

# Systematic Review of Heterogeneity in Psychological Science

Scientists are like sailors who must repair a rotting ship at sea. They trust the great bulk of timbers while they replace a particularly weak plank. Each of the timbers that they now trust they may, in its turn, replace. The proportion of the planks they are replacing to those they treat as sound must always be small. - W. V. Quine

## Introduction

An often overlooked problem in meta-analytic studies within psychology is the handling of heterogeneity (Linden and Hönekopp, 2021). Heterogeneity occurs when the pooled effects contain more variation than expected from random sampling. This is often expressed with the  $\tau^2$  statistic, which represent the between-study variation observed in the sample. The potential cause of the observed heterogeneity can be attributed to the fact that multiple “true” effects are observed in the data and therefore more variation than expected is present within the sample. The presence of multiple true effects can be caused by the presence of subgroups of effects that remain unadjusted for through meta-regression or subgroup analyses. These subgroups of effects can be caused by a multitude of things such as differing research methodologies, populations and other design based matters.

## Motivation and Intention

It is worth to note that this project is the start of a PhD thesis. The objective of this thesis is to explore Bayesian evidence synthesis. One situation where this technique could be useful is when the heterogeneity between studies is too substantial for a valid comparison between effects is possible. Therefore, a qualitative exploration of this situation is needed in order to establish whether the technique provides value in addition to standard meta-analytic techniques.

The reason why this is apt to point out is because this review will largely be a qualitative assessment of the literature on heterogeneity and cover subjective issues related to scientific theory. Being mindful of our prior attitudes in these areas is therefore prudent. One way

of specifying the goal of this paper is to analyze under what conditions studies are excluded from meta analyses on grounds of heterogeneity, and when does heterogeneity become an issue regarding construct validity and external validity. These two types of validity are important to consider in evidence synthesis since we rely on the internal validity that any k study included in the analysis evaluates the questions or underlying concept it states it does, and external validity in the sense that any k study can be generalized to other conditions examining the same concept.

## Objective

The objective of this systematic review is to get a picture of what steps are taken to adjust for observing a high degree of heterogeneity within the field of psychology as a whole, and whether certain fields have more issues than others with observed between-study heterogeneity. Since the underlying tool; pooling effects and analyzing them, are the same across fields, how heterogeneity affects the results should also be the same. In this review i want to...

- Compare degree of heterogeneity across various fields
- Compare tools/statistics for measuring and accounting for heterogeneity
- Evaluate how well heterogeneity is accounted for/reduced
- Evaluate when researchers deem heterogeneity to be an issue which challenges the validity of their results.
- Define when heterogeneity matters and at level of heterogeneity that requires a different mode of evidence synthesis.
- Discuss the objective of psychological meta-analyses with regards to estimation and or significance testing as well as how that relates to issues with heterogeneity.
- Discuss how questionable research practices could affect heterogeneity

## Methods

### Search strategy

To find databases I propose speaking with the people at the university library about this, the same can be said about search terms. I have used the following search terms for cursory searches in scopus:

```
(TITLE-ABS-KEY(meta analy*) OR TITLE-ABS-KEY(meta-analy*)) AND ( LIMIT-TO (
SUBJAREA,"PSYC" ) ) AND ( LIMIT-TO ( LANGUAGE,"English" ) ) AND ( LIMIT-
TO ( EXACTKEYWORD,"Meta Analysis" ) OR LIMIT-TO ( EXACTKEYWORD,"Meta-
analysis" ) )
```

Resulting in 14,365 documents. And

(TITLE-ABS-KEY(meta AND analy\*) OR TITLE-ABS-KEY(meta-analy\*) AND TITLE-ABS-KEY(heterogeneity)) AND ( LIMIT-TO ( SUBJAREA,“PSYC” ) ) AND ( LIMIT-TO ( LANGUAGE,“English” ) ) AND ( LIMIT-TO ( EXACTKEYWORD,“Meta Analysis” ) OR LIMIT-TO ( EXACTKEYWORD,“Meta-analysis” ) )

Resulting in 1972 results.

### **Exclusion criteria**

- Full text not available
- Is not a meta-analysis
- Does not report heterogeneity measures

### **Sampling procedure**

Since the main goal here is to do a qualitative assessment a complete exhaustion of the literature in terms of including all studies on the topic seems unnecessary, especially given the large pool of meta analyses that seems to be present in the searches. In accordance with the objectives stated above, I would like to have a sample which is rich in variability with regards to procedure and topic so that we gain highly generalizable results. I think this is a good idea because...

