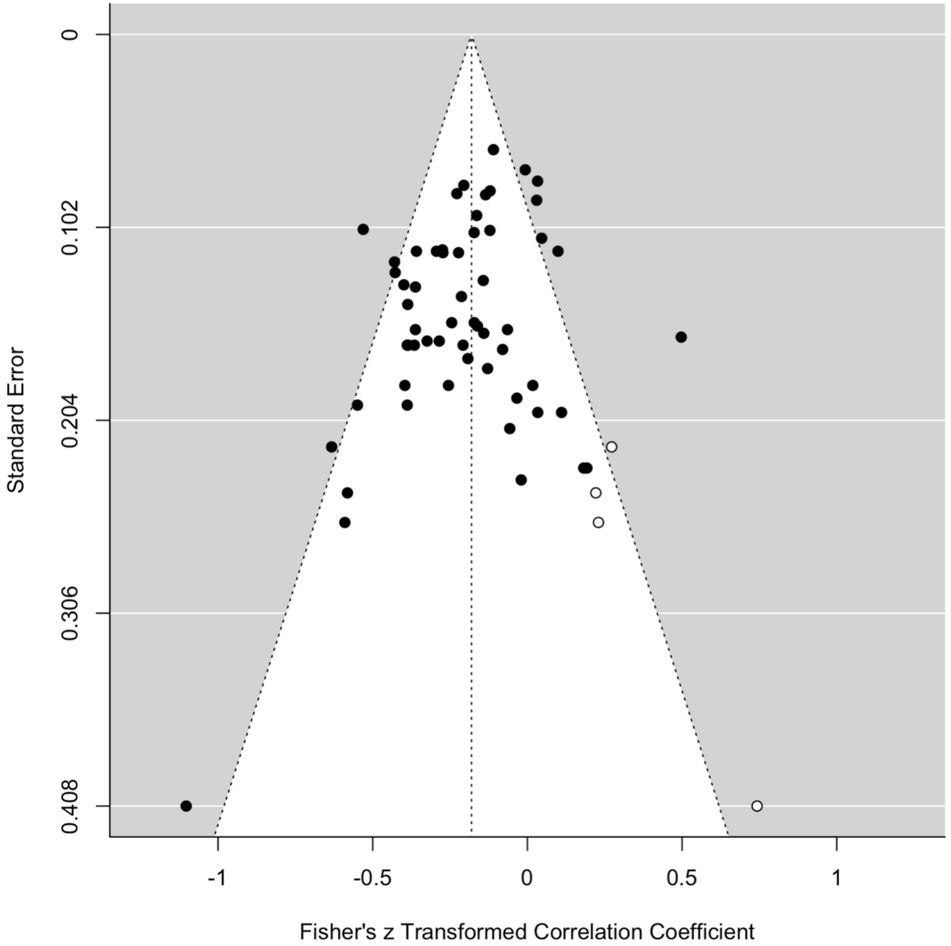
Detecting bias

Meta-analytic tools

Trim and Fill

Tries to account for and imputes missing studies that make a funnel plot (more) symmetrical (white dots)

Lets you estimate the adjusted effect size (in this case, slightly smaller than the original analysis, *r* = -.19 vs. -.18)



