Exploring the Comparability of Psychological Studies: a Mixed Methods Meta-Review via Content Analysis and Scoping Review

B.A Edmar

1. The Question

With this paper we aim to answer how heterogeneity in effects and diversity of methodology impacts our assessment of evidence in systematic reviews and meta-analyses. At some point, the observed heterogeneity in a sample of effects becomes so large that the uncertainty of the inferences that can be made about a construct or theory is called into question. We want to get a better understanding of when this occurs using psychological science as an example. To do this we take a look at successful pooling of evidence in meta-analyses, but also at instances where pooling was not possible due to prominent differences across studies. With this two-pronged approach we aim to paint a nuanced picture of the line that separates qualitative and quantitative evidence synthesis; with the goal of gaining insight into how we can leverage past research for theory building in the present.

2. Background

Systematic reviews and meta-analyses are usually referred to as the gold standard for evidence based inference, meaning that the synthesis of evidence is considered the strongest scientific case one can make for a certain theory. The strength of these methods have their root in their use of multiple independent sources that together makes up a collection of results in which all available evidence can be taken into account when reaching a conclusion. However, taking all available evidence into account is not always easy or free from controversy. How exactly can we determine when different sources provide evidence for the same construct or theory? And when is a body of literature suitable for synthesis?

These questions are explicitly or implicitly answered in the beginning stages of a review process, where the researchers define the research question and the eligibility criteria needed for studies to be included in the review. However, in some cases the literature of interest is not suitable for reviewing or meta-analytic pooling. This occurs when the methods used across studies are so varied that comparing them is not appropriate or possible. This is commonly referred to as the "Apples and Oranges'' problem (Harrer, 2022) but is more precisely described as an issue of between-study heterogeneity. However, it should be noted that there are many interpretations to what we describe as between-study heterogeneity. In the following sections we will describe the key concepts explored in this study along with our definitions of them.

2.1. Heterogeneity

The strictest and most precise definition of heterogeneity is the variability of observed effects. That is, the variance within the pooled effect sizes of the studies included in the given meta-analysis. Explicitly, heterogeneity is the average squared distances from the mean of the true effect. This definition becomes problematic when the variation in effects is so large that their distribution is not indicative of coming from a single true value

given that it is normally distributed. In these cases, viewing the variance of an effect as purely stemming from random sampling is not appropriate. Thus, we make distinctions for situations when we assume that the total variation of effects is attributable to within study variation and when the total variation can be attributable to both between-study variation AND within-study variation. Hence forth, when we say heterogeneity we refer to the total variation in effects, regardless of whether we attribute it to within-study(sampling error) or between-study(heterogeneity) variance. The reason for doing this is that distinguishing between these sources is incredibly difficult since properly defining the "true" effects we are interested in is very hard in psychological settings (Shadish et al., 2002). Thus, what we mean by heterogeneity is the total variation of effect sizes expressed through the τ^2 statistic.

2.2 The Causes of Heterogeneity

With our definition of heterogeneity as the total variation of effects, we can add some nuance to its occurrence by viewing it as a counter-factual effect. Meaning that it is something that occurs under some conditions but not others, that is, it is an effect caused by some underlying factors. Exactly what these factors are will be dependent on the field of study you are in. Multiple guidelines and papers on how to approach heterogeneity when conducting evidence synthesis exist, and they use different terminology to refer to these sources (*Cochrane Handbook for Systematic Reviews of Interventions*, 2019; Li & Reynolds, 1995; Weiss et al., 2014). In general, sources of heterogeneity are categorised as stemming from the units under investigation or the methodology used to study those units. In psychology our units of study are almost exclusively humans and therefore all variability in effects attributable to differences in the humans we study is an example of heterogeneity being caused by the units under investigation. The way we study humans, be that through the design and context of experiments, measurements or modes of inference, is an example of heterogeneity being caused by methodological diversity. It should be noted that the presence of diversity, either in units or method, should not be seen as something that necessitate observed heterogeneity. However, whenever we observe heterogeneity, we must assume it to be caused by either underlying diversity in units and methodology, or through random sampling/variability.

2.3 Validity

One of the reasons behind why isolating true effects in psychology or any social science is hard is because defining and knowing what our measurements are measuring is impossible. Any non-physical unit of measurement will have this issue. This problem is what inspires areas of research interested in *validity*. What constitutes validity is a philosophical questions that targets our subjective beliefs about epistemology and ontology (Macklin & Gullickson, 2022). For the purpose of this paper, validity refers to the reliability and accuracy of an inference. There exists many frameworks for categorising validity, since we are mainly interested in the

validity in terms of inferences we will utilize the Campbellian framework (Macklin & Gullickson, 2022; Shadish et al., 2002).

