Project proposal: Explanations for unexplained effect size heterogeneity

Anton Olsson Collentine

2021-06-01

## Short summary

* Theoretical (mainly) paper consisting of introduction + 7 explanations + empirical section + 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.
  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
  + Methods & results (2 pages)
    - Description of empirical data collection
    - Discuss findings in the context of the section(s) 1- 7 that they affect
  + Conclusion (1/2 page)
* Total: ~15 pages

## Detailed summary

Below I present a detailed summary of

1. the introduction: for a better description of the proposed angle of the paper.
2. Methods & results section: my proposals for the empirical data collection. Needs to be discussed to set the scope 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 usually examined through meta-analysis, and this is also the tool we use throughout this paper to discuss explanations for statistical heterogeneity.** Examining statistical heterogeneity and moderators thereof is usually considered one of the primary purposes of meta-analysis. To enhance the discussion of statistical heterogeneity , we will also collect empirical data relating to heterogeneity estimates in meta-analysis. We first describe and discuss all 7 explanations for statistical heterogeneity, then present a short methods section and discuss relevant sections in the context of our findings, before ending the paper with a summary and conclusion.

### Methods & Results

* This section has a large impact on the scope of the project and needs discussion on what we should actually do.
* I first describe 3 research questions we might try to answer by collecting data, and then make different suggestions to limit the scope.

#### Research questions and empirical data collection:

1. To what extent does meta-analysts’ usage of I2 instead of tau inflate heterogeneity estimates?
   * Extract I2 and harmonic mean
2. To what extent does the smaller average number of studies in medical meta-analyses (compared to in psychology) represent stricter inclusion criteria? (as opposed to smaller population of studies)
   * Extract proportion excluded studies from flowcharts
   * Extract number of exclusion criteria
   * Extract type of exclusion criteria (settings, sample population, measurement or treatment variables, risk of bias)
3. To what extent and which type of moderators explains statistical heterogeneity in meta-analyses?
   * Do the meta-analyses examine moderators? {yes/no}
   * What % of statistical heterogeneity do moderators explain?
   * How many moderators are examined, and what type?
     + Settings, sample population, measurement and treatment variables, risk of bias
     + What % does each type explain?
   * What is the power of the meta-analysis to find significant moderators?

#### Scoping (suggestions for different levels of empirical data collection)

* The moderator part above might be an entire paper by itself and we may not want to do everything. Below 5 suggestions for limiting the scope of data collection in this paper. Other suggestions?
* My inclination would be towards suggestion 3/4 below, which would leave most of the analysis of moderators in meta-analysis to be part of a future project.

Suggestions:

1. Extract:
   * I2 + harmonic mean
   * Proportion excluded studies (medicine vs. psychology)
2. Extract:
   * I2 + harmonic mean
   * Proportion excluded studies (medicine vs. psychology)
   * PLUS: Number and type of exclusion criteria
3. Extract:
   * Proportion excluded studies (medicine vs. psychology)
   * Number and type of exclusion criteria
   * PLUS: How many moderators examined and what type [e.g., psychology might have fewer exclusion criteria, but more moderators]
4. Extract:
   * Proportion excluded studies (medicine vs. psychology)
   * Number and type of exclusion criteria
   * How many moderators examined and what type [e.g., psychology might have fewer exclusion criteria, but more moderators]
   * PLUS: I2 + harmonic mean
5. Extract:
   * Everything, including all the moderator stuff

#### Sample

What would the sample look like?

* A random sample of meta-analyses from medicine and psychology
  + Already collected sample(s) from some project? Our inclusion/exclusion criteria? Sample size?