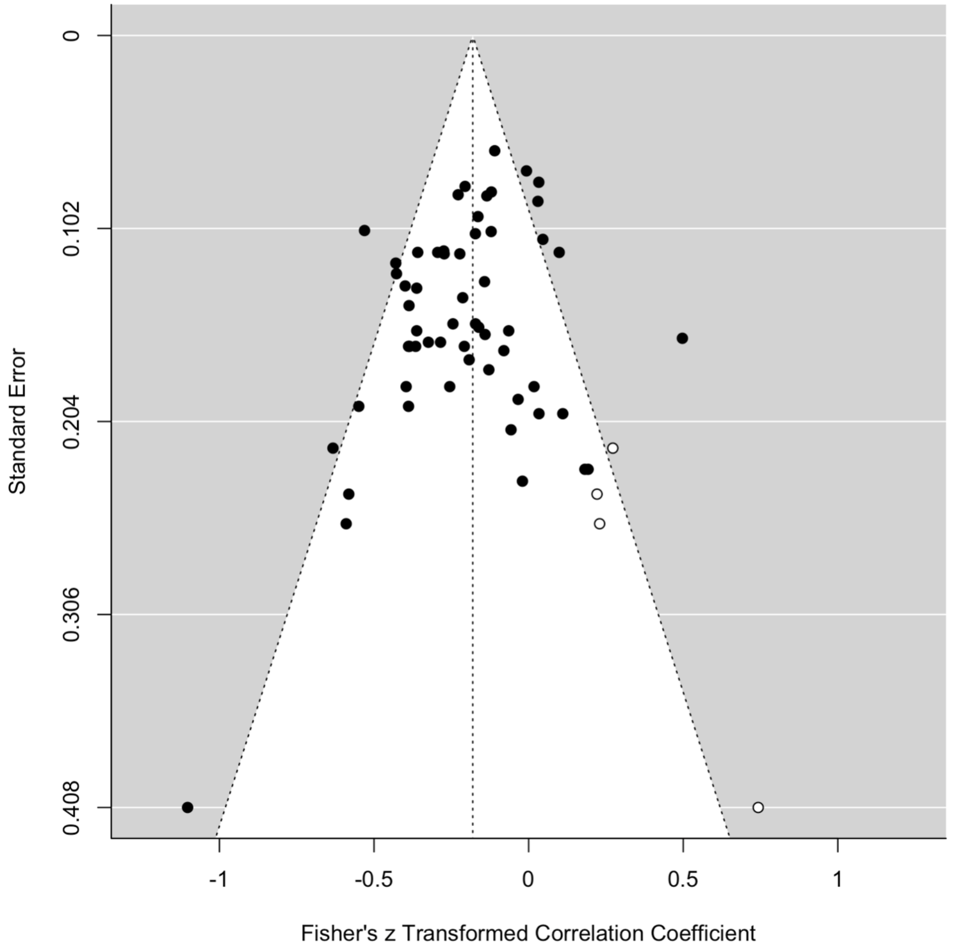
Duvall & Tweedle (2000). *Biometrics, 56*, 455-463.

Detecting bias

Meta-analytic tools

Peter‘s test

Similar to Egger‘s test, but predicts the outcome based on the inverse sample size and uses sample size as weight; In this example gives a lower estimate of the effect size, i.e., *r* = -.11



Peters et al. (2006). *JAMA, 295*(6), 676-680.

