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 clarify the concept of heterogeneity in systematic reviews and meta-analyses within psychology. 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 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 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. There are many types of heterogeneity, and it's typology will often be dependent on what field of study you are in. In this paper we will use the terms as Lorenc et al. (2016), who describes heterogeneity as either statistical or substantive. Statistical heterogeneity refereing to the quantitative measurements of variability across pooled effect measures, and substantive heterogeneity being the individual differences across studies populating the sample of a review. These individual differences include but are certainly not limited to research aims, methodology, participants and outcomes. In the following section I will go into further detail about statistical heterogeneity and the different facets of substantive heterogeneity, as well as how they appear in psychological science.

2.1 Statistical heterogeneity

Statistical heterogeneity is the mathematical operation that quantifies the degree of variation across effect sizes in a study population. Just like the variance of any psychological measure or scale describes the distribution of that measure, heterogeneity describes the distribution of the effect measures included in a review. The statistic meant to capture this variation is τ^2, and is synonymous with σ^2 in samples where the units of measurement is not individual study estimations of an effect size. Another measures of heterogeneity is the I^2 statistic, which is a relative measure capturing how much of the variation between studies that cannot be attributed to random sampling (Higgins & Thompson, 2002). If the heterogeneity is high, we often assume that the studies included in the sample have multiple 'true' effects. In cases like this, treating the effect measures of the studies as random captures the distribution of our sample better than a fixed effect. When we treat effect measures as random, we assume that there exists an underlying 'true' effect which all observed effects are sampled from, but each sampled effect also has a sampling distribution. With random effects, the variation in the 'true' underlying effect distribution is the heterogeneity. Therefore, one way to view statistical heterogeneity is as the counterfactual *effect* resulting from the *cause* of having multiple true effects in a single sample. In the next section on substantive heterogeneity, applied examples are given of how multiple 'true' effects can be introduced into a sample of studies meaning to measure the same psychological phenomenon.

2.2 Substantive Heterogeneity

2.2.1 Cochrane Handbook

A common typology of heterogeneity is that from Cochrane Handbook(Cumpston et al., 2019) which distinguished between clinical, methodological and statistical heterogeneity. An example of how these types of heterogeneity could be applied to a psychological context could be through a review of the effectiveness of cognitive behavioural therapy (CBT) on social anxiety.

Clinical heterogeneity refers to differences in the characteristics of the study populations or the interventions being studied. Thus, clinical heterogeneity could arise from variations in the age, gender, severity of social anxiety, or other demographic(clinical factors) of the participants across th eincluded studies. Additionally, differences in the CBT protocols used, such as the duration and intensity of therapy falls under the category of clinical heterogeneity.

Methodological heterogeneity relates to differences in the design of the studies included in the meta-analysis. This includes variations in study design, outcome measures, assessment tools, and the quality of the research design. For example, if some studies on CBT for social anxiety use self-report questionnaires while others use clinical interviews as outcome measures, methodological heterogeneity could be introduced. Furthermore, differences in randomisation and blinding will contribute to methodological heterogeneity.

2.2.2 Weiss et al. (2014)

Weiss et al. (2014) provides a similar but slightly different taxonomy of heterogeneity. Like Cochrane, they argue that heterogeneity has tree sources: Treatment contrast, within participants moderators, and contextual moderators. An example of how these could be applied to a psychological context could be through a specific treatment effect such as that of the cold pressor test and perceived pain; the cold pressor test being the action off submerging you hand in cold water(e.g Kahneman et al., 1993).

Treatment contrast refer to the strength of the measured effect. For the cold pressor test the degree of coldness can be a source of difference in effects that then contribute to unexpected heterogeneity observed across the included studies examining this treatment. That is, how cold the water is will affect the strength of the measured effect.

Within participants moderators refer to variation coming from individual participants. For our cold pressor experiment we might have a single stimuli of water with a certain degree of coldness, but our sample consists of two unknown subgroups – one with a high pain tolerance and one with low pain tolerance. This difference within participants affects the strength of the effect, and thus is a systematic source of variation which we cannot attribute to normal sampling error – thus providing heterogeneity.