- It gives nuance to when heterogeneity is an issue
- Is applicable to most applied researchers in psychology
- Sets the stage well for future projects

One way of getting a sampling studies could be through a stratified/grouped sample covering different fields of psychology. This way we ensure that a wide range of sub-fields are covered. Additionally and or alternatively, we might follow inclusion based on journals (e.g psych bull, clinical psychology review, neuropsychology review, health psychology review, personality and social psychology bulletin etc). This would apply some rules of sampling from quantitative research such as information saturation and purposeful sampling targeting only information rich cases - though this will come with its own set of biases.

## **Pre-registration**

I think pre-registration would be a good idea. That way I/we need to hammer out what our research questions are thoroughly before doing the data extraction. It also stakes out possible forking paths contingent on possibly problematic results. I think a pre-registration with a PRISMA statement along with the general university guidelines should be enough. I believe the University Library can help with both literature search and establishing guidelines conducive to a high quality review.

## **Questions**

- What, concretely, is the research question. Have i missed it in the mission statement/objective section?
- Should we strive for a very wide scope of studies (fields, models, effect sizes etc)?
- Should we include some meta-analytic measures in the review (sample size, Tau etc)?
- What literature should we look at for when between-study heterogeneity becomes too large for valid/reliable inferences.
- Do we have a vision for what the ideal structure of the paper should be? Who is it for beyond ourselves as a launchpad for the BES project?
- Do you think that i have a good understanding of statistical heterogeneity? what is it I am missing?
- I am meeting Erika Manten on Friday, do we have any documents/information we should provide her with.

## **Reading/material to cover**

### **Books**

- Cook and Campbell. Validity
- Jeffreys, H. (1961). Theory of probability, 3rd edn. Oxford: Oxford University Press. p -387
- Jaynes, E. The Logic of Science(not sure if i can understand it)

## Research papers

Baujat, Bertrand, Cédric Mahé, Jean-Pierre Pignon, and Catherine Hill. 2002. “A Graphical Method for Exploring Heterogeneity in Meta-Analyses: Application to a Meta-Analysis of 65 Trials.” *Statistics in Medicine* 21 (18): 2641–52.

Bogomolov, Marina; Heller, Ruth (2022): Replicability Across Multiple Studies: arXiv.

Borenstein, Michael, Julian PT Higgins, Larry V Hedges, and Hannah R Rothstein. 2017. “Basics of Meta-Analysis: I2 2 Is Not an Absolute Measure of Heterogeneity.” *Research Synthesis Methods* 8 (1): 5–18.

Bryan, Christopher J.; Tipton, Elizabeth; Yeager, David S. (2021): Behavioural science is unlikely to change the world without a heterogeneity revolution. In *Nature human behaviour* 5 (8), pp. 980–989. DOI: 10.1038/s41562-021-01143-3.

Crandall, Christian S.; Sherman, Jeffrey W. (2016): On the scientific superiority of conceptual replications for scientific progress. In *Journal of Experimental Social Psychology* 66, pp. 93–99. DOI: 10.1016/j.jesp.2015.10.002.

Gelman, Andrew (2018): The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. In *Personality & social psychology bulletin* 44 (1), pp. 16–23. DOI: 10.1177/0146167217729162.

Gurevitch, Jessica; Koricheva, Julia; Nakagawa, Shinichi; Stewart, Gavin (2018): Meta-analysis and the science of research synthesis. In *Nature* 555 (7695), pp. 175–182. DOI: 10.1038/nature25753.

Hedges, Larry V.; Schauer, Jacob M. (2019): Statistical analyses for studying replication: Meta-analytic perspectives. In *Psychological methods* 24 (5), pp. 557–570. DOI: 10.1037/met0000189.

Higgins, J, S Thompson, J Deeks, and D Altman. 2002. “Statistical Heterogeneity in Systematic Reviews of Clinical Trials: A Critical Appraisal of Guidelines and Practice.” *Journal of Health Services Research Policy* 7 (1): 51–61.

Higgins, Julian PT, and Simon G Thompson. 2002. “Quantifying Heterogeneity in a Meta-Analysis.” *Statistics in Medicine* 21 (11): 1539–58.

Jackson, Dan. 2013. “Confidence Intervals for the Between-Study Variance in Random Effects Meta-Analysis Using Generalised Cochran Heterogeneity Statistics.” *Research Synthesis Methods* 4 (3): 220–29.

Langan, Dean, Julian PT Higgins, Dan Jackson, Jack Bowden, Areti Angeliki Veroniki, Evangelos Kontopantelis, Wolfgang Viechtbauer, and Mark Simmonds. 2019. “A Comparison of Heterogeneity Variance Estimators in Simulated Random-Effects Meta-Analyses.” *Research Synthesis Methods* 10 (1): 83–98.

