Project proposal: Explanations for unexplained effect size heterogeneity

Anton Olsson Collentine

2021-06-01

## Short summary

* Theoretical (mainly) paper consisting of introduction + empirical data collection + 7 explanations + empirical illustration + conclusion
* Sections are:
  + Introduction (1-2 pages)
    - Content: Comparable studies can have different outcome (heterogeneity), we present 7 explanations, argue that artefactual explanations should be examined before moderators, and summarize the content of the paper.
  + Empirical data collection (1 - 2 pages)
    - RQ: How much do heterogeneity do moderators explain?
    - we will extract Heterogeneity estimates before and after moderators in a sample of meta-analyses
      * Sample: Perhaps van Erp’s dataset, or set up search terms
      * Extract: Q/I2/tau before/after moderators applied
  1. Model and sampling error (1.5 – 2 pages)
     + Present meta-analytic heterogeneity model
     + Explain that it accounts for sampling error
     + Discuss estimation artefact that can increase heterogeneity estimate (I2 vs. tau, central chi-square distribution of Q)
  2. Multiverse effect (2 pages)
     + Discuss based on multi-analyst papers
     + Illustrate using data from meta-multiverse project
     + Highlight that even though multiverse does not affect average estimates in absence of selective reporting, it can affect heterogeneity estimates
  3. Publication bias (max 1 page)
     + Literature review: Hilde’s paper + others
  4. QRPs (minimum 2 pages with illustrations)
     + should usually decreases heterogeneity, assuming people select for significance
     + Literature review on different types of QRPs (talk to Esther)
  5. “Incompetence”/ “poor methodological quality” (1/3 of a page)
     + Common critique when replications ‘fail’
     + This critique is not detailed enough as a critique
     + If an “incompetent” researcher uses a less effective intervention, they should be seen as a design moderator (but must first be better specified)
     + Cochrane recommends talking about ‘risk of bias’ rather than ‘poor methodological quality’
  6. Measurement artefacts (1 – 2 pages)
     + Differential measurement error leads to heterogeneity (Flake et al., examined availability of reliability estimates in primary studies)
     + Range restrictions? Other measurement factors? (see Schmidt & Hunter, 2015)
  7. Moderators (2 pages)
     + Short discussion, extensively described in other papers
     + Types of moderators (contextual factors): settings, sample population, treatment and measurement variables
     + Meta-regression, power, IPD
  + Empirical illustration (2-3 pages)
    - Use a large meta-analysis to illustrate/examine all 7 points
    - Good choice is Chris’ cyberball meta-analysis (see final sentence)
    - Preferably a smaller, homogenous subset from it
  + Conclusion (1/2 page)
* Total: ~18 pages

## Detailed summary

Below I present a detailed summary of

1. the introduction: for a better description of the proposed angle of the paper.

### Introduction

**Comparable studies always have some design heterogeneity, whether they differ in the studies’ settings, their sample populations, the measures they use, or their treatment variables (e.g., Campbell & Stanley, 2015).** In turn, effect size estimates from comparable studies always have some statistical heterogeneity, one potential explanation of which is the studies’ design heterogeneity (i.e., the effect size estimates arising from different true effects). Other explanations for differences in statistical estimates from comparable studies include sampling error, multiverse variability, publication bias, Questionable Research Practices such as selective reporting (including p-hacking), ‘poor methodological quality’ or supposed incompetence of experimenters, and measurement artefacts.

**It can be challenging to disentangle the reasons for why effect size estimates differ.** From a theoretical perspective we are most interested in finding out to what extent these differences arise from design heterogeneity, to which end other explanations are nuisances or artefacts we must first control for. Perhaps this explains why heterogeneity estimates in meta-analysis are almost invariably discussed as representing differences in true effect sizes between included studies, ignoring the many alternative explanations.

**Generally, when comparable studies show different results, there is an undue focus on design heterogeneity as an explanation for these differences.** Design heterogeneity is often one of the first explanations for any ‘failure’ to replicate a study (by the reader’s preferred definition), of which the ‘replication crisis’ in psychology and other domains continue to provide many examples. However, artefactual explanations for statistical heterogeneity by necessity should be partitioned out first before considering design heterogeneity as an explanation. In addition, design heterogeneity is often an unlikely explanation for statistical heterogeneity in at least in cognitive and social psychology (Olsson-Collentine et al, 2020). Fundamentally, anything that affects effect size estimates may also affect statistical heterogeneity estimates. As such, researchers should make a concerted effort to rule out artefactual explanations of statistical heterogeneity before examining design heterogeneity as an explanation.

**The statistical results from comparable studies and their heterogeneity are most usefully examined through meta-analysis.** Examining statistical heterogeneity and moderators thereof is usually considered one of the primary purposes of meta-analysis. We begin this paper with first collecting data on how much heterogeneity meta-analyses in psychology tend to attribute to moderators. We then describe and discuss all 7 explanations for statistical heterogeneity, followed by an illustration of these explanations in a large meta-analysis, before ending the paper with a summary and conclusion.