Contextual moderating factors are factors external to the participants and separate from the stimuli provided to elicit the effect. They cover contextual aspects of the 'study environment' that can cause potential differences in effects. For the cold pressor test the coldness of water could be moderated by the temperature of the room the test is conducted in. This could artificially introduce variation in the effect due to the context of the measurement and again provide heterogeneity.

2.2.3 Li & Reynolds (1995)

Li & Reynolds (1995) takes a different approach to heterogeneity based on the issue of knowing exactly what is being quantified in statistical heterogeneity within the field of ecology. That is, they approach heterogeneity not as a product of an individual effect, but as a product of a system of interest such as the biomass of a plant or the nutrients in soil. To them heterogeneity is the complexity and the properties of a system across space and time. Adapting this perspective to psychology is beyond the scope of this section, but it could provide a fruitful alternative when synthesising evidence of broad psychological systems such as emotional development or decision making under pressure.

2.2.4 A Heterogeneity Framework

It is important to note that the contexts of these perspectives on heterogeneity differs, the Cochrane taxonomy from medicine, Weiss et al. (2014) from policy/political science, and Li & Reynolds (1995) from ecology. This diversity in perspectives and terminology comes with both the pros and cons of nuance; we get a broad and informative framework for how to view sources of variation, but we also become prone to misunderstandings in what we attribute perceived 'heterogeneity' to. Therefore, the term substantive heterogeneity will be used to refer to any type of heterogeneity that is not statistical. To tie the two concepts together as we did in the section on statistical heterogeneity, we can see substantive 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 substantive heterogeneity, but the absence of statistical heterogeneity is *not* indicative of an absence of substantive heterogeneity. This in turn illustrates the importance in evaluating the potential sources of substantive heterogeneity carefully before conducting any type of evidence synthesis.

This conundrum lies at the heart of the goal with this paper. We want to examine the relationship between substantive heterogeneity in systematic reviews, and the overall assessment of heterogeneity in meta-analyses. We do this 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. In the next section we will detail a proposed methodology for answering this question.

3. Methodology

This paper will consist of two dependent studies, one scoping review and one qualitative content analysis. A mixing procedure where the result from both studies are thematically synthesised to provide a nuanced answer to our research question will also be conducted.

3.1 The Scoping Review

The scoping review will target issues of heterogeneity present in meta-analyses. Given that our goal is to clarify the concept of heterogeneity, and assess how the execution of research is impacted by the presence of heterogeneity, a scoping review is preferable to a systematic review(Munn et al., 2018). An alternative to a scoping review could be a rapid review; that being a feasibility constrained systematic review. However, it is not the feasibility of our investigation that informs our method - it is the underlying research question. Since we are mainly concerned with the qualia of methodological heterogeneity, a quantitative assessment of the literature does not target the question; even though the literature itself is quantitative in nature.

3.1.1 Eligibility Criteria

To be included in the systematic scoping review, any document needs to be a meta-analysis within psychology science published after 2021 to capture recently published studies. Additionally, documents must have an applied or theoretical research question that requires the synthesis of evidence across multiple independent studies to be included. This means that studies posing meta-research questions will be excluded. An example of such a question would be an evaluation of the prevalence of open-science practices or replications within a given field. The data within the documents must also have been generated through a systematic search of literature, this excludes sequential studies that aggregate their findings using an 'internal meta-analysis'.

Scoping reviews investigate a population and a concept in a specific context. For this review the population under investigation are published meta-analyses in psychology. The concept that we want to investigate is the handling of statistical and substantive heterogeneity across studies. The context we are interested in is recently published documents. We are not interested how this has been done historically but how it is currently being done.

3.1.2 Methodology

This review will follow the guidelines provided by JBI to the greatest extent possible for the research question at hand(*JBI Manual for Evidence Synthesis*, 2020). The manuscript will

follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018).