- Linden, Audrey Helen; Hönekopp, Johannes (2021): Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science. In *Perspectives on psychological science : a journal of the Association for Psychological Science* 16 (2), pp. 358–376. DOI: 10.1177/1745691620964193.
- Lucas, Jeffrey W. (2003): Theory-Testing, Generalization, and the Problem of External Validity. In *Sociological Theory* 21 (3), pp. 236–253. DOI: 10.1111/1467-9558.00187.
- Makambi, Kepher H. 2004. “The Effect of the Heterogeneity Variance Estimator on Some Tests of Treatment Efficacy.” *Journal of Biopharmaceutical Statistics* 14 (2): 439–49.
- Mathur, Maya B.; VanderWeele, Tyler J. (2019): Challenges and suggestions for defining replication “success” when effects may be heterogeneous: Comment on Hedges and Schauer (2019). In *Psychological methods* 24 (5), pp. 571–575. DOI: 10.1037/met0000223.
- Mathur, Maya B.; VanderWeele, Tyler J. (2019): New metrics for meta-analyses of heterogeneous effects. In *Statistics in medicine* 38 (8), pp. 1336–1342. DOI: 10.1002/sim.8057.
- Moeller, Julia: Generalizability Crisis Meets Heterogeneity Revolution. Determining Under Which Boundary Conditions Findings Replicate and Generalize.
- Olkin, Ingram, Issa J Dahabreh, and Thomas A Trikalinos. 2012. “GOSH—a Graphical Display of Study Heterogeneity.” *Research Synthesis Methods* 3 (3): 214–23.
- Protzko, John; Krosnick, Jon; Nelson, Leif D.; Nosek, Brian A.; Axt, Jordan; Berent, Matthew et al. (2020): High Replicability of Newly-Discovered Social-behavioral Findings is Achievable: Center for Open Science.
- Rücker, Gerta, Guido Schwarzer, James R Carpenter, and Martin Schumacher. 2008. “Undue Reliance on I<sup>2</sup> in Assessing Heterogeneity May Mislead.” *BMC Medical Research Methodology* 8 (1): 79.
- Shrier, I., Platt, R. W., & Steele, R. J. (2007). Mega-trials vs. meta-analysis: precision vs. heterogeneity?. *Contemporary clinical trials*, 28(3), 324–328.
- Silberzahn, R.; Uhlmann, E. L.; Martin, D. P.; Anselmi, P.; Aust, F.; Awtrey, E. et al. (2018): Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. In *Advances in Methods and Practices in Psychological Science* 1 (3), pp. 337–356. DOI: 10.1177/2515245917747646.
- Stanley, T. D.; Carter, Evan C.; Doucouliagos, Hristos (2018): What meta-analyses reveal about the replicability of psychological research. In *Psychological bulletin* 144 (12), pp. 1325–1346. DOI: 10.1037/bul0000169.
- Sidik, Kurex, and Jeffrey N Jonkman. 2002. “A Simple Confidence Interval for Meta-Analysis.” *Statistics in Medicine* 21 (21): 3153–59.
- . 2005. “Simple Heterogeneity Variance Estimation for Meta-Analysis.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (2): 367–84.

- . 2007. “A Comparison of Heterogeneity Variance Estimators in Combining Results of Studies.” *Statistics in Medicine* 26 (9): 1964–81.
- . 2019. “A Note on the Empirical Bayes Heterogeneity Variance Estimator in Meta-Analysis.” *Statistics in Medicine* 38 (20): 3804–16.
- Szucs, Denes; Ioannidis, John P. A. (2017): When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. In *Frontiers in human neuroscience* 11, p. 390. DOI: 10.3389/fnhum.2017.00390.
- Viechtbauer, Wolfgang. 2005. “Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model.” *Journal of Educational and Behavioral Statistics* 30 (3): 261–93.
- . 2007b. “Confidence Intervals for the Amount of Heterogeneity in Meta-Analysis.” *Statistics in Medicine* 26 (1): 37–52.
- Terrin, Norma, Christopher H Schmid, Joseph Lau, and Ingram Olkin. 2003. “Adjusting for Publication Bias in the Presence of Heterogeneity.” *Statistics in Medicine* 22 (13): 2113–26.
- Yarkoni, Tal (2020): The generalizability crisis. In *The Behavioral and brain sciences* 45, e1. DOI: 10.1017/S0140525X20001685.