Briefly put, this framework posits that there exists four types of validity which influences the quality of our inferences. These are external, internal, construct, and statistical validity. External validity is the extent to which inferences holds to and across different settings, participants, persons and outcomes (Lucas, 2003). Internal validity, or more accurately put, local molar causal validity, refers to whether an experiment capture a causal relationship between outcome and treatment. Cook, Campbell and Shadish says: "internal validity is about whether a complex and inevitably multivariate treatment package caused a difference in some variable-as-it-was-measured within the particular setting, time frames, and kinds of units that were sampled in a study". Construct validity refers to the congruence between the variable-as-it-was-measured and the psychological concept/construct we are interested in. Statistical validity refers to whether the statistical conclusions drawn from the analysis accurately represents the observed data.

Together these validity types create a theoretical way of assessing the quality of the inferences we can draw from a paper. External validity targeting the generalisability of units, construct validity targeting the generalisability of the measures, and internal validity assessing whether causal relationships can be established. To fully understand how heterogeneity is caused, we need to evaluate not only the validity of individual studies but the validity of the synthesis and the inferences we draw from our review.

3. Methods

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 enrich the answer to our research question will also be conducted. Though there are surface level similarities between these two analyses, they are categorically different and their results cannot be directly compared. It is through the mixing procedure where an answer to the over-arching research question can be addressed. Simply put, we want to know when heterogeneity is an issue, but to properly answer this question we need to analyse when pooling data was judged to be mostly a statistical issue and when pooling data was judged to not be possible on qualitative grounds. We address the former with a scoping review and the latter with a qualitative content analysis. Figure 1 depicts a flow chart of how our overarching general research questions gets partitioned into more targeted research questions which require different modes of inquiry, and then how the results from both analyses are integrated to answer the initial question.

Figure 1: Design Flow Chart

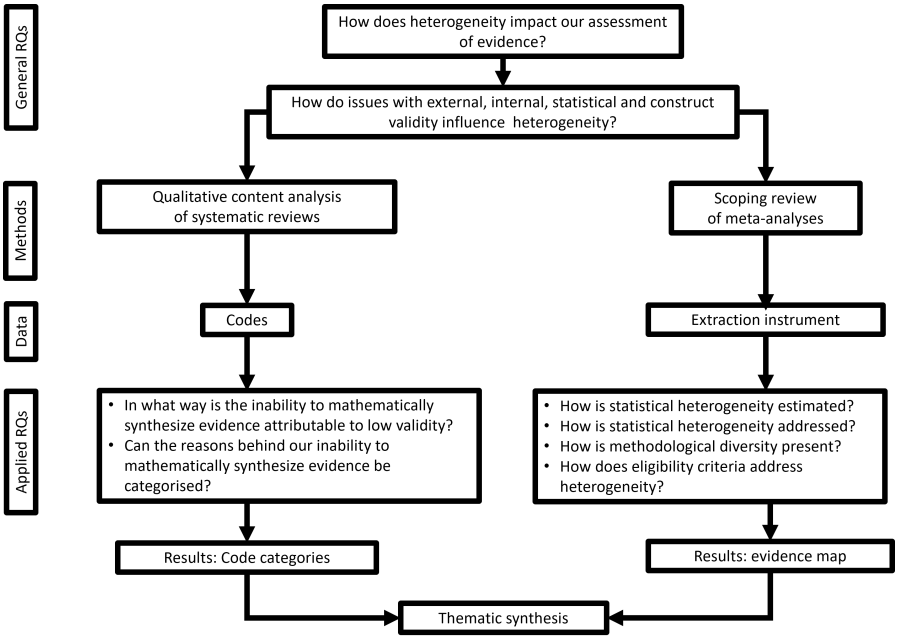

Figure 1: *Note*: This flowchart depicts the research design going from the initial general research questions to the final result of thematic synthesis.

3.1 The Scoping Review

A scoping review is a type of evidence synthesis akin to a systematic review or a traditional literature review. The main difference between a systematic review and a scoping review is that the scoping review is a more iterative process designed to identify knowledge gaps, scope out a body of work, clarify concepts or investigate research conduct. As such, the synthesis of the scoping review is not designed to provide a summary of research findings from individual studies; but generate broader mapping of evidence, not unlike a traditional literature review.

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).