3.1.3 Search Strategy

The JBI(*JBI Manual for Evidence Synthesis*, 2020) guidelines state that at least two databases should be subject to an initial search, thereafter a second search using all identified keywords should be conducted, and then the reference list of all identified sources should be consulted for additional documents. While these guidelines are sound, following this strategy is not feasible in our case due to the breadth of the literature we are interested in. Therefore we deviate from the JBI manual in this area.

Since the field we are interested in is psychology, searching databases outside psychology will have to include field limitations that would not be needed in psychology specific data bases such as PsychInfo. Therefore, we will only search PsychInfo, since that database provides us with a natural exclusion of documents outside the field of psychology. As an extension of this strategy we use PsychInfo classification categories to get a representative sample of papers across the many sub-fields within psychology. To this end we limit our search to papers within the following categories: Developmental Psychology (2800), Social Psychology (3000), Personality Psychology (3100), Organizational Psychology (3600), Cognitive psychology and Intelligent systems (4100), Forensic psychology & Legal Issues (4200). These categories are selected to ensure that a representative sample of different areas of psychology are searched for while limiting the number of search results. Note that health and clinical categories are excluded, this is due to the heavy overlap in medical and psychological research in these areas.

3.1.4 Search Terms

The search was conducted through the data base Ovid in three steps detailed in table 1.

Table 1

	Search Terms
1	(meta-analy* or meta analy*).mp. [mp=title, abstract, heading word, table of contents, key concepts, original title, tests & measures, mesh word]
2	limit 1 to (2800 developmental psychology or 3000 social psychology or 3100 personality psychology or 3600 organizational psychology & human resources or 4100 cognitive psychology & intelligent systems or 4200 forensic psychology & legal issues)
3	limit 2 to yr="2022 -Current"

The initial search resulted in 53 041 hits, the first limitation reduced this to 2015 hits, and the final limitation produced 283 documents.

3.1.5 Document screening

The identified documents will be screened using the software tool rayyan.ai. The screening will be conducted in a two stage fashion where the abstracts from the initial search are screened for eligibility and labelled according to the sub-field of psychology they cover, after which full text screening will be conducted on a subset of the identified studies. As of Thursday the 12th of October 2023 the initial screening of the abstracts is complete, resulting in 165 documents. The next step is to formalise a procedure for how to select studies to analyse from this pool of documents. Since the goal is to produce generalisable results, a diverse and representative sample is required, but given that the analysis is mostly qualitative a smaller purposefully selected sample of documents is needed.

3.1.6 Data Extraction

An excel code book for what elements to extract from each study will be constructed. This code book will cover three main themes: general study characteristics, the presence of substantive heterogeneity in the review, and how the authors dealt with the observed statistical heterogeneity. Across these three themes we will include notation of the aims of the study, the sub-field of the study, the population examined, the methodology used, outcome variables, heterogeneity measures and adjustment, exclusion criteria, findings/conclusion of the study, etc.

3.1.7 Data Analysis

The analysis of the data will consist of summarizing the motivation for the various choices made in the individual data points. These will then be compared across sub-fields to see how the practice of adjusting for heterogeneity varies dependent on the field. These will be presented with simple descriptive statistics but also with a qualitative content analysis of how authors reason about their analysis choices and exclusion criteria with regards to heterogeneity. This content analysis will aim to be descriptive and will not seek to find any emergent or latent themes within the data. The result will be an evidence map of how substantive heterogeneity is accounted for in the exclusion criteria and the theory under investigation as well as how the statistical heterogeneity is handled by the authors.

3.2 The Qualitative Content Analysis

A content analysis, as described by Elo et al. (2014), can be defined as "A flexible method for making valid inferences from data in order to provide new insight, describe a phenomenon through concepts or categories, and develop an understanding of the meaning of communications with a concern for intentions, consequences, and context." This makes it an apt method for our purposes since we aim to find cases of reported problematic heterogeneity in systematic reviews. We mean to locate instances where the apparent goal of the review is to pool data for a meta-analysis but the perceived substantive heterogeneity was too large for a valid synthesis. That is, we are specifically looking of reviews that mention aspirations of meta-analytic pooling, but due to substantive heterogeneity, they could not be compared.