Detecting bias
Meta-analytic tools

Peter‘s test

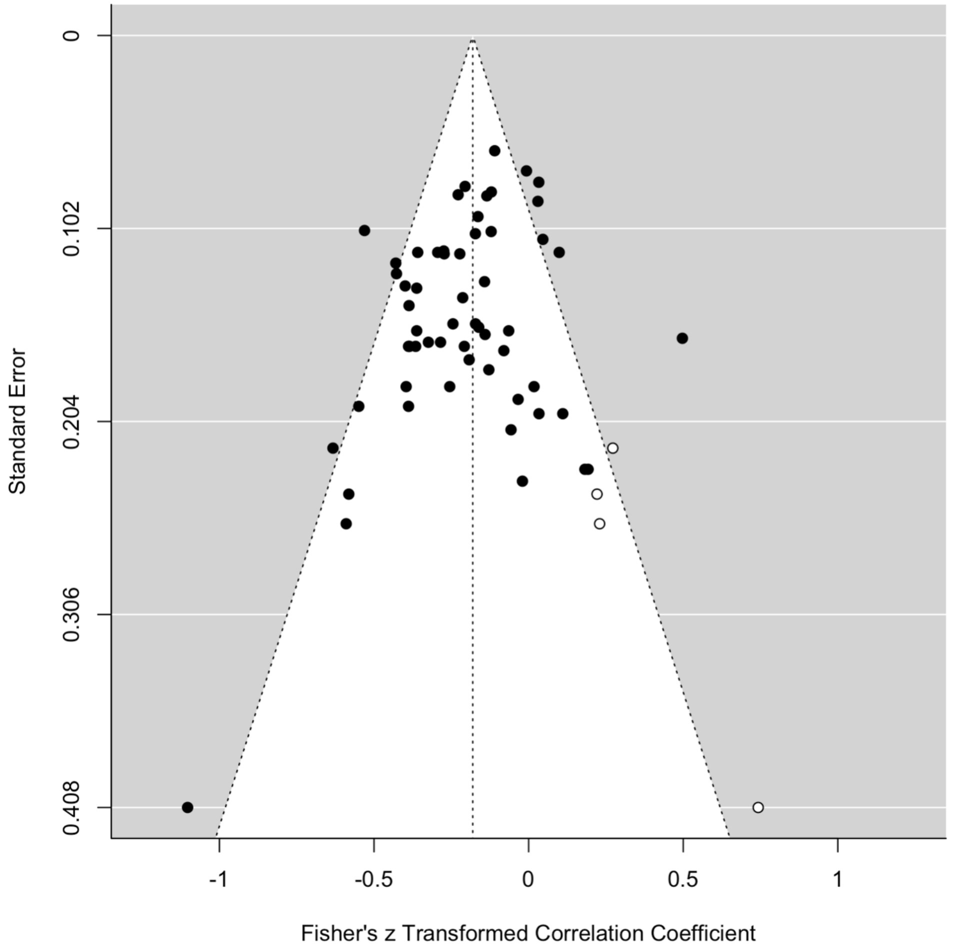
Similar to Egger‘s test, but predicts the outcome based on the inverse sample size and uses sample size as weight; In this example gives a lower estimate of the effect size, i.e., *r* = -.11

PET (precision effect test)

predicts the outcome based on the square root of the sampling variance and uses its inverse as weight, comes up with *r* = -.04 (PET is known to be very conservative)

PEESE (precision effect estimate w/ standard errors)

Predicts the outcome quite similarly to PET, but uses the sampling variance as is as predictor, comes up with *r* = -.12



Stanley (2017). *Soc Psychol Pers Sci, 8*(5), 581-591.

