# Heterogenity in Psychology: a Mixed Methods Meta-Review using a Content Analysis and a Scoping Review

B.A Edmar

"if you don't have any theory about how you're going to explore the (statistical) heterogeneity [...] then [...] exploring it just means you spent more time doing it and not learning much more" – participant 17 (Lorenc et al., 2016)

## 1. The Question

The main goal of this paper is to address how heterogeneity impacts the validity of evidence synthesis. Specifically, we want to examine the problem of comparability across pooled effects and how researchers deal with problematic heterogeneity. Another way to phrase this question could be: at what point are qualitative differences between scientific inquiries into the same underlying psychological construct large enough that we conclude that the effects we are measuring are different; regardless of whether a measurable difference is present? In essence, we want to know what causes the exclusion of a study from a meta-analysis due to perceived methodological diversity or heterogeneity. In order to answer this question, a semi-mixed methods approach is needed. This is because we want to qualitatively examine heterogeneity issues in systematic reviews, and the qualitative AND quantitative heterogeneity issues in meta-analyses. To do this a qualitative content analysis of the systematic review literature, and a scoping review of the meta-analysis literature will be conducted.

## 2. Introduction

Systematic reviews and meta-analyses are usually described as the gold-standard in terms of providing evidence for scientific theories. However, for these types of analyses to be conducted a suitably body of literature is needed. What makes a literature suitable for review is therefore an important question to ask in order to gauge whether a synthesis of evidence is possible. The

answer to this question depends on the field in which the review is conducted, and the type of research question asked.

One of the main reasons to not conduct a meta-analysis even though a suitably large body of literature on a topic exist, is that the individual papers within that literature are so different from one another that the comparability of the individual studies comes into question. This variability across studies is commonly referred to as between-study heterogeneity. In this context heterogeneity refers to the variation in observed effects. While this definition is simple enough, it does not capture the nuances of what variability in observed effects *mean*. In order to set up our working definition of heterogeneity we need to make some clarifications on what we mean by effects and what we mean by variability.

## 2.1 Effects and Variance

With effects we mean the counterfactual difference between states dependent on whether something is present or not, and if present, to what degree it is so. If we assume that an effect has a single true value that we can estimate and that the variation in that effect tends towards zero as our sample size increases we assume that the effect is fixed. If we assume that the effect varies dependent on some underlying parameters, we assume that the effect is random.

With variability, we mean the mathematical concept of variance. Variance describes the average squared distance from a mean. It provides a measure for how varied a set of data points are with large values indicating high variation and low values indicating low variation. Another way to phrase this is that variance describes our degree of uncertainty about a mean. That is, to what degree we can expect to be wrong about estimating a mean value.

## 2.2 Heterogeneity and Uncertainty

One of the main distinguishing features between heterogeneity and variance is that heterogeneity has an accompanying distribution which indicates what sort of dispersion we can expect from studies measuring the same effect. Both variance and heterogeneity are measures of uncertainty, but with heterogeneity we have a framework to discern whether the observed variation is a product of naturally occurring sampling error or indicative of having captured multiple true effects in the pool of studies. This is the foundation of why some refer to heterogeneity as unexpected variation, or uncertainty about a mean value to the extent that it does not fit the expected distribution of a true effect(REF). In this paper we will not distinguish between inherent sampling error of true effects and the measurement of variation in effects that might consist of having multiple true effect sizes in a sample. The reason for this is because the notion of a *true* fixed psychological effect in the context of a meta-analysis is rather abstract. As with all constructs in social science, effects observed in psychology are dependent on multiple unobserved variables that could explain much of the variation in those effects. Therefore we will disregard definitions of heterogeneity that claim to be indications

of ''unexpected variation" or variation due to having captured multiple true effects - since all observed effects of interest in psychology are the product of multiple true effects. This is especially true when we define our effects as latent, which is the case when we use methods like structural equation modelling (SEM) or meta-analysis.

Because we assume that latent psychological effects are conditional on multiple unobserved variables, we are making a claim that are uncertainty about the effect can be reduced if we have more information. That is, our uncertainty is *epistemic*, or, knowable. In the context of a meta-analysis which aims to estimate a latent effect, the inclusion of moderating variables are endeavours to reduce our uncertainty about the effect. In a situation where no more information exists to further explain potential variation in an effect, our uncertainty about it is *aleatory*, meaning that it is random in the sense that it follows a probability distribution in which no parameter can predict its outcome. To our knowledge, this has never happened in the field of psychology, and the often low explained variability in psychological constructs is a testament to this(REF).

Since the concept of heterogeneity is closely tied with that of epistemic uncertainty, how heterogeneity is categorised is often dependent on where the source of the uncertainty is attributed. This varies across fields of study, in medicine the Cochrane typology of clinical and methodological heterogeneity is often distinguished, and in political science lines are drawn between treatment contrasts, participant moderators, and contextual moderators. The only true agreement on what heterogeneity means comes in it's statistical form, namely $\tau^2$, meaning effect variance. Henceforth when we refer to heterogeneity we will be talking explicitly about $\tau^2$ under the assumption that our uncertainty about the effect ($\mu$) is mostly epistemic. That is, the majority of the variation in an effect is attributable to it consisting of multiple true effects, which in theory are knowable.

While this notion of what we can call ''epistemic heterogeneity" may seem slightly arcane to most applied researchers, it is a useful terms since it has statistical underpinnings but is not necessarily an emergent or caused effect like observed statistical heterogeneity. In other words, we can see epistemic heterogeneity as the cause that results in the effect which is observed statistical heterogeneity. This is a theoretical claim that results in the assumption that statistical heterogeneity cannot exist without epistemic heterogeneity, but the absence of statistical heterogeneity is not indicative of an absence of epistemic heterogeneity. This in turn illustrates the importance in evaluating the potential sources of epistemic heterogeneity carefully before conducting any type of evidence synthesis.

This conundrum lies at the heart of the goal with this paper since we want to assess where the comparability between studies breaks down and forces the authors to conduct a purely qualitative evidence synthesis that does not mathematically combine measures across studies.