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). This means that an a priori protocol of the review is formulated which includes the exclusion criteria, search strategy, methodology and reporting of the manuscript. One deviation from the recommendations is present in how we search for literature. 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 JBI guidelines are compatible with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018), that reporting scheme will be followed in the manuscript.

3.1.1 Eligibility Criteria

To be included in the systematic scoping review, any document needs to be a meta-analysis within psychological science published in 2021 or 2022 to capture recently published studies. Additionally, documents must have a 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 a meta-research 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'.

In sum, for a study to be included in the scoping review it needs to: 1) be a meta-analysis conducted using standard literature search strategies; 2) cover a psychological concept; and 3) be published in 2021 or 2022.

3.1.2 Search Strategy

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 APA 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: Physiological Psychology & Neuroscience (2500), Developmental Psychology (2800), Social Psychology (3000), Personality Psychology (3100), Organizational Psychology (3600), 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. Table 1 depicts the iterative searches done through the search engine Ovid and shows how the number of results are reduced as limitations are put on the search terms.

Table 1

	Search Terms	Results
1	meta-analy* or meta analy*{Including Related Terms}	30 330
2	limit 1 to (2500 physiological psychology & neuroscience or 2800 developmental psychology or 3000 social psychology or 3100 personality psychology or 3600 organizational psychology & human resources or 4200 forensic psychology & legal issues)	1188
3	limit 2 to yr“2021 - 2022”	201

3.1.3 Document Screening and Selection

The identified documents was screened using the software tool rayyan.ai, the main purpose of this was to utilise the rayyan screening environment and thus we did not use any automated tools. The screening was conducted in a two stage fashion where the abstracts from the initial search was screened for eligibility. After locating eligible studies, a stratified random selection of 5 studies from each of the six APA classification categories was conducted to get a total sample of 30. However, only 3 of the 5 studies selected from the Personality Psychology category could be located, which gave us a final sample of 28 articles. An over view of the entire search and selection procedure can be seen in the PRISMA flow chart in figure 2.

Figure 2: PRISMA Flow Chart

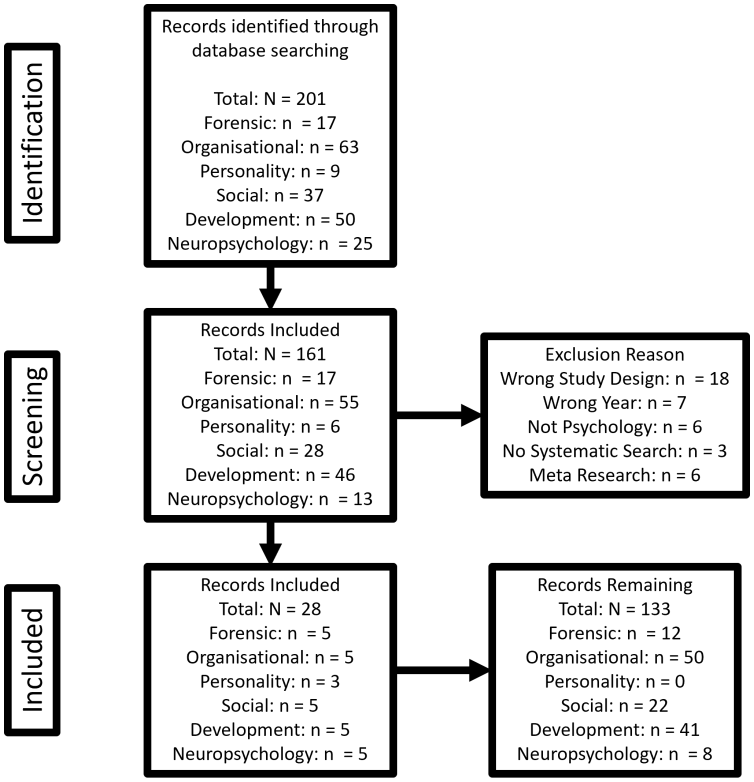

Figure 2: *Note*: This flowchart depicts the entire selection procedure starting with the identification of sources, followed by eligibility screening, and finalised by randomly sampling five studies from each sub-field to be included in the analysis.

3.1.4 Data Extraction

The data from the identified studies will be extracted into an excel code book based on an a priori constructed extraction instrument as is proper for scoping reviews (Munn et al., 2018). The extraction instrument is designed to answer a series of research questions derived from potential sources of heterogeneity based on the four Campbellian validities (Shadish et al., 2002). The code book covers three main themes: general study characteristics, the presence of 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, exclusion criteria, findings/conclusion of the study, etc. The extraction instrument can be found in the supplementary materials.