Detecting bias
Meta-analytic tools

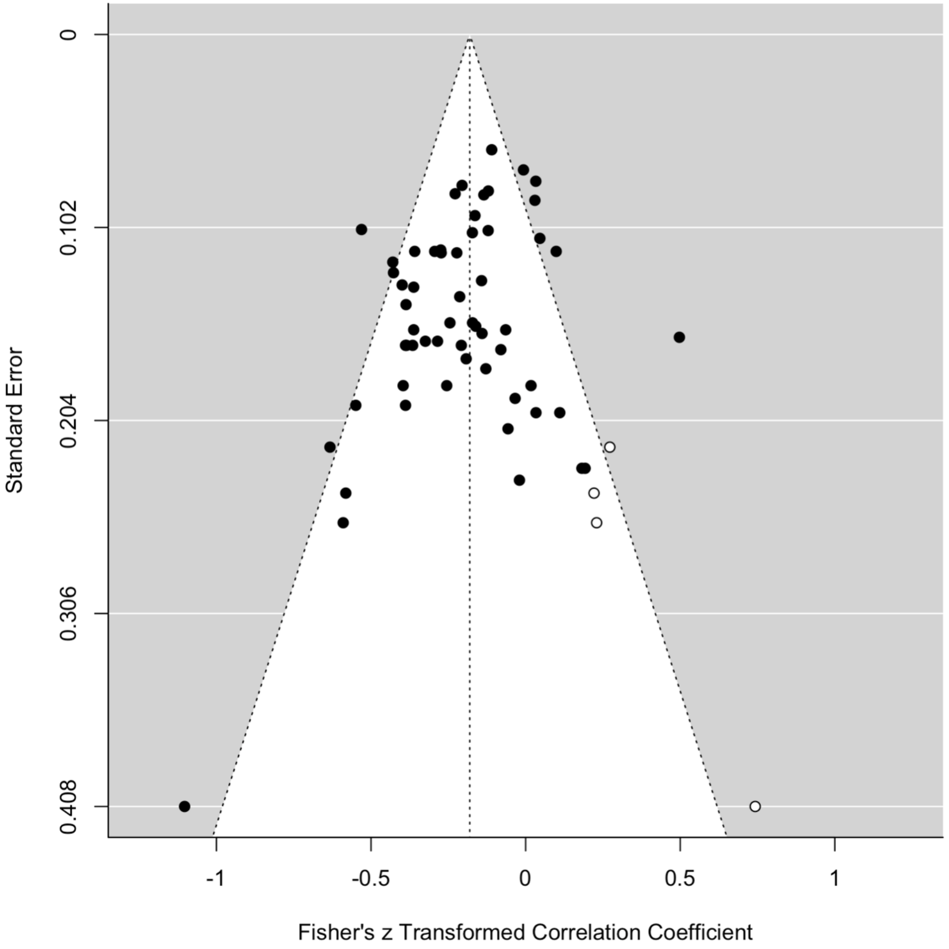
Overall evaluation

In the majority of bias tests performed, the “true” effect size after accounting for bias was only about half as large than the one determined by classical meta-analysis

This fits with results from the Open Science Collaboration that even replicated effect sizes are only half of those initially reported

Yet, simply using half of a reported effect for power analysis size might be overly conservative, especially in fields with growing sample sizes

Using the lower bound of the confidence interval of a reported correlation or a meta-analytically derived one may already give you a good estimate (as long as you do not suspect that QRPs might have been employed)



Avoiding bias

Avoiding bias

Possible measures

Self-reflection and honesty

Reflect on why you chose some analysis decision and be open about it in your research report (or the supplement)

Discussion with colleagues

Would colleagues analyze/interpret your data similarly?

Preregistration

In order to not have to justify your analysis decisions in hindsight, preregister your analysis plan and adhere to it

Ideally, publish your studies as Registered Reports, a growing number of journals offers this option, visit the Center for Open Science for a list: https://www.cos.io/initiatives/registered-reports

Registered Reports to prevent bias

“A new publication format has been developed to prevent selective reporting: In Registered Reports (RRs), peer review and the decision to publish take place before results are known. We compared the results in published RRs (*N* = 71 as of November 2018) with a random sample of hypothesis-testing studies from the standard literature (*N* = 152) in psychology. Analyzing the first hypothesis of each article, we found 96% positive results in standard reports but only 44% positive results in RRs.” (Scheel et al., 2021, p. 1)

Scheel et al. (2021). *Adv Meth Pract Psychol Sci, 4*(2), 1-12.

Avoiding bias
Possible measures

Self-reflection and honesty

Reflect on why you chose some analysis decision and be open about it in your research report (or the supplement)

Discussion with colleagues

Would colleagues analyze/interpret your data similarly?

Preregistration

In order to not have to justify your analysis decisions in hindsight, preregister your analysis plan and adhere to it

Forking path/multiverse/specification curve analyses

Analyze your data along all the forking paths that arise from all reasonable analysis decisions and check how much your results depend on individual decisions

Example forking path analysis

You have a dataset with five variables that you simply want to correlate with each other

Depending on

— whether or how you detect and treat multivariate outliers

— which correlation method you choose

— whether and how you correct for multiple testing

you may end up with up to 32 possible analysis paths already for such a simple analysis!

Wacker (2017). *Front Psychol, 8*, 1332. | Steegen et al. (2016). *Persp Psychol Sci, 11*(5), 702-712. | Simonsohn et al. (2019). *SSRN.*

Avoiding bias
Possible measures

Self-reflection and honesty

Reflect on why you chose some analysis decision and be open about it in your research report (or the supplement)