In order to make valid inferences about this phenomenon, we need to take context, intentions and communication from the authors conducting the study into account, thereby making a qualitative content analysis an appropriate method.

3.2.1 Search for Secondary Sources

To locate studies of this kind, a search for literature similar to that of the scoping review will be conducted with the addition of searching for heterogeneity in the abstracts, titles or keywords. Since the only type of heterogeneity present within systematic reviews is substantive, this additional keyword will help locate instances where the heterogeneity of the study population is challengingly high. On October 18th a literature search was conducted using the terms and limitations as can be seen in table 2.

Table 2

	Search Terms
1	systematic review AND heterogeneity {Including Related Terms}
2	limit 1 to (2800 developmental psychology or 3000 social psychology or 3100 personality psychology or 3600 organizational psychology & human resources or 4100 cognitive psychology & intelligent systems or 4200 forensic psychology & legal issues)
3	limit 2 to yr="2018 -Current"

This first search yielded 22 662 results, the first limitation reduced this to 1658 and the final limitation resulted in 644 documents. A preliminary screening conducted the same day as the initial search found 13 eligible systematic reviews mentioning difficulties in pooling due to substantive heterogeneity.

After locating a set of systematic reviews mentioning between-study heterogeneity in their abstract, key-words or title as a problematic aspect of their synthesis, full text content analysis will be conducted.

3.2.2 Methodology

The qualitative content analysis will be conducted using an inductive analysis with a critical realist framework. However, since the topic under investigation has a long history of study, we will be using the theoretical scaffolding of heterogeneity typology and the considerable work on validity stemming from the philosophy of science (Shadish et al., 2002), thus the study will not be purely inductive.

The analysis will follow the standard procedure of identifying studies, immersing oneself within the data, and developing the codebook as an information saturated picture of the concept under study is painted. The hermeneutic circle will be utilized in the reading and coding in order to update our understanding as new information is acquired. An a priori Epoché (Bracketing) of the analysis will be provided in order to ground the analysis in theory as well as provide some transparency regarding the authors motivation and expectations from the analysis.

3.3 Mixing

Following the guidelines provided by JBI [*Chapter 8* (2020)], the mixing procedure will follow a convergent design where each study informs the other to answer our singular research question of how substantive heterogeneity is present in the literature. We will utilise 'qualitising' transformation where the data from both the scoping review and the qualitative content analysis will be integrated using a qualitative synthesis. Meaning that the scoping review of the meta-analyses will be translated to fit the codebook of the qualitative content analysis of the systematic reviews. Ideally, the results will then be integrated using a thematic synthesis through which our original research question can be answered [Thomas & Harden (2008)].

References

Chapter 8: Mixed methods systematic reviews. (2020). JBI. https://doi.org/10.46658/jbimes-20-09

Cumpston, M., Li, T., Page, M. J., Chandler, J., Welch, V. A., Higgins, J. P., & Thomas, J. (2019). Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.ed000142

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative Content Analysis: A Focus on Trustworthiness. *SAGE Open*, *4*(1), 215824401452263. https://doi.org/10.1177/2158244014522633

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

JBI Manual for Evidence Synthesis. (2020). JBI. https://doi.org/10.46658/JBIMES-20-01

Li, H., & Reynolds, J. F. (1995). On definition and quantification of heterogeneity. *Oikos*, *73*(2), 280. https://doi.org/10.2307/3545921

Lorenc, T., Felix, L., Petticrew, M., Melendez-Torres, G. J., Thomas, J., Thomas, S., O'Mara-Eves, A., & Richardson, M. (2016). Meta-analysis, complexity, and heterogeneity: A qualitative interview study of researchers' methodological values and practices. *Systematic Reviews*, *5*(1), 192. https://doi.org/10.1186/s13643-016-0366-6

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin; Company.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1). https://doi.org/10.1186/1471-2288-8-45

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A CONCEPTUAL FRAMEWORK FOR STUDYING THE SOURCES OF VARIATION IN PROGRAM EFFECTS. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760