3.1.5 Data Analysis

The analysis of the data will consist of summarizing the motivation for the various choices made in the individual data points. These results will be presented with simple descriptive statistics but also with a qualitative analysis of how authors reason about their analysis choices and exclusion criteria with regards to heterogeneity. These results will then be compared across sub-fields to see if the practice of adjusting for heterogeneity varies dependent on the field. The qualitative 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 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

While we have a similar goal with our content analysis as with our scoping review, these methodologies are different from one another. In a content analysis, no systematic protocol is needed, and no synthesis of evidence is conducted. Instead, a fully qualitative approach is taken where the results of the analysis cannot be fully divorced from it's author(s), meaning that its reproducibility is weak. This does not mean that transparency falls by the wayside or that measures to ensure inter-rater-agreement is reduced, but rather that the point of the analysis is not to give a systematically reproducible analysis of a literature.

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 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 a high degree of diversity in methodology or units, that could not be done.

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 was 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 non-statistical, this additional keyword helped us locate instances where the heterogeneity of the study population

was challengingly high. No eligibility criteria for inclusion exists, this search procedure is simply a means of generating data from secondary sources and should not be confused with a systematic literature search. Thus, the search period was loosened to include studies published after 2017 in order to capture current research without being constrained to particular years as was done in the scoping review. On October 18th 2023, the following search was conducted using the terms and limitations provided in table 2.

Table 2

	Search terms	**Results**
1	systematic review AND heterogeneity {Including Related Terms}	33 783
2	limit 1 to (2500 physiological psychology & neuroscience or 2800 developmental psychology or 3000 social psychology or 3100 personality psychology or 3600 organizational psychology & human resources or 4200 forensic psychology & legal issues)	2373
3	limit 2 to yr="2018 -Current"	955

A screening of the 955 documents on November first 2023 found 14 documents to be suitable for analysis since they indicated aspiration of mathematical synthesis, but due to high between-study heterogeneity, did not follow through on the analysis.

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. That is, we will be using a deductive-inductive approach (Hong et al., 2017).

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 (Cho & Lee, 2014). However, three initial categories consisting of external, internal, and construct validity will be taken from the extraction instrument of the scoping review in order to structure the codes. This will in turn allow us to create a categorisation matrix where we can bolster inter-rater agreement and provide a semblance of systematic objectivity in the coding of the articles. These categories also serve as an a priori Epoché

(Bracketing) of what code categories we expect to find in the systematic reviews. Through this we aim to ground the analysis in theory as well as provide some transparency regarding the authors motivation and expectations from the analysis.

After coding the articles and assigning the codes to our pre-defined categories, the hermeneutic-circle will be employed to redefine the categories and create new ones where needed. This will be done to extract as much information as possible from the systematic reviews and provide more nuance to the issue of heterogeneity in those types of evidence syntheses. This specific procedure is selected to suit our research question and the overall purpose of this paper. However, It should be noted that no absolute guidelines for how to conduct a qualitative content analysis exists, making it a flexible but challenging research method (Hong et al., 2017).

3.3 Mixing

Following the guidelines provided by JBI (*JBI Manual for Evidence Synthesis*, 2020), the mixing procedure will follow a convergent design where each study informs the other to answer our singular research question of how heterogeneity hinders evidence synthesis. We will utilise a 'qualitising' transformation where the evidence map from the scoping review will be integrated into the categories identified in the qualitative analysis (Thomas & Harden, 2008), (Hong et al., 2017). This will provide us with a thematic synthesis that can highlight the issues posed by the presence of heterogeneity in the assessment of evidence, both from a qualitative and quantitative perspective.

References

Cho, J., & Lee, E.-H. (2014). Reducing Confusion about Grounded Theory and Qualitative Content Analysis: Similarities and Differences. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2014.1028

Cochrane Handbook for Systematic Reviews of Interventions. (2019). https://training.cochrane.org/handbook

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

Harrer, M. (2022). *Doing meta-analysis with R: A hands-on guide* (First edition). CRC Press.

Hong, Q. N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews*, *6*(1), 61. https://doi.org/10.1186/s13643-017-0454-2

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

Lucas, J. W. (2003). Theory-Testing, Generalization, and the Problem of External Validity. *Sociological Theory*, *21*(3), 236–253. https://doi.org/10.1111/1467-9558.00187

Macklin, J., & Gullickson, A. M. (2022). What does it mean for an evaluation to be "valid"? A critical synthesis of evaluation literature. *Evaluation and Program Planning*, *91*, 102056. https://doi.org/10.1016/j.evalprogplan.2022.102056

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), 45. 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