Discussion with colleagues

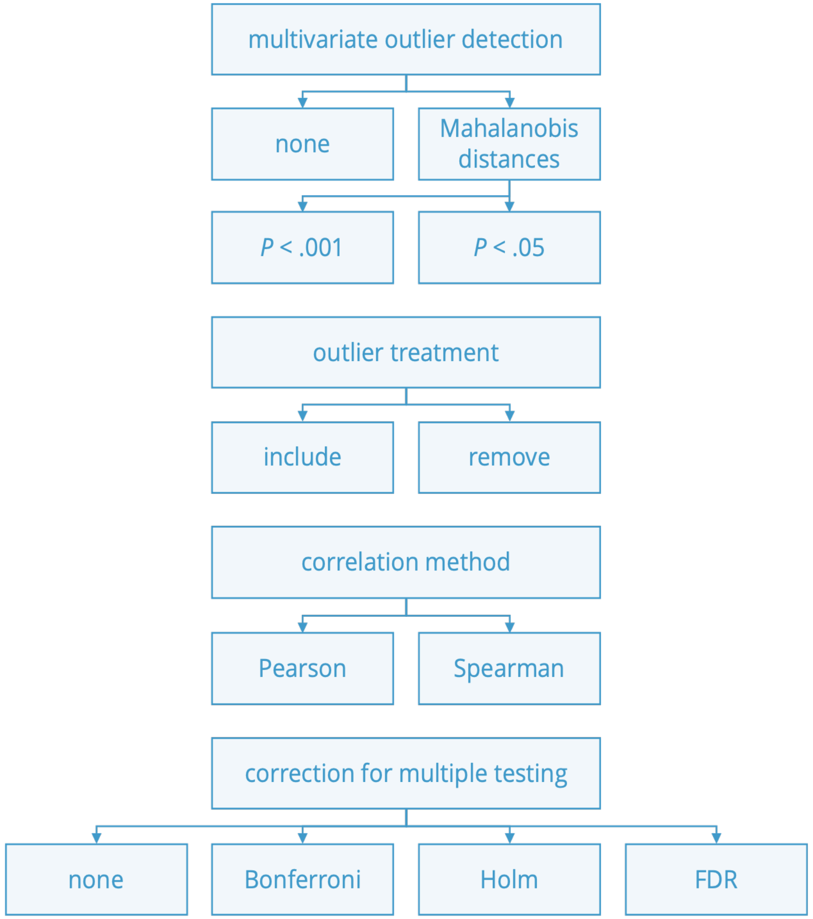
Would colleagues analyze/interpret your data similarly?

Preregistration

In order to not have to justify your analysis decisions in hindsight, preregister your analysis plan and adhere to it

Forking path/multiverse/specification curve analyses

Analyze your data along all the forking paths that arise from all reasonable analysis decisions and check how much your results depend on individual decisions



Wacker (2017). *Front Psychol, 8*, 1332. | Steegen et al. (2016). *Persp Psychol Sci, 11*(5), 702-712. | Simonsohn et al. (2019). *SSRN.*

Summary

Summary
Detecting and avoiding publication bias

Publication bias

is an ubiquitous phenomenon in the psychological literature that among others is due to reviewer and file drawer bias that foster the use of QRPs which in turn exacerbate bias

Detect bias

Critically evaluate studies of interest: is the reported effect size a reasonable one in a given field, is the sample size/power adequate, is the effect plausible in the first place and is it not sold as “sexy”?

If there are several studies in a field of your interest, use meta-analytic tools to correct for bias, e.g., via Trim-and-Fill, Peter’s test, PET-PEESE

Avoid bias

Self-reflect on your analysis, be honest about your analysis decisions, discuss them with colleagues, preregister your analysis plan and/or run a forking path analysis

Thank you!

Exercises

1) Generate random data and randomly name the variables to match your field's variables!

By "believing" these were real data, try to *p*-hack your way down to come up with significant results and then engage in some HARKing and screen the literature to justify expecting what you found (and note, how successful you will be and how bad this feels …)

2) Evaluate a paper you recently read and try to predict its replicability!

Try to answer the questions outlined on slide 11 and give an overall assessment of how much you would rely on the paper's results!

3) If you have several effect sizes of interest, try to apply some meta-analytic bias-estimates!

You might want to work through the paper by Saunders and Inzlicht (2020) in the *Resources* folder using the R code provided there as well.

4) Read about Registered Reports

e.g., on https://www.cos.io/initiatives/registered-reports

Saunders & Inzlicht (2020). *Int J Psychophysiol, 155,* 87-98. | https://osf.io/8m6a2/